IFTA 講演・寄稿論文

AI とマシンラナニング (機械学習) による株式投資戦略

野村アセットマネジメント 資産運用先端技術研究部 リサーチフェロー 中川 慧

要旨

AI で株式投資戦略を立てる方法を説明する。予測方法として、(1) テクニカル分析の理念である株価変動パターンに焦点を当てる、(2) 従来のクオンツ分析の基本となる要因に焦点を当てる、の2種類について紹介する。

もともと、AI と機械学習は異なる概念だったが、最近は同様に用いられている。機械学習は 人間の脳の機能を模倣するアルゴリズムで、一般的には①教師なし学習、②強化学習、および ③教師あり学習の3種類がある。学習とは、データをうまく説明できるモデルを見つけること を意味する。

将来の株価を予測するには、画像認識や翻訳とは異なり、どのデータを抽出するかが重要になる。価格や量などの市場データ(テクニカルデータ)と、GDP や利益などの経済・金融データ(ファンダメンタル)であり、猫の画像や「This is a pen.」の英文(翻訳)と違って、事前に特徴がわからない。

予測の手法は「時系列予測」と「断面予測」の2種類がある。時系列予測は、単一の株式に焦点を当て、履歴データから将来のリターンを予測する。一方、断面予測はすべての株価に焦点を当て、いくつかの基準で高いスコア(=データと株価が連動する)の株式を検索する。

時系列予測は絶対リターンを予測することが目的で、投資期間は短期 (月単位より短い) の場合が多い。断面予測では相対的なリターンを予測し、投資期間は長期 (月単位より長い) の場合が多い。

テクニカルデータと教師あり学習による 時系列予測

機械学習による形成解析:過去の価格変動パターンを用いて、将来の価格を予測する。パターンはモメンタム(勢い)や類似の値動きを使う。 投資家も実際に用いる手法である。

株価変動が現在に最も近い過去の期間を抽出するために、インデックス DTW (IDTW) 方式を用いる。そして、将来の株価予測には k* 近傍法(k*-NN) アルゴリズムを使用する(図 1)。

IDTW:類似性測定のために、次の3点を決め

る。(1) 測定期間と頻度:投資家にとっての基本 単位である1カ月で、日足で見る。(2) 価格水準 の違いを考慮する:月末値を1として指数化する。 (3) 値動きの類似性を定義する:Indexation DTW (動的時間ワープ) 法を使う。時間軸をゆがめるこ とで2つの時系列と最適な対応を行う(月によって営業日数が異なることにも対応できる)。類似 性を定義する場合、通常はユークリッド距離また は相関係数を使用する。

k*-NN:解析の対象データに一切の分布を仮定しないノンパラメトリック手法に最も基本的なアルゴリズムである k-NN を最適化したもの。

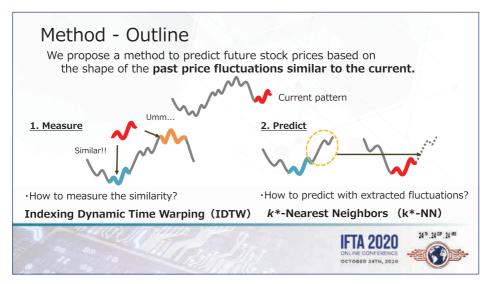


図 1. 時系列予測の研究例

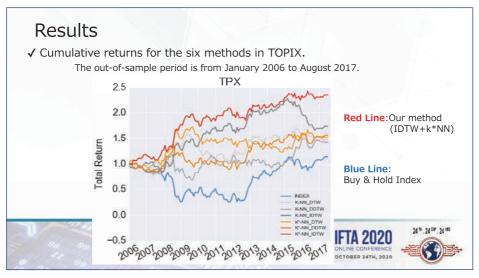


図 2. TOPIX の累積リターン

【実証研究】IDTW と k^* -NN を用いて、5 つの世界の主要な株価指数 (TOPIX、S&P500、CAC40、DAX および FTSE100) のパフォーマンスを計測する。データ期間は 1989 年 1 月~2017 年 8 月 (DDTW は DTW の修正版)。作業は 4 ステップで、(1) 月ごとの類似性を計算する。(2) 翌月のリターンを予測する。(3) 翌月リターン予測がプラスの場合、当月末に指数の 1 単位を購入する。(4) これらを全期間 (の翌月まで) で繰り返す。

【結果】TOPIX について、類似性分析と予測を 6 手法で行い、累積リターンを示している(図 2)。 われわれの提案した IDTW+k*-NN(赤線)が最もパフォーマンスがよかった(青線は実際の指数のパ

フォーマンス)。ほかの株価指数でも同様である。

表 1 (Table 1) は、すべての年の平均精度を示す。われわれの手法 (太字赤枠) は最も正確だった。表 1 (Table 2) は累積リターンで、われわれの手法 (太字赤枠) は FTSE100 を除いて最も収益性が高かった。

単純なモメンタム戦略と比較した結果(表 2)で も、われわれの手法のパフォーマンスが優れており、 また単純戦略とは相関が低いことがわかる。

【要約】機械学習では多くのパラメーターが必要だが、k*NN ならパラメーターは 1 つでよい。投資家が実際に株価を予測する手法を近似でき合理的である。そして、われわれの手法がベストである。

表 1. 平均精度と合計リターン

Results

Table 1. The average accuracy of all years for each method.

The state of the s		CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
DTW	k-NN	46.76%	51.80%	46.04%	52.52%	52.52%	49.93%
	k^+ -NN	52.52%	53.96%	50.36%	47.48%	50.36%	50.96%
	k-NN	51.80%	53.24%	51.80%	60.43%	51.08%	53.67%
	k*-NN	50.36%	56.12%	54.68%	58.99%	55.40%	55.11%
IDTW	k-NN	49.64%	53.24%	55.40%	61.87%	55.40%	55,11%
	k*-NN	57.55%	59.71%	57.55%	66.91%	60.43%	60.43%

Table 2. The total return of each method.

		CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
DTW	k-NN	0.98%	146.03%	34.32%	148.15%	146.79%	95.25%
	k*-NN	122.22%	164.03%	66.98%	97.12%	152.67%	120.60%
DDTW	k-NN	157.14%	74.38%	98,21%	185.76%	142.32%	131.56%
			154.90%				
IDTW	k-NN	126.59%	124.26%	142.10%	212.87%	174.20%	156.00%
	k*-NN	222.24%	212.91%	140.74%	235.29%	234.53%	209.14%



表 2. 単純なモメンタムとの比較

Results

Here, Prediction in Step 2 is replaced by **simple 1Mom**: 1 month return and **12 Mom**: 12 month return excluding the last month (12 Mom).

Table 3. Comparison with Momentum Strategy(Return)

Method	CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
IDTW+k*-NN	222.24%	212.91%	140.74%	235.29%	234.53%	209.14%
1Mom	129.70%	120.53%	33.52%	135.23%	186.36%	121.07%
12-1Mom	110.58%	127.34%	135.90%	201.64%	124.96%	140.08%

Table 4. Correlation with Momentum Strategy

Method	CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
1Mom	0.09	0.02	-0.01	0.07	-0.02	0.03
12-1Mom	0.05	-0.15	0.06	-0.14	0.07	-0.02





テクニカルデータと教師なし学習による 時系列予測

過去の類似の価格変動パターンが将来の株価予測に有効とわかったので、クラスタリング(データ分類)による予測に向けた価格変動パターンの可視化に取り組んでみる。

クラスタリングには通常、k-means 法(k 平均法)を使うが、DTWでは使用できない。そこで、k-medoids 法を用いる。k-means 法が単純距離でクラスタリングするのに対し、k-medoids はデータの類似性を重視する。類似性は IDTW で測定される。

【実証研究】TOPIXを用いて予測に適したパターンを可視化する。月次価格変動に基づいてk-medoidを実行。当月の価格変動を含むクラスター内に正のラベルがある場合は、月末にインデックス1単位を購入する。

【結果】表3は合計リターンと平均精度を示す。 収益性が最適なクラスター数は5個程度と言える。

5つのクラスターを可視化する。黒い線は medoid として選択された価格変動を、赤い線 はクラスターに属する価格変動を示している。 Sample の数字はクラスター内のサンプルの数、()内は翌月のリターンの上昇確率である。左図 は下降傾向だが、翌月の上昇確率は高く、反転上

Results ✓ From the viewpoint of profitability, the number of k is 5 Table 5. Profitability and accuracy INDEX by number of clusters k*-NN Total Return[%] AR IDTW DTW IDTW DTW DTW(k=11) Cumulative Return 1.0 IDTW(k=5) k = 2113 53.23 51.61 k = 398 57.26 53.23 k = 4148 66 61.29 k = 5162 64 63.71 54.03 151 64.52 54.03 k = 7131 61.29 51.61 k = 8132 62.1 53.23 k = 9147 45 62.9 54.84 k = 1058.06 114 41 55 65 k = 11104 79 58.87 59.68 k = 12

表 3. 平均精度と合計リターン

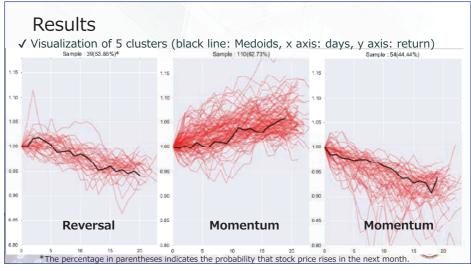


図3.5つのクラスターの可視化①

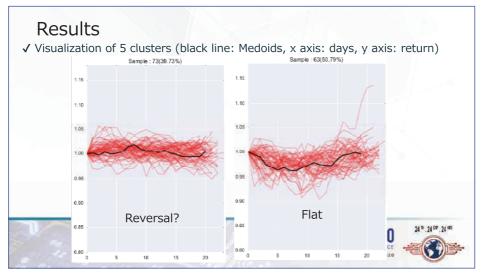


図 4.5 つのクラスターの可視化①

昇を示す。中央の図は上昇の勢いが強く、翌月の 上昇確率も高い。中間図では、価格の上昇が強く、 次の月の上昇確率も高い。右図は翌月の下落確率 が高く、下落幅も大きいと予測している。

図3は中央の黒い線が平坦で、翌月の下落確率 が高い。図4では変動はほぼフラットで、翌月も 持続しそうだ。

【要約】価格変動パターンのクラスター数は TOPIX では約5であり、可視化すると大きな変動 が続く傾向にあることがわかった。

テクニカル&ファンダメンタルデータと 教師あり学習による時系列予測

次に、株価予測のために時系列決定木 (TSDT) を改善する。時系列データのみを使用したディシジョンツリーで、理解しやすく便利である。アンサンブル学習の一種である勾配ブースティングを用いてより正確なモデルを構築する。

【実証分析】5つの世界の主要な株価指数(TOPIX、S&P500、CAC40、DAX および FTSE100)を用いて、そのパフォーマンスを分析する。計測期間は 2018 年6月までの10年間。時系列データに1カ月間の日足を、ファンダメンタルデータとして各指数の月末値ベースの PER、PBR、モメンタム、配当利回りおよびROEを使用した。時系列決定木(TSDT)と時系列勾配ブースティングツリー(TSGBT)についてファンダ

メンタルデータの有無で計4手法を比較した。

【結果】表 4 (Table 6) は平均精度を示している。右端の列は各手法の単純平均である。太字は4手法の中で最も正確な値を示しており、ファンダメンタルデータを持つ TSGBT が最も正確だった。表 4 (Table 7) は4手法の合計リターンを示している。やはりファンダメンタルデータを持つTSGBT が最も収益性が高かった。

【要約】ファンダメンタルデータを追加すること で精度と収益性が向上した。

テクニカル&ファンダメンタルデータと 教師あり学習による断面予測

これまで時系列予測の研究例を見てきたが、次 に断面予測の研究事例を見る。

さまざまなファクターを組み合わせることで、 将来の相対的な株式リターンを予測するので、横 断的な投資戦略と呼ばれている。ディープラーニ ング(深層学習)で株価を説明する多数のファク ターの非線形関係を捉えたい(図 5)。

【実証分析 1】2006 年 4 月から 2016 年 3 月までの TOPIX について、80 ファクターを使用して相対的なリターンを予測し、毎月投資する。リスク、クオリティー、モメンタム、バリュー、サイズの 5 分類からなる 16 のファクターを過去 5 時点分 (16×5=80) 使用した。

Results

Table 6. The average accuracy of all years for each method.

表 4. 主要株価指数の平均精度と合計リターン

ACC	TOPIX	SP500	DAX	CAC	FTSE	avg
TSDT	50.00%	47.73%	51.81%	45.24%	53.47%	49.65%
TSDT+CS	64.04%	61.82%	56.63%	63.10%	61.39%	61.39%
TSGBT	57.87%	61.36%	59.04%	54.76%	57.43%	58.09%
TSGBT+CS	69.10%	65.91%	62.65%	64.29%	66.34%	65.66%

Improvement by adding features

Improvement by boosting

Table 7. The total return for each method.

Total Return	TOPIX	SP500	DAX	CAC	FTSE	avg
TSDT	34.99	10.86	-6.16	0.15	35.68	15.11
TSDT+CS	255.84	198.36	69.20	117.28	67.45	141.62
TSGBT	111.51	88.14	112.78	86.72	75.21	94.87
TSGBT+CS	229.79	224.30	100.93	140.42	120.73	163.23



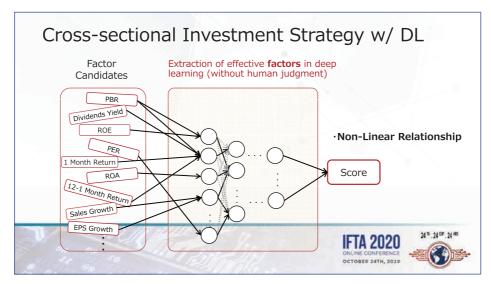


図 5. 断面予測の研究事例

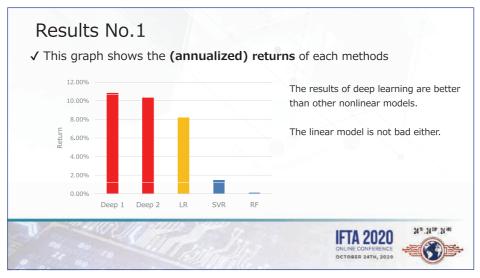


図 6. 各手法の年率リターン (TOPIX)

線形回帰は、線形モデルベンチマーク、SVR およびランダムフォレストを非線形モデルベンチマークとして使用する。5分位の最上位群の株式を買い、最下位群の株式を売る。ネットゼロ投資のためのロング/ショート・ポートフォリオを作り、年間平均リターンとシャープレシオを計算してパフォーマンスを測定する。

【結果】各手法の年次リターンである。ディープラーニングの結果は、他の非線形モデルよりも優れている。また、線形モデルも悪くない(図6)。

以下は各方法のシャープレシオを示している。 やはりディープラーニングは他のモデルよりも優れている。重要なのは線形モデルが不良な非線形 モデルよりも優れていることだ(図7)。

【実証分析 2】日足ベースで 2013 年 4 月から 2017 年 3 月までの TOPIX500 指数について、33 因子を用いて相対的なリターンを予測する。リターンや取引量などの多くのテクニカル指標がある。

線形モデルのベンチマークとして3つのリッジ回帰を使用し、3つのランダムフォレストを非線形モデルベンチマークとして使用する。ディープラーニングでは、6種類の完全接続ネットワークを使用している。5分位のロング/ショート・ポートフォリオを作成し、年間リターンとシャープレシオを計算する。

【結果】各手法の年次リターンを見ると、線形モ

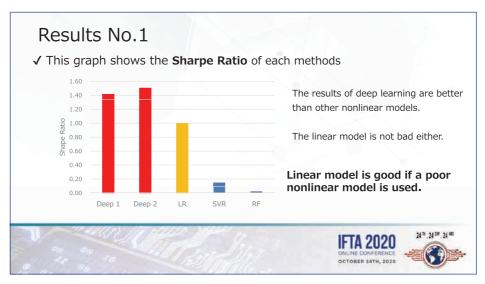


図 7. 各方法のシャープレシオ (TOPIX)

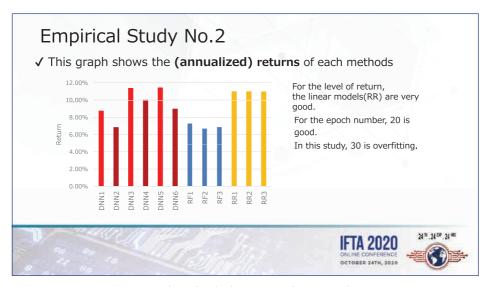


図 8. 各手法の年率リターン (TOPIX500)

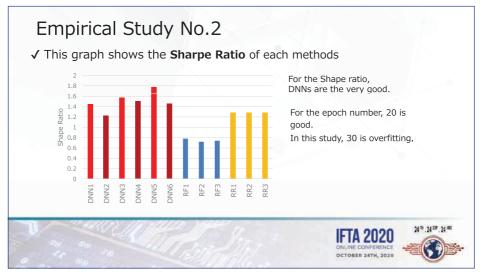


図 9. 各手法のシャープレシオ (TOPIX500)

デル (RR) が非常に相性が良い。ディープラーニングのエポック数に対しては、20 が良い。この研究では、30 は過剰適合となる (図8)。

シャープレシオで見ると、形状比の場合の ディープラーニングは非常に良い。また、日次予 測のリターンとシャープレシオは、これまでの研 究の月次予測のリターンよりも高い(図9)。

【実証分析 3】MSCI のグローバルな指数 (全世界、北米、欧州&中東、太平洋)を使う。20のファクターを用いて相対的なリターンを予測し、毎月投資する。

【結果】各手法のリターンのグラフ。赤い線はディープラーニングを示しており、全地域で他の手法よりも優れている。さらに興味深いことに、予測モデルに関係なく、太平洋→欧州&中東→北米の順にパフォーマンスが低下する(図 10)。

シャープレシオでも、全地域でパフォーマンス がよいのはディープラーニング。地域別には、太 平洋→欧州&中東→北米の順に低下する(図 11)。

市場の効率が高いほど、リターンを得ることはより困難になる。

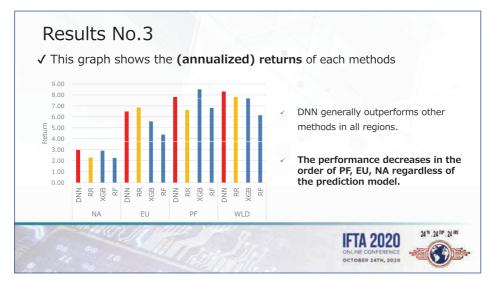


図 10. 各手法の年率リターン予測 (MSCI 指数)

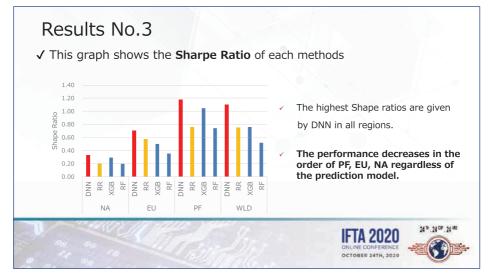


図 11. 各手法のシャープレシオ (MSCI 指数)

●プロフィール-

中川 慧 (なかがわけい)

京都大学経済学部卒。三井住友アセットマネジメントでファンドのリスク管理、国内/海外株式、アセットアロケーションのクオンツファンド運用/助言やモデル開発を行う。2018年より野村アセットマネジメント資産運用先端技術研究部にて、AIやビッグデータを組み合わせた先端的なクオン



ツ運用戦略の研究開発に取り組む。金融工学、人工知能、機械学習に関する学術研究論文多数。筑波大学大学院経営学修士 (MBA)、博士号 (Ph.D) を取得。東京工業大学、筑波大学、早稲田大学で教鞭をとる。

【結論】

- ・時系列予測も断面予測も、機械学習によって強 化することができる。
- ・時系列予測ではデータ量が限られるので、k-NN のような簡単な方法が優れている。
- ・断面予測では多くのデータが存在するため、 ディープラーニングのような複雑なモデルが優れている。
- ・問題の処方に応じて、異なる手法を採用すべき である。