人工知能学会研究会資料
SIG-FIN-015-08

板情報を用いた高速取引の挙動分析

Analyzing Quotes using Order Book Information

吉田 健一 [1*]　櫻井 彰人 [2]
Kenichi YOSHIDA[1]　Akito SAKURAI[2]

[1] 筑波大学 大学院 ビジネス科学研究科
[2] 慶應義塾大学 理工学部 管理工学科

Abstract: 本研究では効率的市場仮説が仮定する情報伝搬の均一性が高速取引では成り立っていない事を板情報のデータを用いて示す。具体的には、1) 板情報を用いて標準的な教師なし学習の枠組みの中で短期の価格推移予測において 82.9%の精度をもったモデルを構築できる事を示し、2) 効率的市場仮説が示唆するランダムウォークとは見做せない事を報告し、3) その原因が情報伝搬の不均一性にある事を議論する。構築した価格予測モデルは直接的に証券取引に利用できるものではないが、板情報が証券価格の形成を分析するにあたり重要な情報源となりえる事、従来標準的に用いられていた時系列データの分析手法が証券市場の短期挙動の分析には適さない事などを示唆している。

1 はじめに

ファイナンス研究の分野では一般に効率的市場仮説 (Efficient Market Hypothesis, EMH [1]) が広く受け入れられており、株価はランダムウォークし予測は困難である事とされる事が多い [2]。一方、この仮説に異を唱える研究事例も存在する (例えば [3, 4, 5, 6, 7])。近年 Twitter など Social Network Service (SNS) に流れる情報をベースに株価を予測する研究 (例えば [8, 9, 10]) も増えている。SNS を使った株価の予測はランダムウォーク仮説が仮定する「あらゆる情報が効率的（迅速かつ正確）に価格に反映される」が、成り立たない期間 (情報の伝搬が不均衡である期間) の市場の動きを分析しているように思える。

本研究では、高速取引 (High frequency trading, HFT[11]) の過程でも「情報の伝搬が不均衡である期間」が存在すると考え、板情報を使って高速取引過程の分析を試みる。板情報は証券市場において「幾らであれば該当証券を購入したいか？販売したいか？」「市場参加者が証券の購入についてどう判断しているか？」の情報を直接表したデータであり、情報を入手した市場参加者の挙動を最も早く観測可能なデータ源である。

板情報として 2010 年 1 月から 2014 年 5 月まで 53ヵ月にわたる東京証券取引所 arrownet[12] のデータを用い、元データの加工方法を工夫すれば、標準的な教師なし学習の枠組みの中で短期の価格推移分析において 82.9%の分析精度をもったモデルを構築できる事を示す。この事は、板情報の使用を前提にすれば、価格の短期推移がランダムウォークとは見做せない事を示しており、高速取引の過程を分析する事で情報の不均衡な伝搬状態が観察できた事を示唆している。

構築した価格予測モデルは直接的に証券取引に利用できるものではないが、板情報が証券価格の形成を分析するにあたり重要な情報源となりえる事、従来標準的に用いられていた時系列データのデータ表現形式や分析手法 (例えば [13]) が証券市場の短期挙動の分析には適さない事なども示唆している。

以下、分析の基本的な枠組み、実験結果を報告した後、実験結果について分析を試みる。本研究が報告する価格推移のモデル化手法は予測と言うよりは計測に近い。最後の章で情報システムとしての高速取引システムの本質的な特性 (市場参加者の情報入手／判断のタイミングより高速な取引システムが情報伝搬の不均衡を表わにする特性) を計測している可能性について議論する。

2 証券取引情報とそのデータ表現形式

2.1 証券取引と板情報

東京証券取引所では、2010 年より売買・相場報道等の各システムと証券会社など取引参加者の間で取引情報を高速に交換するための高速ネットワークシステム arrownet を運用している (図 1)。証券取引の情報は、このネットワーク上の UDP パケットを使ってリアルタイムで取引参加者のシステムに送られている。この

*連絡先：筑波大学
〒 112-0012 東京都文京区大塚 3-29-1
E-mail: yoshida@gssm.otsuka.tsukuba.ac.jp

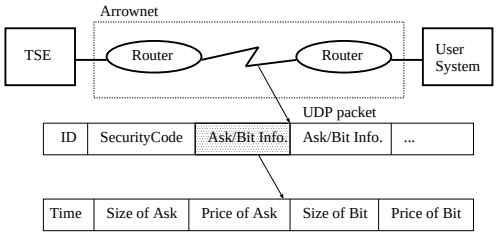

図 1: Arrownet における情報の流れ

表 1: 板情報の例

売数量	値段	買数量
...	...	
800	2026	
600	2015	
300	2000	
	1965	100
	1955	200
	1950	1800
	...	...

UDP パケットの情報には、いわゆる板情報 (表 1) として、価格と数量からなる売買希望の情報を含み高速取引の情報交換に使われている。

証券市場を分析する従来の研究では、株価 p の推移を単位時間間隔の時系列データ p_t として表現するアプローチが代表的である。例えば図 1 で示されたネットワーク上を UDP パケットが流れ、個々の UDP パケットが売買の希望価格と数量からなる注文の情報を伝えている状況 (図 2 Arrownet(Real World Trade Status)) は、図 2 Traditional Time Series Representation) のように成立した取引価格の時系列データ p_t として表現される。

図 2 Traditional Time Series Representation に例示したデータ表現では、固定間隔の時系列データ $p_1 = Q1, p_2 = Q1, , , p_7 = Q1, p_8 = Q2, 9 = Q2, , ,$ を作成するために実際の取引列 (A1, A2, B1, A3, B2, A4, B3, A3, B1) には含まれない時刻のデータは直前のデータから補完されるのが基本である。この表現形式では、何時取引が行われたかや、単位時間以下の取引間隔の情報は失われている。

この点に着目して本研究では注文が発生する毎にデータを記録する方法 (図 2 Event base Representation) について検討する。表現形式 1 では売り/買いの注文が発生する度に、その時刻と売り買いの価格/数量/最新の約定価格を記録し、その推移を分析する。表現形式 2 では売買成立のタイミングのみで表現形式 1 と同じデータを記録し、その推移を分析する。

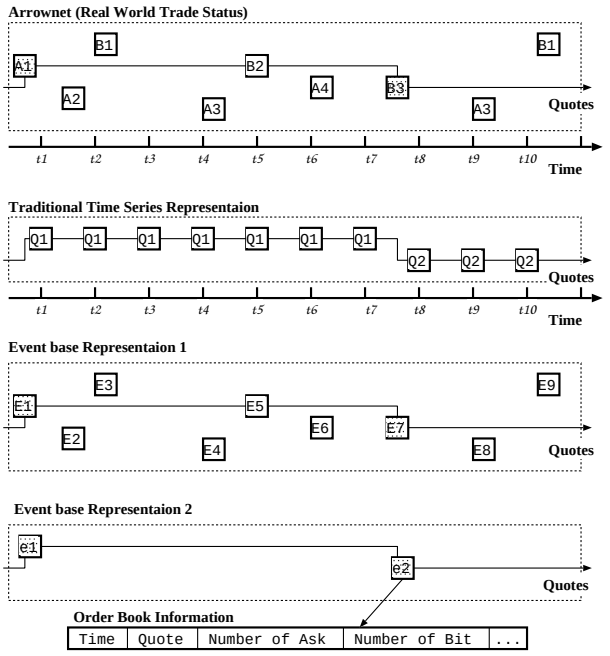

図 2: Arrownet と、そのデータ表現

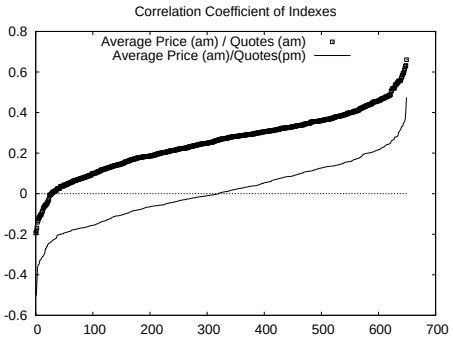

図 3: 板平均と歩み値の相互相関

2.2 予備実験の結果

表現形式 1,2 に記録する板情報としては板に含まれる売り買い価格の平均値 (詳細な計算式は次章 e_1 参照) などを考える。売り買い価格の平均値を分析に組み入れた背景には図 3 および 4 に示す予備実験の結果がある。

図 3 は売り買い価格の平均値の午前中の変化と約定価格の午前／午後の変化の Spearman の順位相関係数を示している。具体的には、2010 年 1 月から 2013 年 1 月まで取り引き量の多い証券 100 銘柄について証券毎に午前と午後の売り買い価格の平均値の変化と約定価格の変化を計算した後に、その順位相関係数を計算した。図は計算結果を数値順にソートして示している。図に示すように売り買い価格の平均値の午前中の変化と約定価格の午前中の変化には非常に弱い相関が見ら

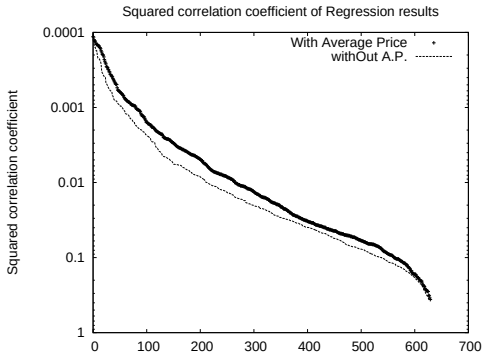

図 4: 板平均が SVR による歩み値予測誤差に及ぼす影響

れるものの、売り買い価格の平均値の午前中の変化と約定価格の午後の変化に相関は見られない。この結果は売り買い価格の平均値の持つ情報量を否定する結果にも見える。しかしながら、午前と午後のデータに相関が見られず、午前中のデータ間に弱い相関が見られる事を、我々は「約定価格の単期間の挙動について、売り買い価格の平均値は、何らかの情報を持つ」と解釈した。

図 4 は、SVM regression を用いて約定価格のモデル化を試みた結果である。データの表現形式は図 2 に Traditional Time Series Representation として示した形式を用いて、10 分後の約定価格を被説明変数に、遡って 30 分間の毎分の約定価格と売り買い価格の平均値を説明変数に用いた。売り買い価格の平均値を説明変数に加えた場合、加えなかった場合に比べモデルに基き計算した値と実際の約定価格の相関係数 (正確には Squared correlation coefficient を図に示す) は悪化している。我々はこの結果は「約定価格の単期間の挙動について、売り買い価格の平均値は、情報を持たない」のではなく、「データ表現形式として Time Series Representation は適切でない」と解釈し、より適したデータ表現形式として 2 種類の Event base Representation を設計した。

3 解析結果

3.1 データ表現、分析手法の詳細

本章では、2010 年 1 月から 2014 年 5 月まで 53ヵ月にわたる東京証券取引所 arrownet のデータ (2.9 Tera byte) から取り引き量の多い証券 100 銘柄のデータを抜き出し、図 2 Event base Representation を使って表現したものを一般的な教師なし学習手法を使って分析した結果を示す。

解析にあたっては libSVM[14] と weka の C4.5 実装である J45[15] を、それぞれのソフトの標準パラメータで利用した。被説明変数は、それぞれの約定価格が直前の約定価格から上昇したか、下降したか、変化なしかの 3 クラス問題とした。表現形式 1 では約定価格が未定のデータについては、直後の約定の価格を遡って用いた。これは元データである UDP パケットから見た場合、未来のデータを被説明変数として使用する事になるが、訓練時には既に過去データになっており、また、評価時には学習したモデルの精度評価の目的に用いるだけであるので、問題はない。

説明変数としては次のものを利用した。

板中の平均価格: arrownet の UDP パケットには、価格順に 8 価格までの売り買いの注文価格と数量が記憶されており、注文価格を数量を考慮して荷重平均した価格と最新の約定価格の差額を各パケット毎に計算し v とし、約定後、その UPD パケットまでの v を積算したものを説明変数 e_1 の値とした。

買い注文/売り注文の数: 約定後、その UDP packet を含めて買い注文／売り注文のパケット数をそれぞれ説明変数 e_2, e_3 とし、その差を e_4 とした。

最新の約定価格より安い注文／高い注文の数: 約定後、その UDP packet を含めて最新の約定価格より安い注文/高い注文のパケット数をそれぞれ説明変数 e_5, e_6 とし、その差を e_7 とした。

最新の約定価格より安い売り注文／高い買い注文の数: 約定後、その UDP packet を含めて最新の約定価格より安い売り注文／高い買い注文のパケット数をそれぞれ説明変数 e_8, e_9 とし、その差を e_{10} とした。

最新約定価格の買い注文/売り注文の数: 約定後、その UDP packet を含めて最新の約定価格での売り注文/買い注文のパケット数をそれぞれ説明変数 e_{11}, e_{12} とし、その差を e_{13} とした。

最新の約定価格と１つ前の約定価格の差: 最新の約定価格と１つ前の約定価格の差を説明変数 e_{14} とした。

3.2 板情報に基く基礎的分析結果

図 5 に訓練事例として証券毎に 10,000 注文ずつ学習データとして取り出し、引き続く 1,000 注文の中に含まれる約定価格が上るか/下るか/ 変化しないかを分析するモデルを学習した場合の平均精度を示す。

ZeroR は 10,000 注文の中で多かったケースが起こるとした場合の精度である。データ中、前回の約定価格と同額で取引が成立するケースはほぼなく、残りは半々

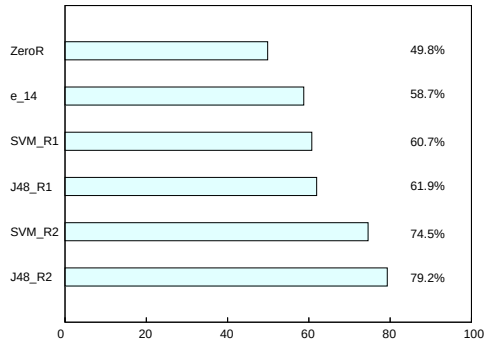

図 5: 板情報に基く基礎的分析結果

の確率で上るか下るかであった。従って ZeroR の結果は凡そ 50%(正確には 49.9%) であった。

e_14 は変数 e_{14}(最新の約定価格と 1 つ前の約定価格の差) のみ J45 の説明変数として入力した場合の結果 (58.7%) である。予備実験では 2 ～ 5 回前まで約定価格の差を使った結果も求めたが、今回利用した学習手法では精度向上が見られなかったので、以降の実験でも 1 つ前の約定価格のみ考える。ZeroR より改善は見られるものの、改善効果が大きいとは言えず、ランダムウォークを否定する結果としては弱い。

一方、表現形式 2 で説明変数として $e_1 \sim e_{13}$ を使い SVM, J45 で学習した結果が SVM_R2, J45_R2 である (それぞれ 74.6%, 79.3%)。J45_R2 の結果は、約定価格の情報を陽には含まない説明変数 $e_1 \sim e_{13}$ により、次の約定価格が 8 割近い精度で分類できる事を示している。表現形式 2 を実際の証券取引に使う事はできない。即ち表現形式 2 の評価用事例を入手できる時には新しい約定価格の情報も得る事ができるので、実務上の有用性はない。しかしながら、この実務上の有用性欠如は説明変数 $e_1 \sim e_{13}$ と約定価格の間の関係性を否定するものではない。板情報の使用を前提にすれば、価格の短期推移がランダムウォークとは見做せないと解釈するのが自然である。即ち、arrownet のような高速取引の過程では全取引参加者の間で情報が均等に伝わっておらず、情報の不均衡な伝搬状態により、短期的にランダムウォークとならなかったと解釈できる。

更に同じ説明変数を表現形式 1 を使い解析した結果を SVM_R1, J45_R1 として示す (それぞれ 60.7%, 61.9%)。J45_R2 の枠組みは、結果の精度や取引手数量の問題を別にして実取引にも利用できる予測の枠組みである (評価用事例に、実際にも約定価格の情報が未定のものを含み、未定の情報が将来実際にどうなるかで精度を評価している)。約定価格の情報を陽には含まない説明変数 $e_1 \sim e_{13}$ の方が過去の約定価格の推移 e_{14} よりも将来の約定価格の推移を精度良く推定できる事は興味深い。

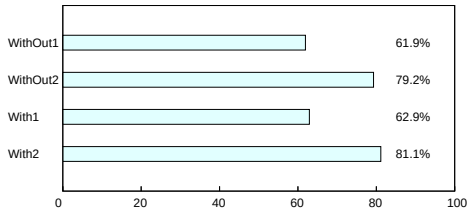

図 6: 取引履歴を使った分析精度の向上

図 3 および 4 に示した予備実験の結果では、板情報が約定価格の分析に役立つと言う結果は得られなかったが、図 5 の結果は、板情報が約定価格の分析に有用な情報を提供する事を示している。これは使用したデータ形式を従来標準的に用いられていた形式 (図 2 Traditional Time Series Representation) から注文が発生する毎にデータを記録する方法 (図 2 Event base Representation) に変更した事が主因である。

更に J45 の予測精度が SVM より高い事にも注意を要する。SVM は被説明変数の値を計算するための説明変数を利用した数学的な関数を求める枠組みであると言う意味で、ARMA モデルなど従来時系列解析で使われていた表現形式に近い。式を定義するパラメータの計算手法が違うだけとも見做せる。一方 J45 が使用している分類木表現は数学的な関数を求める枠組みではない。モデル化の枠組みとしては異る部分が多い。差の約 5%をどう見るか?SVM や同類の関数を使った手法の改良は可能か?など今後検討すべき課題も多いが、従来標準的に用いられていた時系列データの分析手法が証券市場の短期挙動の分析には適さない可能性には注意を要する。

3.3 分析精度の向上の試み

本研究の結果が実務に及ぼす影響を考察する上で、予測精度の向上は重要である。以下、幾つかの補助的な手法で予測精度の向上を試みた結果を報告する。

まず説明変数として $e_1 \sim e_{13}$ に加えて e_{14} を使った場合の結果を図 6 に示す。With_R1/2 が e_{14} を使った場合の表現形式 1 の結果、WithOut_R1/2 が比較のための e_{14} を使わなかった場合の結果 (図 5 の J45_R1/2 と同じもの) である。表現形式 1 の場合で 61.9%から 62.9%に、表現形式 2 の場合で 79.2%から 81.1%に、精度が向上している。

次に学習時に表現形式 2 と同じく売買成立のタイミングのみのデータを使い学習し、結果を全ての UDP パケット到着時のデータで評価した結果を図 7 Training に示す。Rep.1 と Rep.2 は比較のために示した図 6 With_R1/2 であり、学習時のデータを約定売買成立の

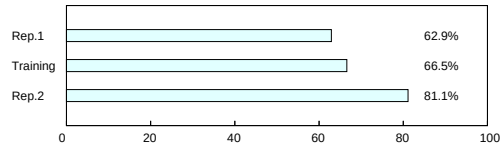

図 7: データ表現形式の違いを考慮した分析精度の向上

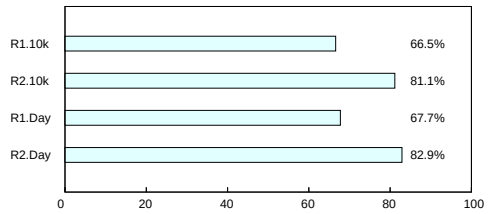

図 8: 処理単位を考慮した分析精度の向上

タイミングのみにする事で表現形式 1 の精度が 62.9%から 66.9%に向上している。結果の評価時に限らず学習時のデータのみを表現形式 1 のものから表現形式 2 に該当するものに制限する事で予測精度が向上するのは興味深い現象である。売買成立のタイミングでの情報の重要性を示唆していると考えるが、この現象の詳細な分析は今後の課題である。

図 8 は 10,000 データを使って学習し 1000 データを使って評価するかわりに、1 日分のデータを使って学習し翌日分のデータを使って評価した場合の精度を示したものである。表現形式 1 の場合は 66.9%から 67.7%に表現形式 2 の場合は 81.1%から 82.9%に精度が向上している。本研究の実験では、取り引き量の多い証券 100 銘柄について実験を行ったが、取り上げた 100 銘柄はいずれも 1 日の取引量が 10,000 より多い。学習に使うデータ数を増やす事で精度が向上したと考える。

4 本分析の本質に関する考察

図 9 に説明変数 $e_{1,4,7,10,13}$ を説明変数 e_{14} と組合せて 2 説明変数のみで約定価格を予測した場合の精度を示す。板に含まれる売り買い価格の平均値 e_1 も説明変数 e_{14} 単独の予測精度を向上するのに寄与するが、説明変数 e_{13} は説明変数 e_{14} と組合せて使用した場合、その他の説明変数全てを使った場合と同等の予測精度を達成した。

この事は最新約定価格での買い注文と売り注文の数の差と、最新の約定価格がその直前より上ったか下ったかを観察していれば、次に約定価格が上るか/下るかを高い精度で予測できる事を意味している (図 10)。この事は arrownet を使った高速取引では約定価格の挙動

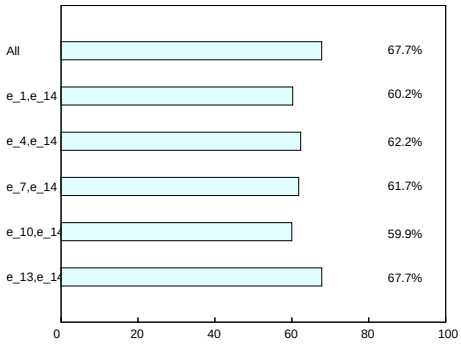

図 9: 歩み値での入札と歩み値変化の関係

は効率的市場仮説が示唆するランダムウォークとは見做せない事を意味するだけではない。高速取引に real time で参加可能な機関投資家が、設備を持たず高速取引に参加できない一般の市場取引者に対して有利な場面がある事を意味し、公平性の観点から検討を要する事を示唆している。

5 結論

本研究では板情報として 2010 年 1 月から 2014 年 5 月まで 53ヵ月にわたる東京証券取引所の取引データを用い、

- 板情報を用いて標準的な教師なし学習の枠組みの中で短期の価格推移予測において 82.9%の予測精度をもったモデルを構築できる事を示し、
- 効率的市場仮説が示唆するランダムウォークとは見做せない事を報告し、
- その原因が情報伝搬の不均一性にある事を議論した。

上記価格予測モデルは直接的に証券取引に利用できるものではないが、板情報が証券価格の形成を分析するにあたり重要な情報源となりえる事、従来標準的に用いられていた時系列データの分析手法が証券市場の短期挙動の分析には適さない事などを示唆している。

また、証券取引可能な形でのモデル化も試み 67.7%の予測精度も持つ事示した。この結果は、高速取引に real time で参加可能な機関投資家が、設備を持たず高速取引に参加できない一般の市場取引者に対して有利な場面がある事を意味し、公平性の観点から検討を要する事を示唆している。

```
J48 pruned tree
------------------

e14 <= -1
|   e13 <= -1: n (111.0/24.0)
|   e13 > -1: p (878.0/166.0)
e14 > -1
|   e13 <= 0: n (845.0/135.0)
|   e13 > 0: p (125.0/26.0)

Number of Leaves :      4
Size of the tree :      7

Time taken to build model: 0.14 seconds
Time taken to test model on training data: 0.11 seconds

=== Error on training data ===

Correctly Classified Instances      1608     82.0827 %
Incorrectly Classified Instances     351     17.9173 %
Kappa statistic                        0.6417
Mean absolute error                    0.1956
Root mean squared error                0.3128
Relative absolute error               58.6688 %
Root relative squared error           76.615  %
Coverage of cases (0.95 level)       100      %
Mean rel. region size (0.95 level)    66.6667 %
Total Number of Instances           1959

=== Error on test data ===

Correctly Classified Instances       942     81.9843 %
Incorrectly Classified Instances     207     18.0157 %
Kappa statistic                        0.6393
Mean absolute error                    0.1978
Root mean squared error                0.3145
Relative absolute error               59.3132 %
Root relative squared error           77.014  %
Coverage of cases (0.95 level)       100      %
Mean rel. region size (0.95 level)    66.6667 %
Total Number of Instances           1149
```

図 10: weka の学習例

参考文献

[1] S. Basu, “Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis,” The Journal of Finance, vol.32, no.3, pp.663–682, 1977.

[2] B.G. Malkiel, A random walk down Wall Street: including a life-cycle guide to personal investing, WW Norton & Company, 1999.

[3] L.A. Gallagher and M.P. Taylor, “Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks,” Southern Economic Journal, pp.345–362, 2002.

[4] E. Gilbert and K. Karahalios, “Widespread worry and the stock market.,” ICWSM, pp.59–65, 2010.

[5] R.D. Gay Jr *et al.*, “Effect of macroeconomic variables on stock market returns for four emerging economies: Brazil, russia, india, and china,” International Business & Economics Research Journal (IBER), vol.7, no.3, 2011.

[6] P.H. Hsu, Y.C. Hsu, and C.M. Kuan, “Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias,” Journal of Empirical Finance, vol.17, no.3, pp.471–484, 2010.

[7] L. Menkhoff, “The use of technical analysis by fund managers: International evidence,” Journal of Banking & Finance, vol.34, no.11, pp.2573–2586, 2010.

[8] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” Journal of Computational Science, vol.2, no.1, pp.1–8, 2011.

[9] T.O. Sprenger, A. Tumasjan, P.G. Sandner, and I.M. Welpe, “Tweets and trades: The information content of stock microblogs,” European Financial Management, 2013.

[10] H. Sul, A.R. Dennis, and L.I. Yuan, “Trading on twitter: The financial information content of emotion in social media,” System Sciences (HICSS), 2014 47th Hawaii International Conference on, pp.806–815, IEEE, 2014.

[11] M. Chlistalla, B. Speyer, S. Kaiser, and T. Mayer, “High-frequency trading,” Deutsche Bank Research, pp.1–19, 2011.

[12] 東京証券取引所, “arrownet,” 2012. [accessed 19-Aug-2014].

[13] J.D. Hamilton, Time Series Analysis, Princeton university press, 1994.

[14] C.C. Chang and C.J. Lin, “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol.2, pp.27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, “The WEKA Data Mining Software: An Update,” SIGKDD Explorations, vol.11, pp.10–18, 2009. [accessed 19-Aug-2014].