

吉野貴晶 のクオンツ トピックス : NO10

クオンツとAI/機械学習の融合 (AI/機械学習によるモメンタム強化)

AI/機械学習によるモメンタム効果の改良

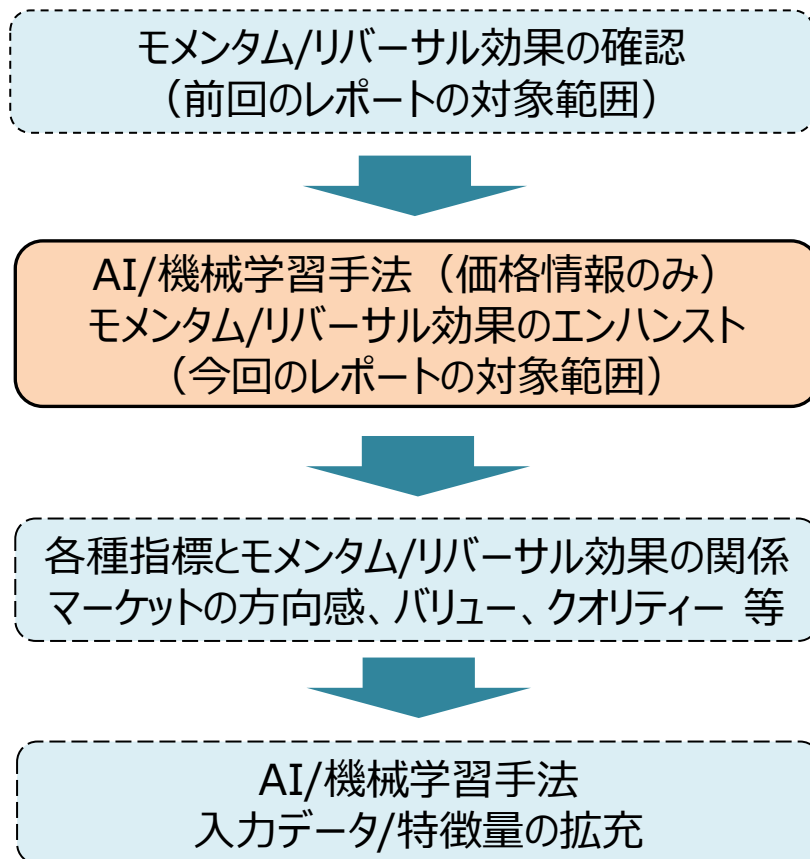
- 連載形式でAI (人工知能) と投資手法の関係性を紹介。
- 引き続き、クオンツ手法の一つであるモメンタム/リバーサルが今回のテーマ。

最近、AI (人工知能、以下AI) に関連するニュースが増えています。投資の分野でも研究開発が盛んに行われており、実際に投資手法として利用可能な段階まで進展しています。本レポートでは、AIと投資手法の関係性をご紹介したいと思います。

前回レポートから引き続き、「クオンツとAI/機械学習の融合」をテーマにします。

投資の世界には従来からクオンツという職種が存在します。具体的には、統計解析等の数学的手法を用いて、マーケットの分析や投資戦略の構築をするのが主な仕事になります。前回のレポートでは、クオンツ領域において有名な投資手法であるモメンタム/リバーサル効果の概論をご紹介しました。今回のレポートでは、モメンタム/リバーサル効果の強化 (エンハンスト) 案として、AI/機械学習の適用を模索していきます。数回の連載形式を予定しています。

図1. モメンタム/リバーサルに係る連載予定



モメンタム効果とは？

1. モメンタム/リバーサル現象とは？（前回レポートの再掲）

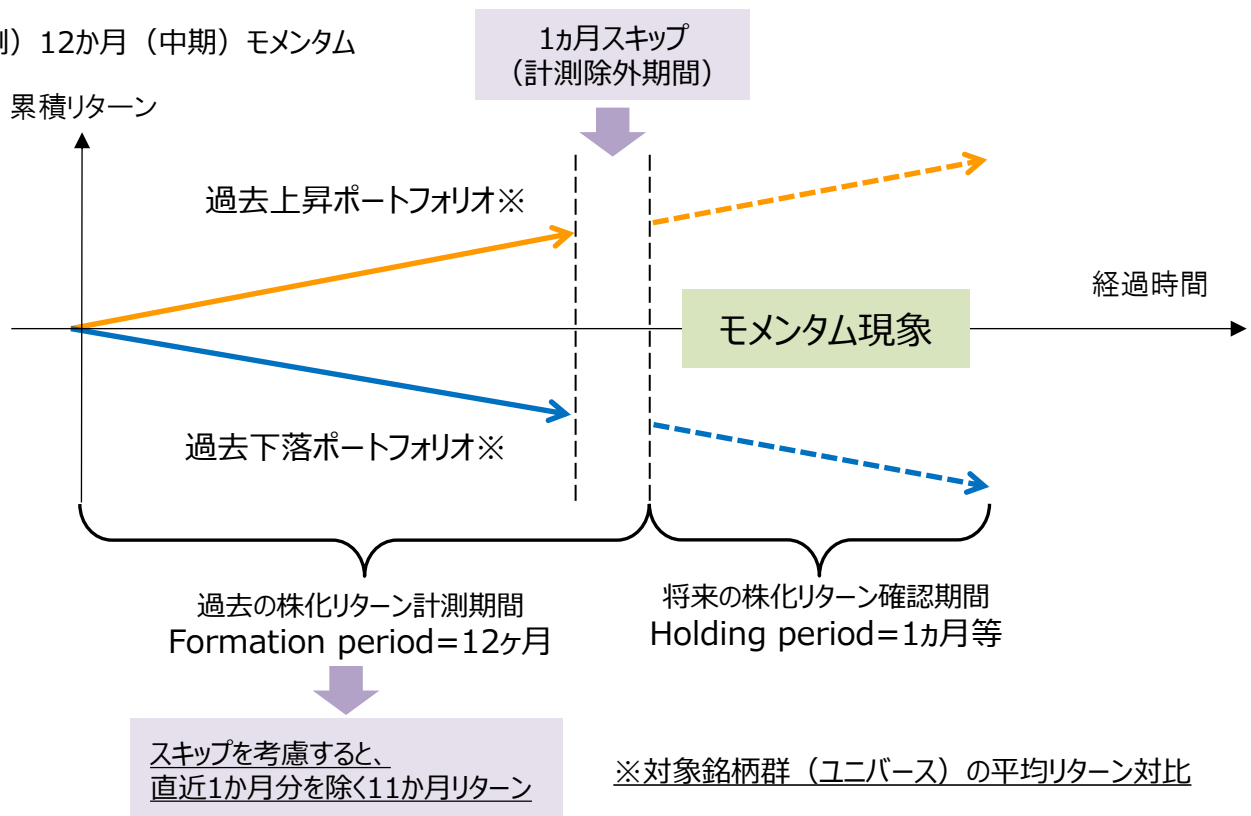
代表的な投資手法として、モメンタム/リバーサル現象が知られています。モメンタム/リバーサル現象とは、マーケットで観測されるアノミーであり、効率的市場仮説では説明が難しいとされています。具体的にモメンタム現象とは、過去の一定期間に、他の銘柄と比べてリターン上昇が大きかった（小さかった）銘柄群は、その後の一定期間においても相対的な上昇（下落）傾向を持続することであり、順相関傾向を示す事象の事です。リバーサル現象は、逆相関傾向を示す事象です。

過去一定期間における上昇傾向の銘柄群を買い（ロング）、下落傾向の銘柄群を売り（ショート）した場合の合成ポートフォリオをモメンタムポートフォリオと言います。用語を以下に整理しますが、上昇傾向銘柄群を過去上昇ポートフォリオ、下落傾向銘柄群を過去下落ポートフォリオと以下で定義しています。

図2. モメンタム/リバーサルの定義

モメンタム/リバーサル区分	過去と将来の関係	パターン①		パターン②	
		過去	将来	過去	将来
モメンタム	順相関（同じ向き）	↑	↑	↓	↓
リバーサル	逆相関相関（逆向き）	↑	↓	↓	↑

（例）12か月（中期）モメンタム



★モメンタムポートフォリオ

= 過去上昇ポートフォリオ - 過去下落ポートフォリオ

★スキップとは？

Formation periodにおいて、直近数か月分のリターンを除外する処理

モメンタム/リバーサルの計測では一般的。理由としては、短期リバーサル、中期モメンタム、長期リバーサル等の効果を分けて考えるため。

AI/機械学習で結果を向上（エンハスト）できるか？

2. AI/機械学習の導入

前回のレポートで検証したモメンタム/リバーサル効果について、AI/機械学習を用いて結果を向上（エンハスト）できないか検討していきます。R&Dのイメージを共有して頂くために、成功事例のみ載せるのではなく、失敗と試行錯誤も掲載します。一般的に、通常の機械学習モデルでは、複数（大量）の入力データ系列を活用したモデル構築を目指します。しかし、いきなり複雑なモデル構築に挑戦するのは、最初から応用問題を解こうとしている状況に等しいと言えます。まずは、データもモデルも最もシンプルな仕様から始め、データとモデルの傾向や癖を掴みつつ改善方法を模索していきたいと思えます。

2_1. モデルとベースラインの決定

シンプルな機械学習モデルとして、今回は決定木モデル、その中の分類木モデルを利用します。決定木モデルとは、文字通り木構造を用いており、段階的にデータを分割していきます。分類木モデルでは、最終的にどの分類値になるかを推計するモデルです。直感的に理解する上で、図3をご覧ください。実際のモデル構築では、インプットデータとして12ヶ月リターン（1ヵ月スキップ）を利用し、予測値は将来1ヵ月予想値の3値分類です。3値分類ですので、モデルがでたらめでランダムに予想値を吐き出すと33.3%の正答率となります。よって、超えるべき分類正答率のベースラインはこの33.3%となります。

データの期間としては、モデルを構築するために使う期間（学習期間）は、月次頻度で2000/1から2009/12の10年間とします。また、上記の学習期間で作成されたモデルを使い予測値を算出する期間（テスト期間）は、同じく月次頻度で2010/1から2018/11迄とします。なお、対象となる銘柄は東証1部とし、データが取れる銘柄は全て対象とします。

図3. 分類木モデルのイメージ（入力データは0.0～1.0の範囲、目的ラベルが1、2、3の3種類の場合）

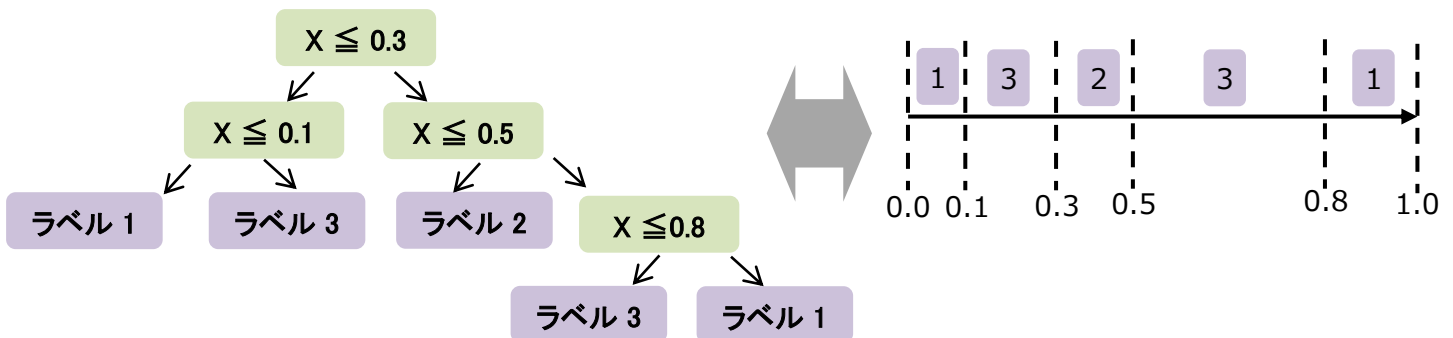


図4. データ詳細

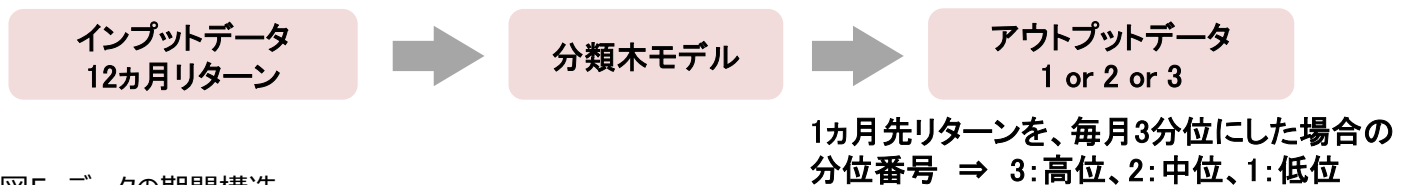
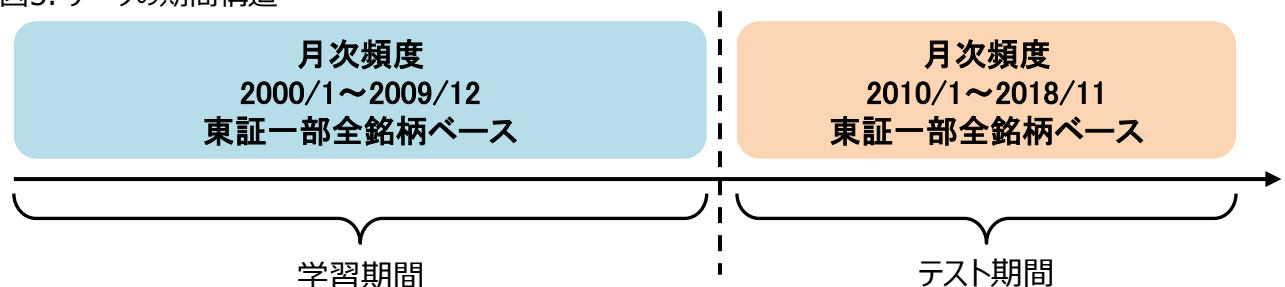


図5. データの期間構造



●当資料は、市場環境に関する情報の提供を目的として、ニッセイアセットマネジメントが作成したものであり、特定の有価証券等の勧誘を目的とするものではありません。●当資料は、信頼できると考えられる情報に基づいて作成しておりますが、情報の正確性、完全性を保証するものではありません。●当資料のグラフ・数値等はあくまでも過去の実績であり、将来の投資収益を示唆あるいは保証するものではありません。また税金・手数料等を考慮しておりませんので、実質的な投資成果を示すものではありません。●当資料のいかなる内容も将来の市場環境の変動等を保証するものではありません。

AI/機械学習で結果を向上（エンハスト）できるか？

2_2. 決定木モデルの結果

さて、以下に決定木モデルでの結果を記載します。この結果から分かることは、累積リターンで見ると全くワークしていないモデルになっていることが分かります。これならばむしろAI/機械学習はやらない方が良いと言えます。

簡単な分析として、各予測ラベルにおける正答率の表を作成しています。3分位のうち、中位に当たる部分は当てられているのですが、下位1/3は正答率が悪くなっています。上位1/3に関しては正答率が5%を切るほど悪いです。結果として、全体の正答率はベースライン（33.3~33.4%）を越えていても、パフォーマンスは上がらないという結果を引き起こしています。

図6. 決定木モデルの結果

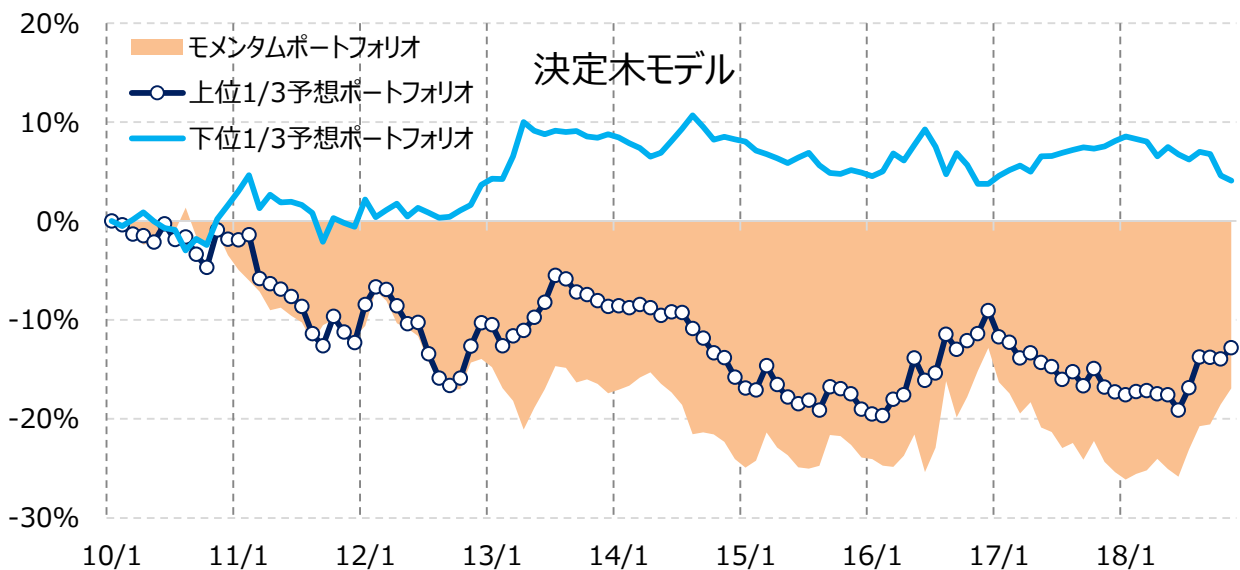


図7. モデルのラベル別正答率

正答率はベースラインを超えてはいるが...

計測期間	期間平均							
	全体 ベースライン	全体 正答率	実績下位1/3		実績中位1/3		実績上位1/3	
			正解	不正解	正解	不正解	正解	不正解
2010/01~2018/11	33.4%	36.0%	29.6%	70.4%	74.5%	25.5%	3.9%	96.1%
2010/01~2010/12	33.3%	35.1%	27.4%	72.6%	74.1%	25.9%	3.8%	96.2%
2011/01~2011/12	33.4%	36.3%	23.4%	76.6%	82.7%	17.3%	2.7%	97.3%
2012/01~2012/12	33.4%	36.6%	24.3%	75.7%	82.1%	17.9%	3.3%	96.7%
2013/01~2013/12	33.4%	34.2%	40.5%	59.5%	57.1%	42.9%	5.1%	94.9%
2014/01~2014/12	33.3%	36.7%	30.1%	69.9%	76.0%	24.0%	4.0%	96.0%
2015/01~2015/12	33.4%	36.4%	34.7%	65.3%	70.6%	29.4%	3.9%	96.1%
2016/01~2016/12	33.4%	35.9%	23.4%	76.6%	80.3%	19.7%	3.8%	96.2%
2017/01~2017/12	33.3%	35.7%	31.0%	69.0%	71.7%	28.3%	4.2%	95.8%
2018/01~2018/12	33.4%	37.4%	31.9%	68.1%	76.0%	24.0%	4.3%	95.7%

特に上昇銘柄が全然当てられない

●当資料は、市場環境に関する情報の提供を目的として、ニッセイアセットマネジメントが作成したものであり、特定の有価証券等の勧誘を目的とするものではありません。●当資料は、信頼できると考えられる情報に基づいて作成しておりますが、情報の正確性、完全性を保証するものではありません。●当資料のグラフ・数値等はあくまでも過去の実績であり、将来の投資収益を示唆あるいは保証するものではありません。また税金・手数料等を考慮しておりませんので、実質的な投資成果を示すものではありません。●当資料のいかなる内容も将来の市場環境の変動等を保証するものではありません。

データにおけるトライ&エラー

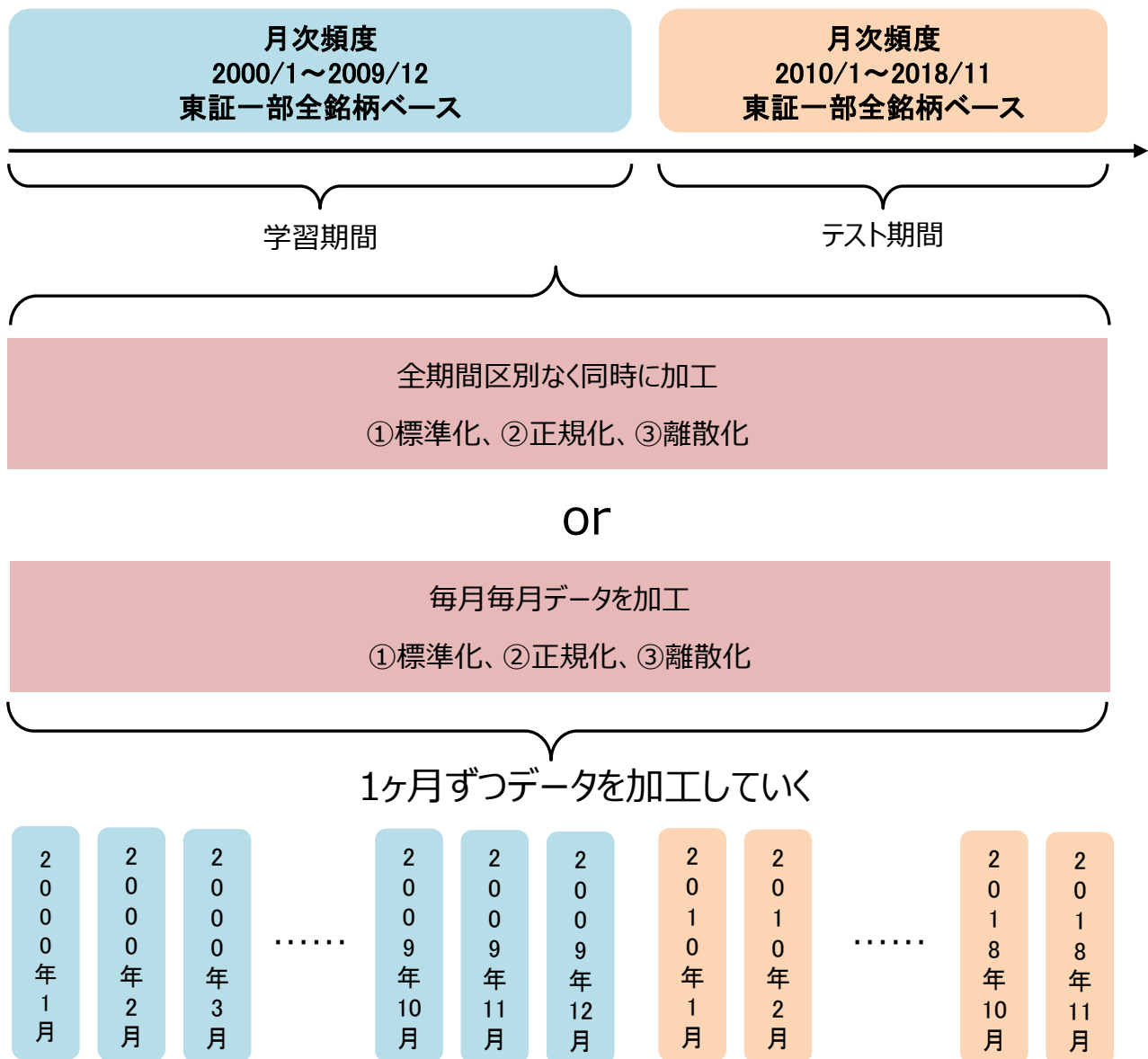
3. 結果の改善に向けた試行錯誤（データ編）

先述の決定木モデルでは、分類正答率こそベースラインから超えているものの、その分類予測値を利用した銘柄群のリターン累積は明らかに失敗と言える結果でした。ここからは改善方法を考えたいと思います。

3_1. 入力データの加工（特徴量エンジニアリング）

機械学習において、モデルに併せてデータに様々な加工処理を施す作業は非常に重要であり、特徴量エンジニアリングと呼ばれています。本レポートでも、入力データの加工が結果に与える影響を確認したいと思います。色々な手法が考えられますが、今回は3パターン試してみます。①入力データを標準化（z-score normalization）、②正規化（Min-Max normalization）、③離散化、の三つです。さらに、上記3つの手法を実施する上で、全てのデータに対して同時に加工するパターンと、毎月毎月加工を実施していくパターンを考えます。

図8. 入力データ加工のイメージ



モデルにおけるトライ&エラー

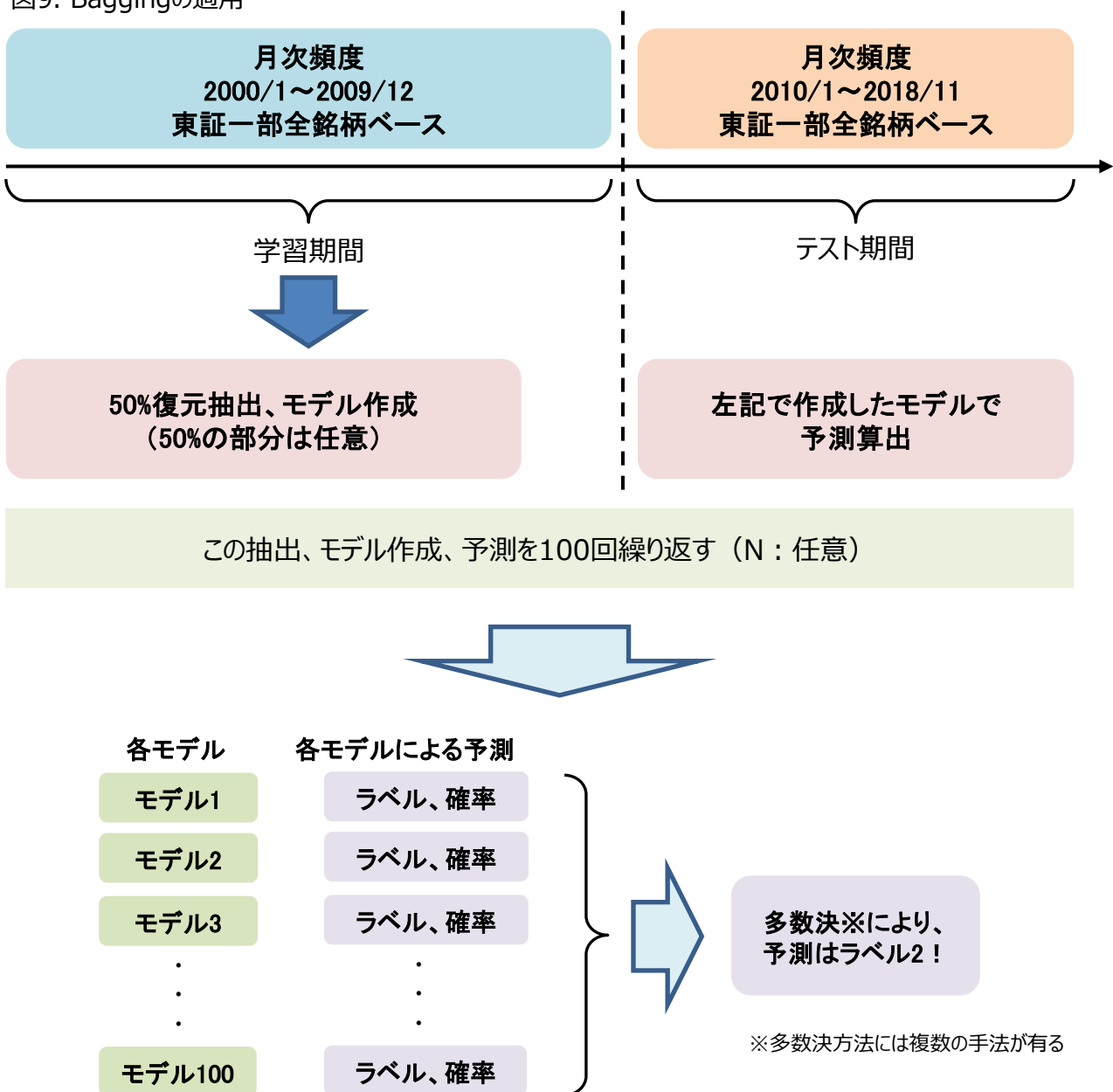
4. 結果の改善に向けた試行錯誤（モデル編）

入力データの加工だけではなく、モデルにも改良を加えてみます。他のAI/機械学習系モデルでも言えますが、決定木のマイナスな特徴として、そもそも決定木は過剰適合しやすい、という点があります。そこで、過剰適合への対策として、Baggingを導入します。

4_1. Bagging

今回はBaggingという方法を導入します。具体的には、学習モデル（弱学習器）を複数作成して、そのモデルたちの多数決※で予測値を決定します。複数作るために、学習データから復元抽出して学習用データセットを作成、このデータセットでモデルを一つ作成します。この抽出処理をN回繰り返して複数のモデルを生成します。今回は、データセットの抽出条件を50%、作成するモデルの数を100個とします。

図9. Baggingの適用

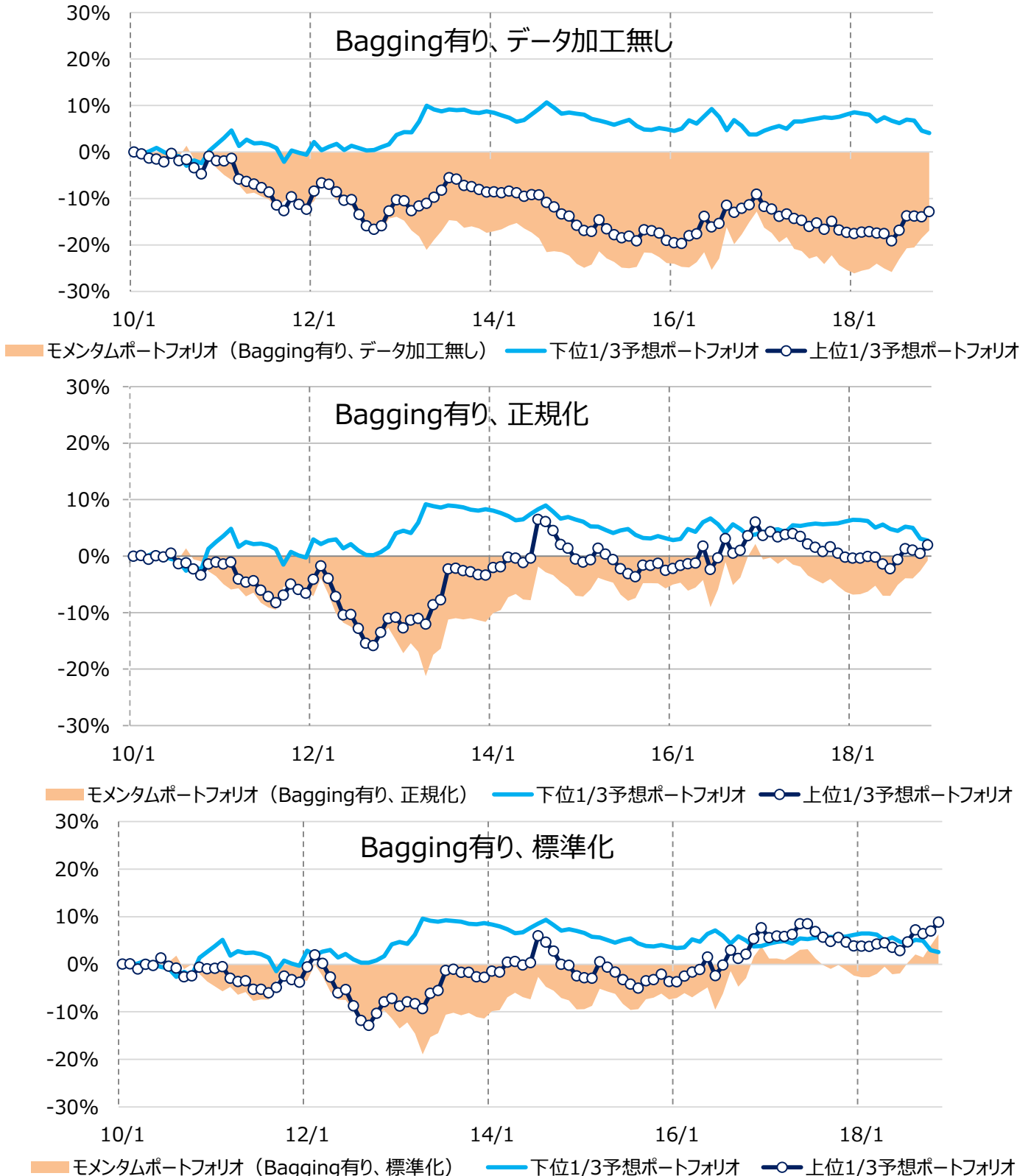


AI/機械学習で結果を向上（エンハスト）できるか？

5. 改良結果

データの加工及びBagging導入後の結果を見てみましょう。データ加工無しに比べて、入力データに対して一括で正規化と標準化を加えた場合は幾分改善してはいますが、モメンタムポートフォリオの累積リターンはマイナスまたはゼロ近傍です。違う改良が必要なようです。

図10. 入力データ一括加工処理のパフォーマンス結果

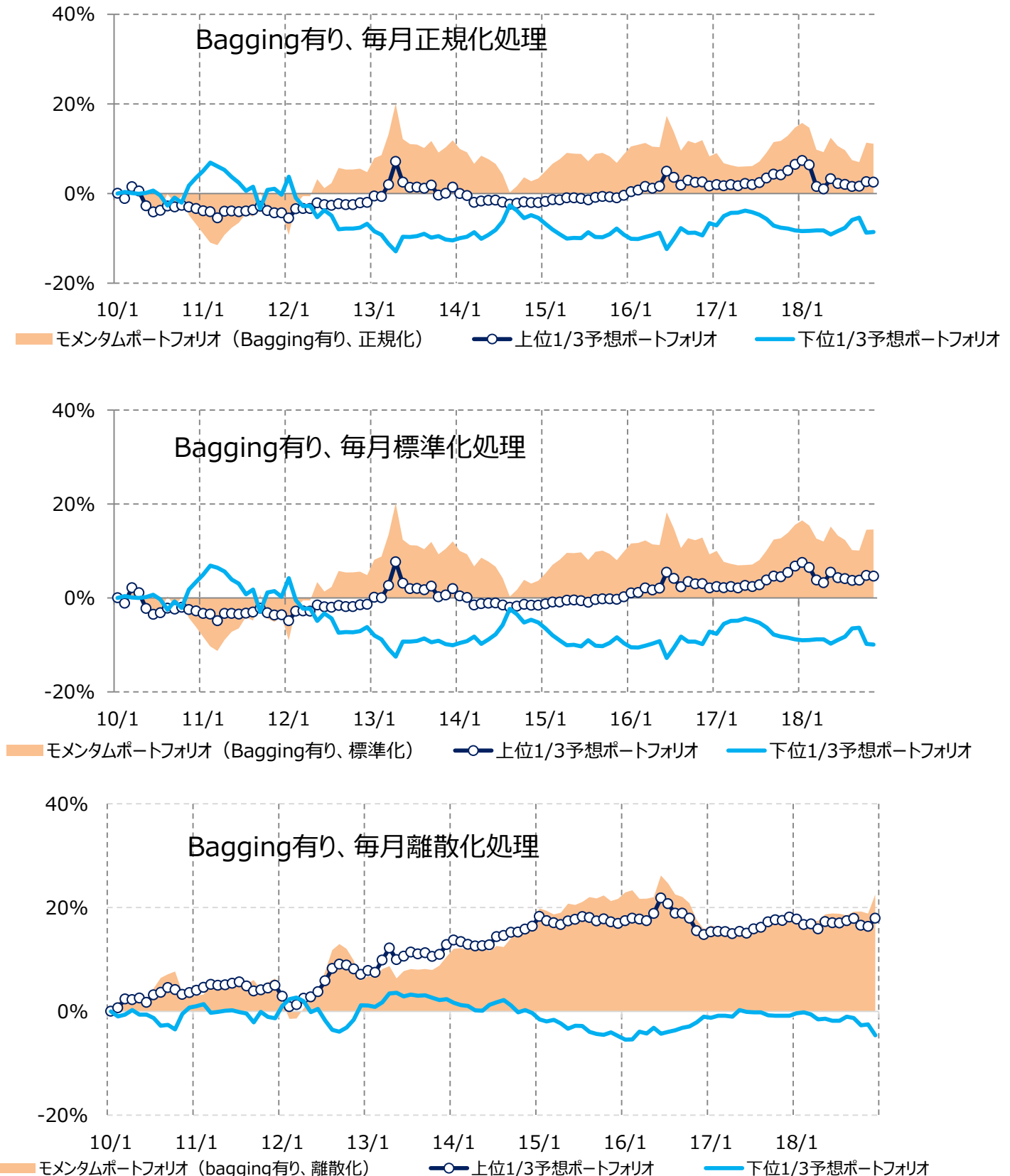


●当資料は、市場環境に関する情報の提供を目的として、ニッセイアセットマネジメントが作成したものであり、特定の有価証券等の勧誘を目的とするものではありません。●当資料は、信頼できると考えられる情報に基づいて作成しておりますが、情報の正確性、完全性を保証するものではありません。●当資料のグラフ・数値等はあくまでも過去の実績であり、将来の投資収益を示唆あるいは保証するものではありません。また税金・手数料等を考慮しておりませんので、実質的な投資成果を示すものではありません。●当資料のいかなる内容も将来の市場環境の変動等を保証するものではありません。

AI/機械学習で結果を向上（エンハスト）できるか？

次は各月毎にデータ加工処理を施した結果になります。以前のグラフと比べると、明らかに結果が改善しています。特に毎月離散化処理が良さそうです。

図11. 入力データ毎月加工処理のパフォーマンス結果



●当資料は、市場環境に関する情報の提供を目的として、ニッセイアセットマネジメントが作成したものであり、特定の有価証券等の勧誘を目的とするものではありません。●当資料は、信頼できると考えられる情報に基づいて作成しておりますが、情報の正確性、完全性を保証するものではありません。●当資料のグラフ・数値等はあくまでも過去の実績であり、将来の投資収益を示唆あるいは保証するものではありません。また税金・手数料等を考慮しておりませんので、実質的な投資成果を示すものではありません。●当資料のいかなる内容も将来の市場環境の変動等を保証するものではありません。

AI/機械学習で結果を向上（エンハスト）できるか？

離散化に関しては、図13においてモメンタムポートフォリオベースでバギング無しとも比較検討していますが、bagging有りの方が効果が優っています。

これまでの結果から分かることは、「データの加工処理は毎月処理が良い」、「正規化、標準化、離散化なら離散化が良い」、「Bagging有り無しなら有りの方が良い」となります。毎月処理かつ離散化かつBagging有りの結果は、AI/機械学習を何も使っていないシンプルな中期モメンタムモデル（図12）と比べると、累積リターンが改善しています。よって、機械学習によるエンハストと効果はありそうだ、と言えるかと思えます。

図12.（前回レポートの再掲）リーマンショック後期間での検証

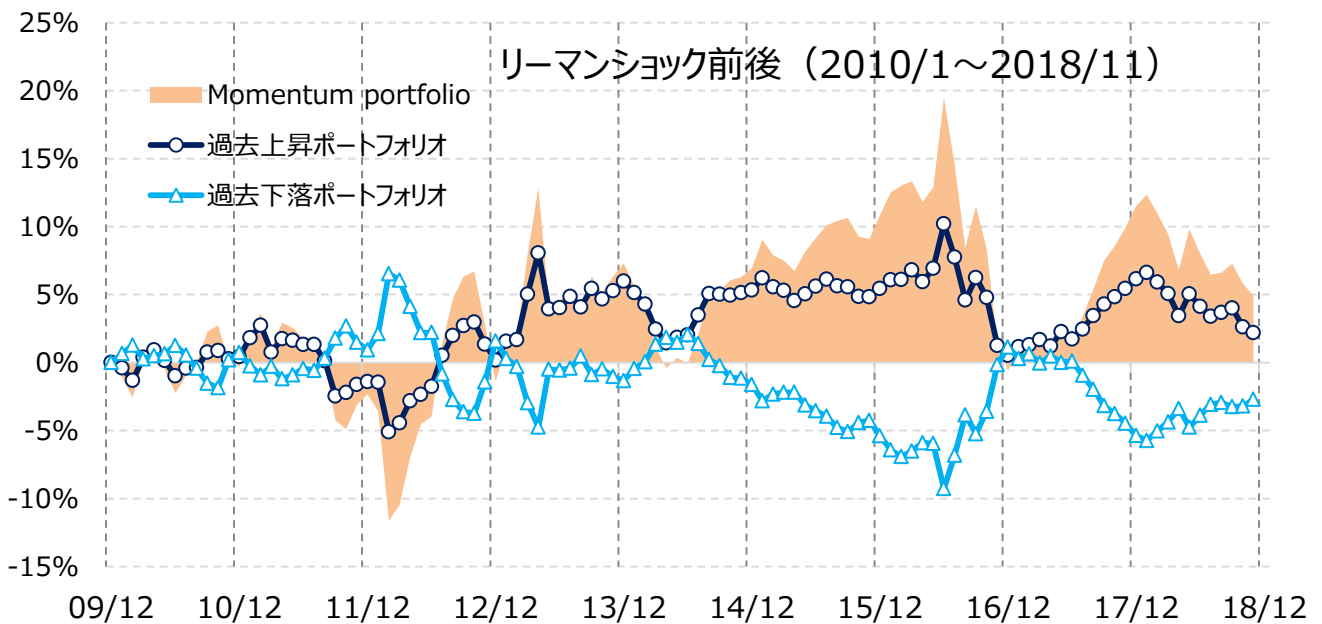
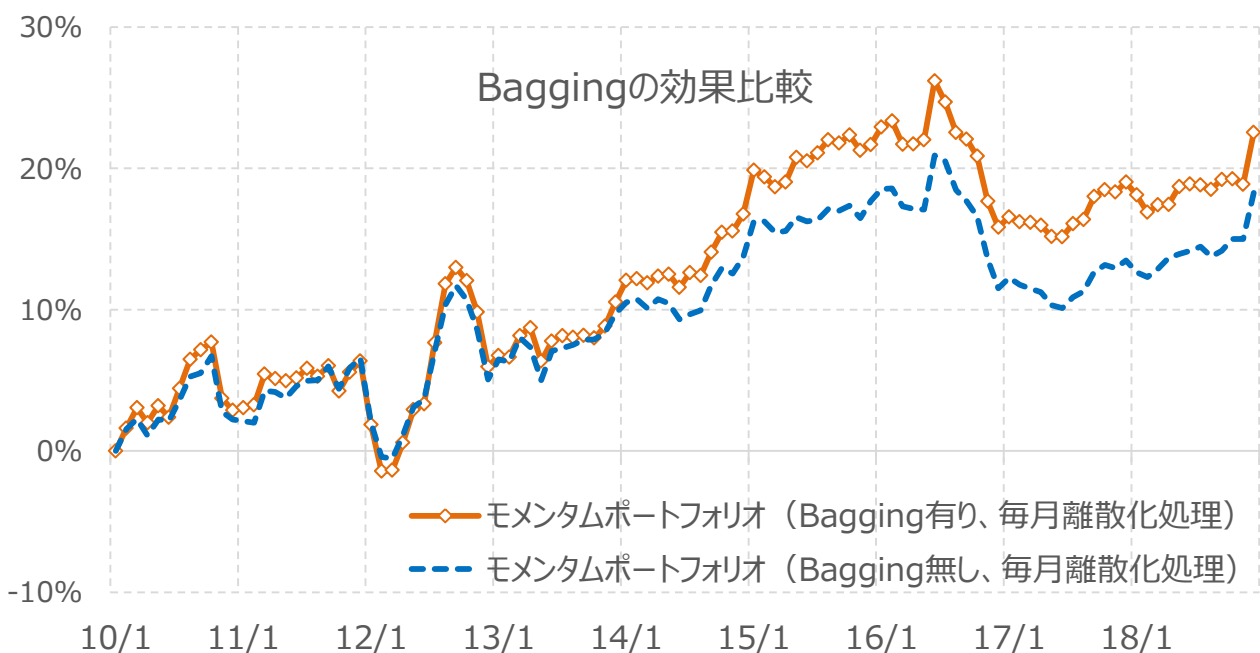


図13. Bagging有り無しにおけるパフォーマンスの比較（モメンタムポートフォリオ部分）



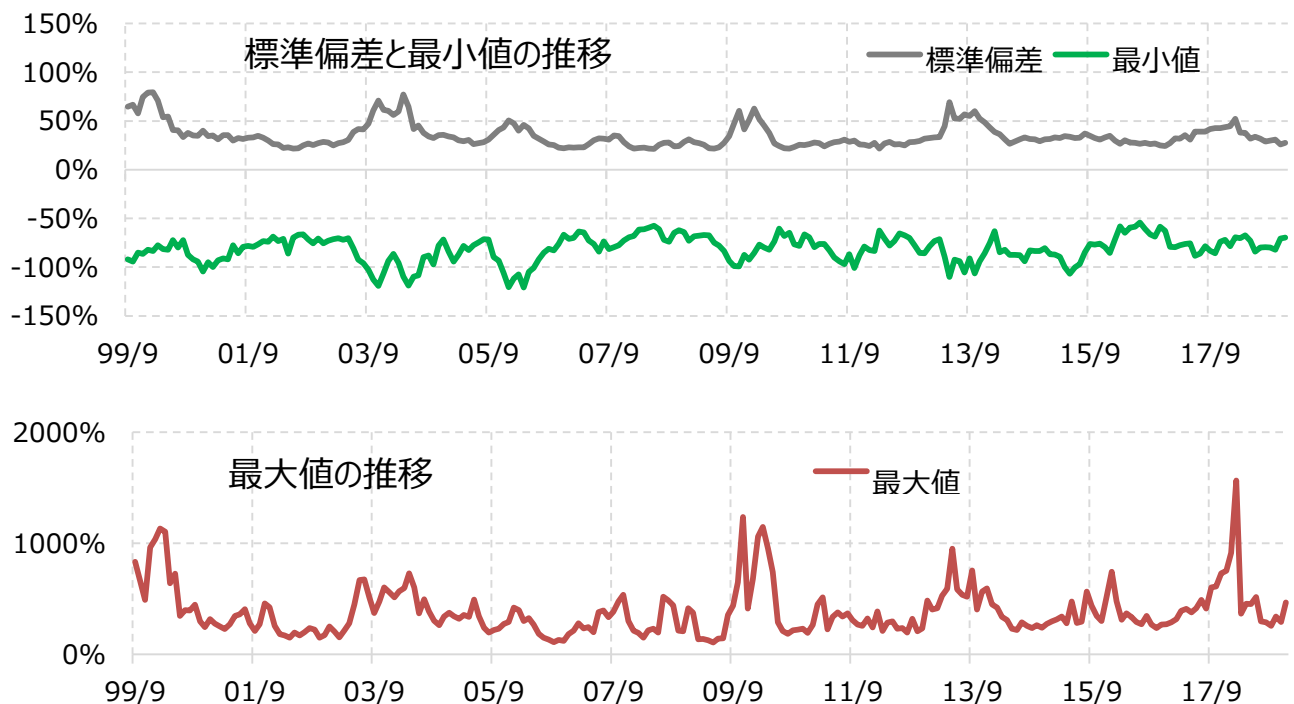
●当資料は、市場環境に関する情報の提供を目的として、ニッセイアセットマネジメントが作成したものであり、特定の有価証券等の勧誘を目的とするものではありません。●当資料は、信頼できると考えられる情報に基づいて作成しておりますが、情報の正確性、完全性を保証するものではありません。●当資料のグラフ・数値等はあくまでも過去の実績であり、将来の投資収益を示唆あるいは保証するものではありません。また税金・手数料等を考慮しておりませんので、実質的な投資成果を示すものではありません。●当資料のいかなる内容も将来の市場環境の変動等を保証するものではありません。

時系列金融データにおける分散の不安定性

6. データを毎月加工処理することのメリットとは？

複数のパターンを検証しましたが、大まかに言える事の一つとしては、月次でデータ加工を施すと総じて結果が改善する、と言う事かと思えます。これは、金融時系列データにおいて毎月の分散が一定ではない点に起因している可能性があります。以下の図を見ると分かるように、12ヶ月リターン（1ヶ月スキップ）の毎月のばらつき（ここでは標準偏差）は一定ではありません。値自体の最大値、最小値も安定しません、中でも最大値は月によって大幅なばらつきがあります。これは、株価の性質上、下落する場合には株価が0近傍までしか行きませんが、上昇する場合には青天井になるためです。このように、説明変数に分散のばらつきがある一方、教師データの予測ラベルである1ヶ月先リターンは、毎月毎月3分位化の処理を実施しています。この差異が機械学習モデルの学習を難しくしていた、と推測されます。※個別銘柄の12ヶ月リターン（1ヵ月スキップ）に対して、毎月マーケット平均リターンを控除しているため、理論上最小値は-100%を下回る場合がある。

図14. 入力データの特徴の時系列推移



7. 今後の展開

次回以降も、引き続きモメンタム効果に焦点を当てた分析を継続します。投資手法としてのモメンタム効果において、他の様々な要因、例えばサイズ、マーケットの方向感、バリュエ等の他の代表的指標が、モメンタム効果と関連性があると言われていています。これらの要因とモメンタム効果の関連性を確認することで、機械学習モデルに利用する新たな特徴量（入力データ）の作成を目指します。この特徴量にモデル改良を加えて、AI/機械学習を活用したモメンタム効果の強化（エンハンス）を目指します。（筆者の都合でテーマが変わる場合があります。）

～執筆者の紹介～

吉野貴晶（写真：右）

「日経ヴェリタス」アナリストランキングのクオンツ部門で16年連続で1位を獲得。ビッグデータやAIを使った運用モデルの開発から、身の回りの意外なデータを使った経済や株価予測まで、幅広く計量手法を駆使した分析や予測を行う。



高野幸太（写真：左）

ニッセイアセット入社後、ファンドのリスク管理、マクロリサーチ及びアセットアロケーション業務に従事。17年4月に投資工学開発室に異動後は、主に計量的手法やAIを応用した新たな投資戦略の開発を担当する。

●当資料は、市場環境に関する情報の提供を目的として、ニッセイアセットマネジメントが作成したものであり、特定の有価証券等の勧誘を目的とするものではありません。●当資料は、信頼できると考えられる情報に基づいて作成しておりますが、情報の正確性、完全性を保証するものではありません。●当資料のグラフ・数値等はあくまでも過去の実績であり、将来の投資収益を示唆あるいは保証するものではありません。また税金・手数料等を考慮しておりませんので、実質的な投資成果を示すものではありません。●当資料のいかなる内容も将来の市場環境の変動等を保証するものではありません。