

データマイニングによる投資支援システムの検討

事業推進室

松井 武史

1. はじめに

昨今、インターネットの普及、クラウド技術や SNS 技術の発展によって、様々なデータが以前よりはるかに大量かつ迅速に収集できる環境にある。

こういった環境を背景として、データマイニング技術が注目されつつある。

本稿では、インターネット上で容易に収集できるデータを用いてマイニングを行い、投資支援を行うシステムについて検討を行った。

2. データマイニングとは

データマイニングとは、統計学等のデータ解析技術を用いて、大量のデータの山から、それまで認識されていない、有益な知識や情報を取り出すことを言う。「マイニング」は「採掘」の意。データウェアハウスなどに蓄積された膨大なデータを「鉱山」に見立て、そこから未知の知識や情報という“鉱石”を「発掘する」という意味が込められている。

3. 支援システムについて

株価等の相場予測方法としては、世の中に様々なものが出回っており、ソフトウェアも数多く存在するが、本稿では、敢えて相場予測のテクニックを使わずに、データ解析のみによって、投資判断を支援するシステムを構築し、実際の投資シミュレーションを行う。

3.1 投資対象

投資対象はいろいろな金融指標に影響を受けていると考えられる日経 225 先物取引とした。日経 225 先物取引とは、日経平均株価を原資産とする株価指数先物取引であり、大阪証券取引所等に上場されている。

3.2 売買のルール

本稿における投資支援システムにおける売買ルールは以下の通りとする。

- ① 取引は日中(9:00~15:15)のみとする。
- ② 支援システムの結果より、AM9:00 時点で売り買いを行う。ただし、売り/買い両方のシグナルが出た場合は、売り/買い両方の取引を行い、どちらのシグナルも出なかった場合は、取引を行わないものとする。
- ③ 買いのポジションを取った場合、購入金額より一定割合(0.8%)以上上昇した場合に売る。売りポジションの場合は一定割合以上下落した場合に買い戻す。
- ④ 一定割合(0.8%)以上の変動が無かった場合は終値で精算する。

3.3 支援システムの概要

インターネット上で入手できる時系列の金融指標データを基に日毎に、始値から 0.8%以上上昇するか、0.8%以上下落するかの分類を行うモデルを機械学習(後述)により構築する。

構築したモデルに、対象となる日のデータを入力してその日の売り買いを判断する。

3. 3 使用する指標

使用する指標は、表 1 に示すとおり。また、それぞれに移動平均、指数平滑移動平均の変化率を日毎に求め、予測のためのパラメータとする。

移動平均:

対象となる日から一定期間の平均値。本稿では 20 日間の移動平均を用いている。

指数平滑移動平均:

直近の数値に比重を置いて過去になるほど比重を軽くして平均化した値。以下の式で表される。

指数平滑移動平均=前日の指数平滑移動平均
+ $\alpha \times \{\text{当日終値} - \text{前日の指数平滑移動平均}\}$

表1 使用する指標

円/ドル、円/ユーロ、NY 金、NY 原油 米国 10 年金利、NY ダウ(米)、NASDAQ(米)、 FTSE100(英)、上海総合(中)、HANSEN(香)、 CME(シカゴ・表品取引所)日経 225 先物 大証日経 225 先物日中取引 大証日経 225 先物ナイトセッション
--

4. 支援システムの構築

支援システムは、汎用的な統計解析ツール R に組み込まれている機械学習関数のアルゴリズムを用いて、過去データを基に学習し、日毎の売り買いを判定するモデルを構築する。また、検証用データを用いて再現性の評価を行う。

4. 1 統計解析ツール R

支援システムの構築には、オープンソースの統計解析ツール R を用いた。

R は R-project(<http://www.r-project.org/>)という国際共同プロジェクトで開発されており、誰でも無料

で自由に利用できるように、研究機関等を中心に広く普及している統計ソフトである。

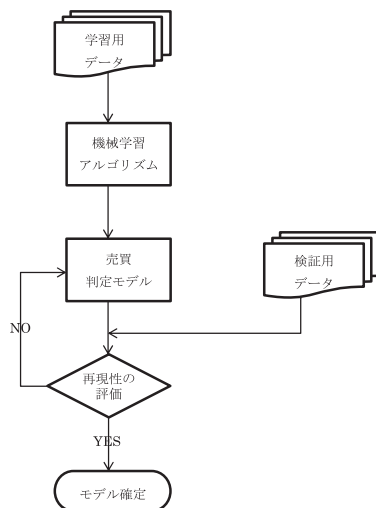


図1 システム構築のフロー

4. 2 機械学習について

機械学習とは、人工知能における研究課題の一つで、人間が自然に行っている学習能力と同様の機能をコンピュータで実現しようとする技術・手法のことで、センサやデータベースなどからある程度の数のサンプルデータ集合を入力して解析を行い、そのデータから有用な規則、ルール、知識表現、判断基準などを抽出し、アルゴリズムを発展させる手法である。

機械学習の技法は数多く提案されており、それぞれ特徴を持っている。(得手不得手がある。)

本稿では、本稿で作成したデータと比較的相性のよかった(良好な結果が得られた)、以下の技法を用いた。

(1) ニューラルネットワーク(NN)

生物の神経ネットワークの構造と機能を模倣するという観点から生まれた学習アルゴリズム

である。人工神経を相互接続したもので計算を構造化し、コネクショニズムの計算技法で情報を処理する。現代的ニューラルネットワークは非線形な統計的データモデリングツールである。入力と出力の間の複雑な関係をモデル化するのに使われ、データのパターン認識や観測された変数間の未知の同時分布における統計的構造を捉えるなどの用途がある。

(2) ランダムフォレスト (RF)

決定木を弱学習器とする集団学習アルゴリズムであり、この名称は、ランダムサンプリングされたトレーニングデータによって学習した多数の決定木を使用することによる。対象によっては、同じく集団学習を用いるブースティングよりも有効とされる。

(3) サポートベクターマシン (SVM)

統計分類や回帰分析に使われる一連の教師あり学習技法である。訓練例のラベルは 2 値 (2 つに分類される) であり、訓練アルゴリズムによってモデルを構築し、新たな例がどちらに分類されるかを予測する。

個々の技法のアルゴリズムについては紙面の制約により割愛した。それぞれの技法の詳細については、文末の参考文献を参照されたい。

4. 3 システムの構築

表 1 で示した各指標の 2011/1/1～2013/6/30 のデータをベースデータとして、売りポジション用データ、買いポジション用データをそれぞれ作成する。売りポジション用データはその日の安値が始値を 0.8% 下回る場合、“SHORT”(売り)のフラグ立てる。また、買いポジション用データはその日の高

値が始値を 0.8% 上回る場合、“LONG”(買い)のフラグを立てる。このように売り、買いのデータを分割するのは 1 日の変動の中で、安値が始値より 0.8% 安くなる場合と、高値が始値より 0.8% 高くなる場合が同時に現れる日があるためである。

売りポジション用データ及び買いポジション用データをそれぞれ 2 群に分割し、一方を学習用データ、もう一方を検証用データとして、前述の 3 通りの機械学習技法を用いて、図 1 に示すフローにしたがって、技法別・売り買い別にシステムを構築した。

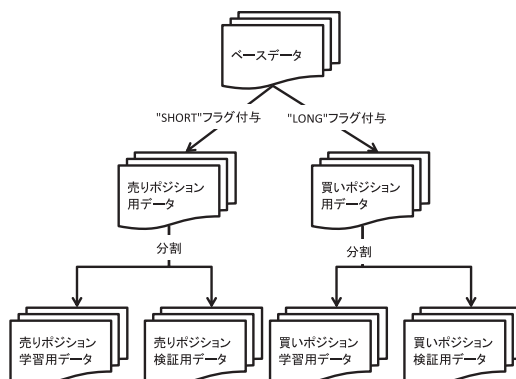


図 2 データの関連性

(1) ニューラルネットワーク (NN)

ア) NN の概要

NN の関数として、R のパッケージ nnet の関数 nnet を利用した。nnet は、単一中間層の階層型ニューラルネットワークの関数である。

本稿で使用した NN の特徴としては、パターン認識・分類、ノイズが混在しているデータの処理が得意である反面、計算時間がかかる、乱数で決定される初期値の影響を受けやすく、解が安定しない等の欠点がある。(本稿では、パラメータチューニングのために初期の乱数を固定して検討した。)

イ) システムの再現性

売り、買いそれぞれのシステムによるデータの検証結果を表 2 に示す。行方向は実データから計算された売り買いのフラグ、列方向はシステムで計算された売り買いのフラグを集計している。表 2(1) では、実データでは売りフラグが 81 個、フラグ無しが 214 個、計算から求められた売りフラグが 56 個、フラグ無しが 239 個、実データでも計算結果でも売りフラグとなっているものが 31 個、いずれもフラグ無しとなっているものが 189 個あることを示している。また、正答率とは、実データのフラグと計算結果のフラグが一致する確率を示している。売り/買いいずれのケースについても 75% 程度の一致が見られる。

表 2(1) 売りポジション用システムの再現結果

	売	無	合計
売	31	50	81
無	25	189	214
合計	56	239	295
正答率	74.6%		

表 2(2) 買いポジション用システムの再現結果

	買	無	合計
買	41	38	79
無	34	182	216
合計	75	220	295
正答率	75.6%		

(2) ランダムフォレスト(RF)

ア) RF の概要

RF の関数として、R のパッケージ randomForest の関数 randomForest を利用した。

RF の特徴としては、高次元のデータ解析が得意であり、計算速度が速いことが挙げられる。ただし、比較的新しい手法であるため、参考書籍等が少ないことが欠点である。

イ) システムの再現性

売り、買いそれぞれのシステムによるデータの検

証結果を表 3 に示す。売り/買いいずれのケースについても 75% 以上の一致が見られる。

表 3(1) 売りポジション用システムの再現結果

	売	無	合計
売	22	59	81
無	11	203	214
合計	33	262	295
正答率	76.3%		

表 3(2) 買いポジション用システムの再現結果

	買	無	合計
買	22	57	79
無	8	208	216
合計	30	265	295
正答率	78.0%		

(3) サポートベクターマシン(SVM)

ア) SVM の概要

SVM の関数として、R のパッケージ kernlab の関数 ksvm を利用した。

SVM の特徴としては、高次元の分類問題が得意であり、計算速度が速いことが挙げられる。ただし、今回用いた関数は二値分類の技法であり、3 以上のカテゴリのデータを分類することはできない。

イ) システムの再現性

売り、買いそれぞれのシステムによるデータの検証結果を表 4 に示す。売り/買いいずれのケースについても 80% 近くの一致が見られる。

表 4(1) 売りポジション用システムの再現結果

	売	無	合計
売	33	48	81
無	19	195	214
合計	52	243	295
正答率	77.3%		

表 4(2) 買いポジション用システムの再現結果

	買	無	合計
買	36	43	79
無	19	197	216
合計	55	240	295
正答率	79.0%		

(4) 技法間の比較

システム構築を行った 3 種類の技法に関して行った比較結果を以下に示す。

① 正答率

正答率に関しては、いずれも 75%~80%の範囲に収まっており、大きな差は見られなかったものの、NN<RF<SVM の順になっている。

② 売り/買いシグナル検出数

投資支援システムにおいては、データから売り/買いのシグナルをできるだけ多く、正確に検出できることが重要となる。

売り/買いのシグナル検出数については、各表の左上の数値が大きいものほど、数多く検出できていることになる。

売りで見た場合は、RF<NN<SVM、

買いで見た場合は、RF<SVM<NN

の順になる。

③ シグナル誤検出率

シグナルが出ていない日に、シグナルを検出してしまうと投資の失敗に繋がる。投資を成功に導くためには、売り/買いシグナルを正確に検出する必要がある。それぞれの誤検出率は、以下の通りとなる。誤検出率は売り/買いともに、RF<SVM<NN となる。

NN

売り:34/75=45.3%、買い:31/56=44.6%

RF

売り:11/33=33.3%、買い:8/30=26.7%

SVM

売り:19/52=34.5%、買い:19/55=36.5%

これらの要素を総合して評価すると、本稿においては SVM が総合的に高精度で予測できていると評価できる。

5. 投資シミュレーション

5.1 シミュレーションの条件

前章で構築した支援システムを用いて、3.2 で示したルールに従い、2013/7/1~2013/7/31 の1ヶ月間の投資シミュレーションを行った。

5.2 シミュレーションの結果

シミュレーションの結果を表5に示す。

左から、日付、技法毎のシグナル、当日の日経225先物の始値、高値、安値、終値、売り/買いで精算する閾値、技法毎の収支を表している。閾値に関しては日経 225 先物ラージ¹の刻み幅に合わせて10円単位で丸めている。

システムで実データのシグナルと比較すると、売りは比較的うまくシグナルをとらえられているが、買いでは、実データで13回のシグナルが出ているにもかかわらず、比較的多くとらえている RF でも5回(うち1回は間違い)、SVM は1回、NN に至ってはシグナルを1度もとらえることができていない。

売りでは NN の7/4、7/23、7/24、買いでは RF の7/29 で予想が外れマイナスとなっているが、その他のシグナル発現時は全て実データと一致しており、検証時よりも誤検出が非常に少なくなっていた。

最終的な結果は、NN で+480、RF で+930、SVM で+750 となっており、日経 225 先物ラージで取引したとすると、取引単位が指数の1000倍となるので、48万~93万の利益が得られたことになる。

対象期間においては、誤検出率が NN で3/10=30%、RF で1/12=8.3%、SVM で0/7=0%と非常に低かったが、検証時の結果から、誤検出は35%程度発生すると思われるため、実際にはこれほどの利益は見込めないと思われる。

¹ 大阪証券取引所の日経 225 先物はラージとミニが上場されており、取引単位はラージが呼値の刻みが10円、指数の1000倍、ミニが呼値の刻みが5円、指数の100倍。

表 5(1) シミュレーション結果(売りポジション)

日付	NN	RF	SVM	実データ	OPEN	HIGH	LOW	CLOSE	精算額	NN	RF	SVM
7月1日	無	無	無	売	13760	13880	13560	13870	13600	—	—	—
7月2日	無	無	無	無	13960	14140	13900	14130	13800	—	—	—
7月3日	売	無	無	売	14170	14180	13970	14020	14100	70	—	—
7月4日	売	無	無	無	13980	14100	13950	13990	13900	-10	—	—
7月5日	無	無	無	無	14170	14350	14150	14340	14100	—	—	—
7月8日	売	売	売	売	14480	14510	14060	14100	14400	80	80	80
7月9日	無	無	無	無	14290	14500	14180	14460	14200	—	—	—
7月10日	売	無	売	無	14450	14570	14270	14430	14300	150	—	—
7月11日	無	無	無	無	14270	14510	14260	14470	14200	—	—	—
7月12日	無	無	無	無	14460	14590	14410	14470	14300	—	—	—
7月16日	無	無	無	無	14610	14660	14560	14600	14500	—	—	—
7月17日	無	無	無	無	14500	14620	14450	14600	14400	—	—	—
7月18日	無	無	無	無	14660	14840	14650	14800	14500	—	—	—
7月19日	無	売	売	無	14920	14970	14410	14650	14800	—	120	120
7月22日	売	売	売	売	14770	14790	14520	14690	14700	70	70	70
7月23日	無	無	無	無	14580	14830	14540	14780	14500	-200	—	—
7月24日	無	無	無	無	14720	14770	14620	14730	14600	-10	—	—
7月25日	無	売	売	無	14730	14760	14520	14520	14600	—	130	130
7月26日	無	売	売	売	14340	14380	14100	14130	14200	140	140	140
7月29日	売	売	売	売	13870	13960	13600	13600	13800	70	70	70
7月30日	無	無	無	無	13640	13920	13610	13890	13500	—	—	—
7月31日	売	売	売	売	13720	13840	13600	13600	13600	120	120	120
トータル収支										480	730	730

表 5(2) シミュレーション結果(買いポジション)

日付	NN	RF	SVM	実データ	OPEN	HIGH	LOW	CLOSE	精算額	NN	RF	SVM
7月1日	無	買	無	買	13760	13880	13560	13870	13900	—	140	—
7月2日	無	無	買	無	13960	14140	13900	14130	14100	—	—	—
7月3日	無	無	無	無	14170	14180	13970	14020	14300	—	—	—
7月4日	無	無	買	無	13980	14100	13950	13990	14100	—	—	—
7月5日	無	買	無	買	14170	14350	14150	14340	14300	—	130	—
7月8日	無	無	無	無	14480	14510	14060	14100	14600	—	—	—
7月9日	無	無	買	無	14290	14500	14180	14460	14400	—	—	—
7月10日	無	無	買	無	14450	14570	14270	14430	14600	—	—	—
7月11日	無	無	買	無	14270	14510	14260	14470	14400	—	—	—
7月12日	無	無	買	無	14460	14590	14410	14470	14600	—	—	—
7月16日	無	無	無	無	14610	14660	14560	14600	14700	—	—	—
7月17日	無	無	買	無	14500	14620	14450	14600	14600	—	—	—
7月18日	無	買	無	買	14660	14840	14650	14800	14800	—	140	—
7月19日	無	無	無	無	14920	14970	14410	14650	15000	—	—	—
7月22日	無	無	無	無	14770	14790	14520	14690	14900	—	—	—
7月23日	無	無	買	無	14580	14830	14540	14780	14700	—	—	—
7月24日	無	無	無	無	14720	14770	14620	14730	14800	—	—	—
7月25日	無	無	無	無	14730	14760	14520	14520	14800	—	—	—
7月26日	無	無	無	無	14340	14380	14100	14130	14500	—	—	—
7月29日	無	買	無	無	13870	13960	13600	13600	14000	—	-270	—
7月30日	無	買	買	買	13640	13920	13610	13890	13700	—	60	60
7月31日	無	無	買	買	13720	13840	13600	13600	13800	—	—	—
トータル収支										0	200	60

6. システムの評価と今後の課題

本稿のシミュレーションでは、買いシグナルの検出率が低かったものの、誤検出率が非常に低く、結果的に大きな利益が出る結果となったが、システムの検証結果からも長期間では誤判定率は35%程度になると推測でき、収支は今回の結果よりは悪化すると思われる。しかし、同時に今後の可能性も見いだすことができた。

今後は、

- ・適切なパラメータの絞り込み
- ・機械学習のチューニング
- ・予測が外れたときの対応(損切り)

について検討が必要であると思われる。

本システムの特徴としては、単純にデータのみに基づいて予測を行い、政治/経済情勢等を一切

考慮しない点である。しかし、逆にそれらの要因はデータの中に既に組み込まれているとも言える。

データから新たな知識を掘り起こすデータマイニングの技術は今後幅広い分野に適用されることが期待される。

<謝辞>

本稿の執筆にあたり、多大なるご協力を頂いた株式会社空間情報の矢野公一氏に深く感謝いたします。

<参考文献>

- 1) 「R によるデータサイエンス」(金明哲, 2007 年 10 月, 森北出版)