価格変動パターンを用いた
株価予測手法の実証研究

筑波大学審査学位論文（博士）

2019

中川 慧

筑波大学大学院
ビジネス科学研究科 企業科学専攻

概要

株価の予測可能性は、学術的にも実務的にも重要な研究テーマである。実務的には、株価の予測はファンダメンタル分析とテクニカル分析に基づき行われてきた。テクニカル分析はその有効性は実務的にも学術的にも認められつつも、主観性、恣意性が強いものであるとの指摘を多く受けている。本研究は、テクニカル分析の哲学・論理である価格変動パターンという観点に着目しながらも、客観的、機械的な予測手法を開発することを目的とする。すなわち、現在の価格変動パターンが過去のある時点と似ていれば、そのパターンを用いて将来の価格の予測ができるだろうという仮説を基本とし、このような幾何学的な価格変動パターンを機械的に抽出し、予測へ活用し、その有効性を検証する。第 2 章において、実務的なファンダメンタル分析、テクニカル分析という観点ではなく、株価予測に関する学術的な先行研究を次の観点から整理する。株価予測の先行研究はクロスセクション分析と時系列分析という方法論の観点、テクニカルとファンダメンタルというデータの観点、そしてパラメトリックとノンパラメトリックという手法の観点がある。それぞれ、クロスセクション分析とは複数企業群のある断面での属性 (例えば、時価総額など) に基づいて比較を行う方法、時系列分析はある銘柄に着目して過去の株価等を時系列データとして捉え、時系列に予測を行う方法である。データについては、株価や出来高などの市場データがテクニカル、財務、経済データをファンダメンタルといい、手法は分布を仮定するパラメトリックな手法か、分布に仮定を置かないノンパラメトリックな手法かという観点をそれぞれいう。本研究の位置付けは、時系列分析の観点から、データに過去の価格変動パターンというテクニカルデータを用いた、ノンパラメトリックな手法の提案である。先行研究として、データに市場データである価格を用いた研究は多数あるものの、価格変動パターンそのものを扱った先行研究は少ない。また、予測手法としてもノンパラメトリックな手法を用いるが、ニューラルネットワークやサポートベクターマシーンのような目的変数の予測過程が理解しづらく、各説明変数の目的変数に対する重要度・寄与度などが評価できないモデルではなく、シンプルな k-Nearest Neighbor 法や決定木をベースとした可読性、解釈性の高いアルゴリズムをベースとした。第 3 章では、

各国の代表的な株式指数に対して、アルゴリズムにより機械的に抽出された過去の価格変動パターンが、将来の株価予測に有効であるかをシンプルな手法を用いて検証する。音声認識の分野で時系列データの類似度計測に用いてきた動的時間伸縮法 (Dynamic Time Warping;DTW[Ita75]) を、テクニカル分析の観点から指数化動的時間伸縮法 (Indexing Dynamic Time Warping;IDTW) として改良し、価格変動パターンの抽出に使用する。DTW は時系列間の長さが異なる場合でも使用でき、かつ時間方向のずれを許容するため、人間の直感にあった時系列間の類似度を計測できるといわれている。そしてノンパラメトリックな予測に手法の中で最もシンプルな k-Nearest Neighbor(k-NN) 法の改良である、k^*-Nearest Neighbor 法と IDTW を組み合わせた株価予測手法を提案した。提案手法を用いて各国の代表的な株価指数を予測したところ、過去の価格変動パターンは有効な特徴量であることが確認できた。また株価変動パターンは単純なモメンタム、リバーサル戦略以上の収益機会が存在することを合わせて確認した。一方で、DTW と k-NN の組み合わせは時系列データマイニングにおいて非常に有効な手法であるが、株価に対してそのまま適用しても予測精度は低いことが確認できた。株価のような金融時系列に対して、DTW を用いた価格変動パターンを用いて予測に有効なアルゴリズムを作成したこと、単純なモメンタム戦略とは相関が低く、より多くの収益が獲得できることを実証したことが本章の貢献である。第 3 章では、価格変動パターンが似ているというシンプルな仮定で予測を行ったが、どのような価格変動パターンが予測に有効であったかは不明である。そこで第 4 章では、3 章で有効であると確認できた過去の価格変動パターンを、クラスタリング手法を用いて可視化する。日本の株価指数 TOPIX を対象に可視化を行い、予測力という観点から、どのような価格変動パターンがクラスタリングによって抽出されたかを確認する。しかしながら前章までと同様に、月間の日次株価変動はデータ数 (営業日) が月毎に異なるため、単純なベクトル空間上のユークリッド距離を用いたクラスタリング手法 (例えば、k-means 法) は適用できない。よって、株価を対象に異なるデータ数においても、欠損値として値の挿入および削除を行うことなく、より自然にデータ間の類似性を測定し、対応するクラスタリング方法を適切に組み合わせる必要がある。そこで価格変動パターンの可視化のため、IDTW を類似度として用いた k-medoids 法によるクラスタリング手法を提案し、最も収益性や予測精度の高いクラスタ数 (5 つ) でクラスタリングを行った。株価変動パターンのクラスタを抽出したところ、当月の株価変動が大きい場合に、上昇、下落ともに強いモメンタム効果が可視化により確認できた。第 4 章までは、価格変動

パターンの類似性に着目した予測を行い、具体的にどのようなパターンが有効であったかを検証した。しかしながら、予測に有効な特徴量として株価変動パターン以外の情報も組み込むことで更なる予測精度の改善を見込むことができると考えられる。第 5 章では、前章までに確認した予測に有効な株価変動パターンと、その他のテクニカル、ファンダメンタルデータを組み合わせ、可読性の高いモデルを構築した。具体的には、木構造に基づく予測手法である時系列勾配ブースティング木を提案した。時系列勾配ブースティング木を構成する弱学習器として、時系列決定木を用いるが、先行研究における時系列決定木は時系列属性以外のデータを考慮できないため、時系列属性以外のデータを組み込んだ時系列決定木を提案した。各国の株式指数を用いた実証分析の結果、 単純な時系列決定木に対して勾配ブースティングを用いることで予測精度が改善し、また、時系列決定木に対してクロスセクションのデータを組み込むことでも予測精度が改善することが確認できた。

目次

図目次

表目次

第 1 章

序論

1.1　研究の背景と目的

株価の予測可能性は、学術的にも実務的にも重要な研究テーマである。株式市場は現代社会においては世界中の国々の経済活動に影響を与え、経済的・社会的に非常に重要な役割を果たす。その重要性のために、これまでに株価を予測するために様々な方法が提案されてきた。

実務的には、株価の予測はファンダメンタル分析とテクニカル分析に基づき行われてきた。ファンダメンタル分析では、経済・産業調査や企業の経営、財務分析を行い、適切な株価を見積り、見積もった株価と現状の株価の比較により売買判断を下す。売上や利益の伸び、保有資産と株価のバランスなどの財務データが主に分析対象となる。また、ファンダメンタル分析で使用されるデータの中には文章などで定性的に表されるものも存在している。

一方、テクニカル分析は、主に過去の株価変動パターン (チャート) を用いて将来の株価変動を予測し、売買判断を下す。株価そのものや出来高などの市場データが主に分析対象となり、ファンダメンタル分析とは異なり数値データのみが分析の対象となる。

ファンダメンタル分析、テクニカル分析はその特性上向き不向きがあり、テクニカル分析は比較的短期的な予測に適しており、ファンダメンタル分析は中長期的な予測に適しているといわれている。

近年、インターネットの普及により株式取引の手数料の低コスト化が実現されるとともに、計算機の性能向上によって、コンピュータプログラムによる株式の自動取引、いわゆるシステムトレードが現実のマーケットにおいてその割合を増加させていると言われてい

る。こうしたシステムトレードが売買判断を下す上で一般的に使用する手法は、その詳しい仕組みやアルゴリズムは公開されていないことが多いものの、主にテクニカル分析であると言われている。

このようなテクニカル分析の哲学・論理は、

(1) 投資家の行動を反映した結果として株価が決まる。
(2) 株価は一定の変動パターンを形成する。
(3) 株価は繰り返す。

の 3 つである。

株価を動かす要因は多数あるが、それらの要因によって投資家の行動が決定され、売買を通じて株価に反映される。投資家を動かした要因を分析しなくても、予測したい株価そのものを分析すれば、将来を含めた株価の動きが分かるというのが (1) である。(2) は、投資活動はランダムに行われるのではなく、多くの場合、投資家には共通する投資パターンがあり、その結果として一定期間、一定方向に株価が動く特徴、変動パターンを示すというものである。(3) は、投資活動は投資家の心理に基づいて行われ、投資家の心理が変わらないのであれば、株価変動は繰り返すというものである。このような観点からテクニカル分析には大きく、

(A)　トレンド分析
(B)　モメンタム分析
(C)　フォーメーション分析

といった技法がある [Mur99, 伊藤 10]。(A) トレンド分析は、価格変動の細かい値動きを無視すれば、ある大きな方向性 (トレンド) を持って動いており、この流れをとらえ、それに乗っていくための分析をトレンド分析という。(B) モメンタム分析は、方向性を探るためのトレンド分析に対し、現在の価格が買われ過ぎなのか、それとも売られ過ぎなのかを判断し、その後の反転などを分析するための手法をいう。(C) フォーメーション分析は、価格推移にはいくつかの幾何学的な類型 (パターン) があり、現在までの価格推移がどのパターンに該当するか判断し、今後の展開を予想するための手法であり、パターン分析ともいわれる。

実務的には国内外を問わずテクニカル分析を重視する投資家は多い。例えば、Billingsley

and Chance (1996)[BC96] は、Commodity Trading Advisor(CTA) と呼ばれる米国における商品先物取引を行うファンドの約 60% が、テクニカル分析よる取引システムを利用していることを調査した。より最近では、Menkhoff (2010) [Men10] はアメリカ、ドイツ、スイス、イタリア、タイの 5 ヵ国のファンドマネージャーに対して、アンケート調査を実施し、彼らが 1 か月程度の短期間においては、ファンダメンタル分析よりもテクニカル分析を重視して投資判断に用いていることを明らかにした。また、日本においては筒井・平山 (2009)[筒井 09] が日本の機関投資家にアンケート調査を行い、10% 程度の投資家が、ファンダメンタル分析や直観的判断よりも、テクニカル分析を重視していると回答している。

しかしながら古くからファイナンスの学術研究において、テクニカル分析は、Campbell et al. (1997)[CLM$^+$97] が述べているように、Cowles (1933)[Cr33]、Fama and Blume (1966) [FB66] など、多くの研究者が懐疑的、否定的に扱い研究されてこなかった。日本においても同様に、袴田 (2002)[袴田 02] によると、テクニカル分析はトレンド推定のための手法として科学的な客観基準によるものではなく、経験に基づいた直感に依るところが大きいと論じている。また刈屋 (2003)[刈屋 03] もテクニカル分析を広い意味での時系列分析であると述べているものの、株価の変動パターンの理解の仕方は、その専門家の判断に依拠した主観性が強いものであると意見を述べている。つまり、各分析手法について機械的、客観的な適用方法が定義されておらず、各人の主観に従って分析が行われ、事後的な検証ができない。例えば、トレンド分析において、いくつかの価格をつないでトレンドを作成するが、どのように価格をつないでトレンドを定義するかは分析を行う各人の裁量による。

海外では 1990 年代以降、テクニカル分析の研究は盛んになり、多数のシンプルなテクニカル分析の有効性を実証する学術論文が発表されるようになったと Campbell et al.(1997)[CLM$^+$97] は指摘している。例えば、Park and Irwin (2007)[PI07] は外国為替・先物・株式市場におけるテクニカル分析を主題とする 95 本の論文を分析し、いずれの市場においても半分以上の論文がテクニカル分析が有効であると示していたと述べている。

一方で、テクニカル分析はその有効性は実務的にも学術的にも認められつつも、主観性、恣意性が強いものであるとの指摘には十分な回答が与えられていない。そこで本研究においては、テクニカル分析の哲学・論理である価格変動パターンという観点に着目し、客観的、機械的な予測手法を開発する。すなわち、現在の価格変動パターンが過去のある

時点と似ていれば、そのパターンを用いて将来の価格の予測ができるだろうという仮説を基本とし、このような幾何学的な価格変動パターンを機械的に抽出し、予測へ活用する手法の開発を行う。

以上、本研究における実務的な意義を述べたが、学術的にも株価予測に関する先行研究は多数存在する。次章において詳細なレビューを行うが、株価予測に関する学術的な先行研究は、クロスセクション分析と時系列分析という方法論の観点、テクニカルとファンダメンタルというデータの観点、そしてパラメトリックとノンパラメトリックという手法の観点から整理できる。詳細は第 2 章で述べるが、それぞれ、クロスセクション分析とは複数企業群のある断面での属性 (例えば、時価総額など) に基づいて比較を行う方法、時系列分析はある銘柄に着目して過去の株価等を時系列データとして捉え、時系列に予測を行う方法である。データについては、株価や出来高などの市場データをテクニカル、財務、経済データをファンダメンタルといい、手法は分布を仮定するパラメトリックな手法か、分布に仮定を置かないノンパラメトリックな手法かという観点をそれぞれいう。

先行研究を踏まえた本研究の位置付けは、時系列分析の観点から、データに過去の価格変動パターンというテクニカルデータを用いた、ノンパラメトリックな手法の提案である。先行研究として、データに市場データである価格を用いた研究は多数あるのものの、本研究では、上述のテクニカル分析の枠組みでは、(C) フォーメーション分析に該当するが、(A) トレンド分析や (B) モメンタム分析に比べて価格変動パターンという定量化しづらいものを含むため、先行研究は多くない。実際、先行研究においてテクニカルをインプットとして用いた予測は、すべて終値をベースに移動平均等で加工したデータを用いており、価格変動パターンそのものではない。一方で実務的なテクニカル分析には恣意性が入り、再現性が確保できないという欠点が挙げられるが、本研究では客観的、機械的に価格変動パターンを抽出し予測に活用する。また予測手法としてもノンパラメトリックな手法を用いるが、ニューラルネットワークやサポートベクターマシーンのような目的変数の予測過程が理解しづらく、各説明変数の目的変数に対する重要度・寄与度などが評価できないモデルではなく、シンプルな k-NN 法や決定木をベースとした可読性、解釈性の高いアルゴリズムを使用する。

1.2 本研究の構成

本論文の構成は次の通りである。まず第 2 章においては、実務的なファンダメンタル分析、テクニカル分析という観点ではなく、株価予測に関する学術的な先行研究をもとに次の観点から整理を行う。株価予測の先行研究はクロスセクション分析と時系列分析という方法論の観点、テクニカルとファンダメンタルというデータの観点、そしてパラメトリックとノンパラメトリックという手法の観点があり、これらの観点に基づいて整理し、その中で本研究の位置付けを明確にする。

次に、第 3 章においては、各国の株式指数に対して、アルゴリズムにより機械的に抽出された過去の価格変動パターンが、将来の株価予測に有効であるかをシンプルな手法を用いて検証する。音声認識の分野で時系列データの類似度計測に用いてきた動的時間伸縮法 (Dynamic Time Warping;DTW[Ita75]) を、テクニカル分析の観点から指数化動的時間伸縮法 (Indexing Dynamic Time Warping;IDTW) として改良し、価格変動パターンの抽出に使用する。DTW は時系列間の長さが異なる場合でも使用でき、かつ時間方向のずれを許容するため、人間の直感にあった時系列間の類似度を計測できるといわれている。そしてノンパラメトリックな予測に手法の中で最もシンプルな k-Nearest Neighbor 法の改良である、k^*-Nearest Neighbor 法と IDTW を組み合わせた株価予測手法の提案を行う。提案手法を用いて各国の代表的な株価指数を予測したところ、過去の価格変動パターンは有効な特徴量であることが確認できた。また株価変動パターンは単純なモメンタム、リバーサル戦略以上の収益機会が存在することを合わせて実証する。

第 3 章では、価格変動パターンが似ているというシンプルな仮定で予測を行ったが、どのような価格変動パターンが予測に有効であったかは不明である。そこで第 4 章では、前章で確認した株価変動パターンの可視化を日本市場を対象に行い、どのような価格変動パターンが予測に有効であったかの可視化とその解釈を行う。しかしながら前章までと同様に、月間の日次株価変動はデータ数 (営業日) が月毎に異なるため、単純なベクトル空間上のユークリッド距離を用いたクラスタリング手法 (例えば、k-means 法) は適用できない。よって、株価を対象に異なるデータ数においても、欠損値として値の挿入および削除を行うことなく、より自然にデータ間の類似性を測定し、対応するクラスタリング方法を適切に組み合わせる必要がある。以上を踏まえて、前章の IDTW を類似度として用い

た k-medoids 法によるクラスタリング手法を提案し、それに基づいて代表的な株価変動パターンのクラスタを抽出する。

第 4 章までで、価格変動パターンが似ているというシンプルな仮定で予測を行い、具体的にどのようなパターンが有効であったかの検証を行う。しかしながら株価変動パターン以外の情報も予測に有効な特徴量として組み込むことで更なる予測精度の改善を見込むことができると考えられる。

第 5 章では、前章までに確認した予測に有効な株価変動パターンと、その他の予測に有効であろうデータを予測に利用するために組み合わせることを試みる。具体的には、モデルの可読性を考慮し、木構造に基づく予測手法である時系列勾配ブースティング木を提案する。また時系列勾配ブースティング木を構成する弱学習器として、時系列決定木を用いるが、先行研究における時系列決定木 [YSYT03] は時系列属性以外のデータを考慮できないため、時系列属性以外のデータを組み込んだ時系列決定木を用いる。各国の代表的な株式指数を対象に、単純な時系列決定木をベンチマークに、予測精度の検証を行う。

第 6 章に結論と今後の研究の方向性について総括する。

第 2 章

株価予測に関する先行研究

2.1　株価予測のための方法論

株価の予測可能性は、学術的にも実務的にも重要な研究テーマであり、伝統的なファイナンスの観点の他、様々な分野からの知見も取り入れられ、これまでに株価を予測するために様々な方法が提案されてきた。株価予測に関する学術的な先行研究を、クロスセクション分析と時系列分析という方法論の観点、テクニカルとファンダメンタルというデータの観点、そしてパラメトリックとノンパラメトリックという手法の観点から整理を行う。

はじめに株価予測の手法は大きく 2 つの方法論に大別できる (図 2.1)。1 つは、ある断面での企業群の属性等を用いてクロスセクション分析を行う方法、もう 1 つは、ある銘柄に着目して過去の株価等を時系列データとして捉え、時系列分析を行う方法である。

クロスセクション分析とは複数企業群のある断面での属性 (例えば、時価総額など) に基づいて比較を行う方法、時系列分析はある銘柄に着目して過去の株価等を時系列データとして捉え、時系列に予測を行う方法である。

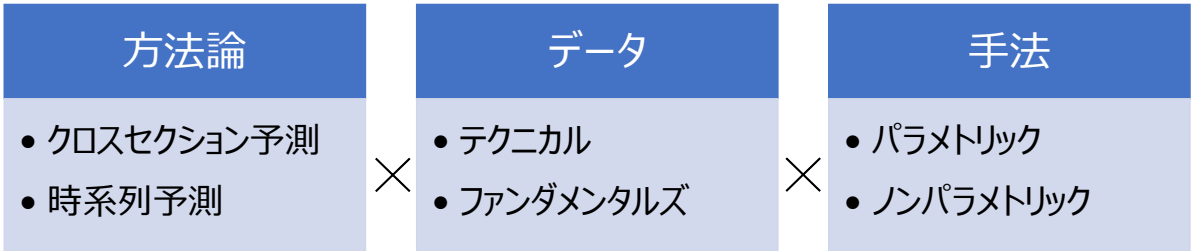

図 2.1　Method of Stock Price Prediction.

クロスセクション分析によって明らかになった株価を説明する属性ないし特徴量を「ファクター」と呼ぶ。ファイナンス分野の長年の実証研究により、クロスセクションの中でどういうファクターを持つ銘柄群が相対的に株価が上昇し、どういう銘柄群が下落するかが明らかにされてきた。その代表的な回帰モデルとして、Fama and French(1992,1993)[FF92, FF93] による 3 ファクターモデルがある。Fama-French の 3 ファクターモデルは、それ以前の代表的なモデルであった CAPM[Sha64, Lin75, Mos66]) に比べ、モデルの説明力が高いことが、米国を含む様々な株式市場において実証 [Gri02, 久保 07] されており、学術と実務の両方でスタンダードな資産価格モデルの一つとして認識されている。これ以降、Fama-French の 3 ファクターモデル以外のファクターが次々と見つけられた。その結果、Harvey et al.(2016)[HLZ16] によると、対象を一流学術誌および高い評価を得ているワーキングペーパーに絞っても、2012 年までに累計 300 を超える大量のファクターが発見された。McLean and Pontiff(2016)[MP16] は、ファクターの有効性を論文出版後という真のアウトオブサンプルでも有効かどうかを検証するため、ファイナンス、会計、経済の専門誌に掲載された 79 の論文から 97 のファクターにおいて、次の期間における予測力を比較した。

1. 元となる論文のサンプル期間 (In-Sample)
2. 元となる論文のサンプル期間後で、論文出版前の期間 (Out-of Sample)
3. 論文出版後の期間 (Post-Publication)

97 の各予測変数に対して、元の論文で高いリターンであった上位 20% を買い (ロング)、下位 20% を売る (ショート) ロング/ショートポートフォリオを構築した。その結果、ロング/ショートポートフォリオのリターンは、2. のアウトオブサンプルでは 26% 減少し、3. の論文出版後には 58% 低下することを確認した。そのためこれらファクターを用いたリターンの予測可能性は完全に消滅するという仮説は棄却されるものの、論文出版後の予測可能性は変わらないという仮説も棄却された。

一方で、一般に時系列分析により株価を予測できるかどうかについて明確な結論は出ていない。伝統的には、弱度の効率的市場仮説 (Efficient Market Hypothsis;EMH[Fam70]) により、過去のデータを分析することで将来の株価を予測することはできないと言われている。しかしながら、次章でレビューするように過去の株価等のデータを用いて、将来の株価の予測が可能であるという研究も多数存在する [YKJ05, AV09, CBS$^+$16]。

時系列予測には、EMH に基づき、どのような情報を用いて予測を行うかという「データ」の観点とどのようなアルゴリズムを用いて予測を行うかという「手法」の観点がある。

予測に利用する「データ」という観点から、EMH にはウィーク、セミストロング、ストロングという 3 つの仮説がある。ウィーク型の EMH は、将来の株価の変動が、過去の株価の変動あるいはいかなるパターンからも独立であると仮定する。つまり、過去の株価情報は、将来の株価の予測には役立たない。ウィーク型の EMH が正しいとすると、過去の価格情報を分析するいわゆるテクニカル分析は否定され、このような分析では収益を上げることができない。セミストロング型の EMH は、過去の株価情報に限らず、財務情報を含むすべての公開情報が即座に現時点の株価に反映されると仮定する。つまり、利用可能なすべての公開情報は、将来の株価の予測には役立たない。セミストロング型の EMH が正しいとすると、公開情報を元に企業価値を分析するいわゆるファンダメンタル分析は否定され、このような分析では収益を上げることができない。最後にストロング型の EMH は、公開されないインサイダー情報さえも瞬時に現時点の株価に反映されると仮定し、これが正しいとすると、いかなる情報をもってしても収益を上げることはできない。

「手法」という観点からは、パラメトリックな (特定の分布を仮定する) 時系列解析による方法とノンパラメトリックな (特定の分布を仮定しない) 方法に分類できる。時系列予測は伝統的に自己回帰 (AR) モデルに代表されるように、過去の価格の線形和で将来の価格を説明する線形モデルかつ、誤差分布に正規分布を仮定するパラメトリックなモデルから出発した。しかし、実際の金融時系列データでは非正規な挙動が数多く観測されてきたこと [Man63] から、一つの方向として、ボラティリティ[*1]に時系列構造を組み込む分散自己回帰 (ARCH) モデル [Eng82] や一般化分散自己回帰 (GARCH) モデル [Bol86] が試みられてきた。特に、GARCH モデルは様々なマーケットのボラティリティの変動をよく説明するとして、分散の変動を記述するスタンダードなモデルになった。

時系列予測の手法のもう一つの方向として、ニューラルネットワーク、決定木やサポートベクターマシンといった分布を仮定しないノンパラメトリックなモデルが株価予測に使用され始めている。これらのノンパラメトリックな手法は市場の効率性や無裁定関係などの経済学的インプリケーションを把握すること自体は目的とせず、実務的に予測精度を上げることを目的としている。こういったアプローチは、近年の計算機能力向上やそれに伴

[*1] 株価の変動率をいい、最も単純なボラティリティとして株価リターンの標準偏差を利用することが多い。

う機械学習やデータマイニングの予測力の向上等から注目が集まっており、また金融市場の分析・シミュレーションに適用する研究は実務的に関心が高く、世界中で競って研究が行われている。

2.2 ファクターに基づく予測に関するサーベイ

ファクターに基づく株価予測の方法論は、銘柄群を特定し、クロスセクションである属性 (ファクター) に注目し、その銘柄間の相対的な順序関係に基づいて株価を予測する。これらのファクターは市場データであるテクニカル、およびそれ以外の主に財務データを中心としたファンダメンタルの両方のデータを用いる。例えば、株価純資産倍率 (PBR) で測った割安度 (バリュー) に基づいて銘柄群を 3、5 または 10 分割し、最も割安な分位を買い、最も割高な分位を売るポートフォリオを構築することで投資を行う。

この方法論により、EMH の反論として 1980 年代以降に様々なファクターが主に米国株式市場において提案、検証されてきた。Basu (1977)[Bas77] は PBR の低いバリュー株ほどリターンが高いという低 PBR 効果を検出し、Banz (1981)[Ban81] は時価総額の小さい小型株ほどリターンが高いという小型株効果 (サイズ) を検出した。さらに、Basu (1983)[Bas83] は株価収益率 (PER) が低いほうが、Bhandari (1988)[Bha88] は資産負債比率 (レバレッジ比率) が高いほうがリターンが高いことをそれぞれ検証した。これらの研究を踏まえて、Fama and French (1992)[FF92] は当時発見されていたこれら 4 つの代表的ファクターであるサイズ、バリュー、レバレッジ比率、PER はサイズとバリューの 2 つに集約されることを実証的に示した。その後、Fama and French(1993)[FF93] は株価の期待収益率を市場ポートフォリオに対するリスクプレミアム (ベータ)、サイズファクター、バリューファクターの 3 つのファクターで記述する Fama-French の 3 ファクターモデルを提案し、米国株式市場において実証的にモデルの有効性を確認した。Fama-French の 3 ファクターモデルが発表された後も更なる新しいファクターが次々と提案された。代表的なファクターとして、Jegadeesh and Titman(1993,2001)[JT93, JT01] によって発見されたモメンタムファクターがある。モメンタムファクターは非常にシンプルな過去の価格に基づいて定義される。彼らはは、一定期間 (3 から 12 ヶ月) において、投資ユニバースの中で相対的にパフォーマンスの良い銘柄を買い、パフォーマンスの悪い銘柄を売るというモメンタム戦略で収益が得られることを示した。このようなモメンタ

ムファクターは Fama-French の 3 ファクターモデルでは説明されないため、モメンタムファクターを Fama-French の 3 ファクターモデルに追加した 4 ファクターモデルが Carhart (1997)[Car97] によって提案され、株式の投資信託のパフォーマンスの持続性を 4 ファクターモデルで説明できることが統計的に示されている。一方で、Daniel and Moskowitz(2016)[DM16] によると、モメンタムファクターには大幅なドローダウン現象がみられることを指摘している。実際にモメンタムファクターのリターンの分布を見ると、歪度が大きなマイナスとなり、平時ではモメンタムは正のリターンをもたらすが、あるときに大幅な負のリターンをもたらす。

モメンタムと対称的なファクターとして、Lehmann (1990)[Leh90] はリバーサルファクターを発見した。これは短期的 (1 ヶ月) には、モメンタムとは逆に、投資ユニバースの中で相対的にパフォーマンスが悪い銘柄を買い、パフォーマンスが良い銘柄を売ることで収益が得られるファクターである。

また、Ang et,al(2006)[AHXZ06] では、過去 1 ヵ月の日次リターンで計測したボラティリティが相対的に高い銘柄はリターンが低く、ボラティリティが低い銘柄はリターンが高いことを報告した。これは低リスクファクターと言われ、ハイリスク・ハイリターンというこれまでのファイナンス理論の常識と反しているため、驚きをもって広まった。同様に、配当についても伝統的には Miller and Modigliani (1961)[MM61] が、市場に摩擦がないと仮定すれば、配当方針は企業価値に影響しないはずだと述べていた。しかし経営者と株主の情報の非対称性や市場の摩擦の存在を踏まえ、一般に配当が高いほど将来の利益成長率が高まり、株価上昇に結びつくことが明らかにされた。実際に、Fama and French(1988)[FF88] は、高配当銘柄の将来リターンが長期的に高いことを実証している。また、2015 年には Fama と French 自身の手により Fama-French の 3 ファクターモデルに加え、企業の営業利益に対するファクター、すなわち収益性を示す RMW (Robust Minus Weak) と前年の企業の総資産変化率に対するファクターすなわち投資を示す CMA (Conservative Minus Aggressive) を付け加えた 5 ファクターモデル [fam15] が発表されている。

以上の研究は米国株式市場を対象としたものであるが、Fama and French(2012)[FF12] において、北米、ヨーロッパ、日本、アジア太平洋の 4 つの地域で、バリュー、サイズファクターの有効性が確認された。一方で、モメンタムファクターは日本では有効でないことも報告された。Ang et,al(2009)[AHXZ09] は米国、カナダ、英国、フランス、ドイツ、イ

タリア、日本の G7 を含む先進国市場において、低リスクファクターが有効であることを検証した。Fama and French(2017)[FF17] は、5 ファクターモデルが [FF12] と同様の北米、ヨーロッパ、日本、アジア太平洋の 4 つの地域で 3 ファクターモデルよりも株価リターンの説明力が向上していることを実証した。

さらに近年では、株式市場からもたらされた知見を株式以外の他の資産へ拡張する取り組みが行われている。Asness et,al(2013)[AMP13] は、バリューやモメンタムが株式市場以外の資産内あるいは資産間でもクロスセクション予測に有効であることを実証した。ただし、モメンタムは各資産の直近 12 か月から 2 か月までのリターンという共通の定義がされるが、バリューについては株式以外は簿価に該当するものを資産ごとに個別に定義している。Frazzini and Pedersen(2014)[FP14] は、ベータで測ったリスクが小さい資産を買い、大きい資産を売ることで収益獲得の機会が存在すること株式を含む、債券、商品、為替で確認した。彼らによると、この背景にはレバレッジ制約のある投資家は、高ベータ資産への選好があり、価格を競り上げるためであるという。Koijen et,al(2018)[KMPV18] は配当利回りを株式以外の他資産へ拡張するために、保有利得である Carry を定義し、株式や債券、為替といったあらゆる資産で Carry ファクターがクロスセクションで予測力をもつことを示した。

これらすべての研究は回帰分析の枠組み、従ってパラメトリックな分析で行われている。クロスセクション予測をノンパラメトリックに行った研究として、日本株式市場を対象に様々なファクターを深層学習を用いて組み合わせ、予測を行った Abe and Nakayama (2018)、Nakagawa et,al(2018)、Nakagawa et,al(2019)[AN18, NUA18, NIAI19] の研究がある。単純に様々なファクターを線形回帰で組み合わせるよりも、深層学習を用いて非線形に組み合わせるほうが、予測精度、収益性ともに良くなることが報告されている。

2.3 時系列予測に関するサーベイ

2.3.1 パラメトリックな予測モデル

パラメトリックな予測として代表的な手法は時系列解析に基づくものである。時系列分析は、観測可能なデータがその時点で一度しか観測できない。株価を例にあげると、株価の終値は 1 日ごとに一度だけ観測するができるが、当然ある日の終値の期待値を一点の観測データから推定することはできない。また、将来の予測を行う場合には、将来の観測点

表 2.1　Summary of representative factors

ファクター	米国市場	グローバル市場	その他資産
バリュー	Basu(1977)、Fama(1992)	Fama(2012)	Asness(2013)
サイズ	Banz(1981)、Fama(1992)	Fama(2012)	-
モメンタム	Jegadeesh(1993)	Fama(2012)	Asness(2013)
低リスク	Ang(2006)	Ang(2009)	Frazzini(2014)
配当利回り	Fama(1988)	Koijen(2018)	Koijen(2018)
収益性	Fama(2015)	Fama(2017)	-
投資	Fama(2015)	Fama(2017)	-

が得られないため、存在しない値と過去の値との相関関係を評価する必要がある。このため、将来の値を含めて予測を行うためには、分析対象の時系列自身になんらかの構造を仮定し、その構造を用いて予測を行う。

時系列分析では、観測された時系列データをある確率変数列からの一つの実現値とみなす。この確率変数列のことを確率過程 (Stochastic Process) もしくはデータ生成過程 (Data Generating Process;DGP) と呼び、時系列分析ではこの確率過程の構造のことを時系列モデルと呼ぶ。時系列分析の困難は、データがある時点において一度しか観測できないにもかかわらず、DGP を推定しなければならないという点にある。例えば、株価を例にあげると、ある企業の昨日の終値は一度だけ観測可能であるが、昨日の終値の平均的な値、つまり期待値を推定するには、株価になんらかの構造を仮定し、その構造を利用して推定する必要がある。その代表的な構造に、定常性 (Stationarity) というものがある。定常性には、弱定常性 (Weak Stationarity) と強定常性 (Strong Stationarity) の 2 つがある。弱定常性とは、過程の期待値と自己共分散が時間を通じて常に一定であることを意味し、強定常性は、任意の時点間において、過程が常に同一の同時分布をもつことを意味する [Ham94, 沖本 10]。一般的な時系列モデルでは弱定常性を仮定することが多い。このような時系列解析のモデルとしては、条件付き平均モデル、条件付き分散モデルがある。条件付き平均モデルの代表例として自己回帰 (Autoregressive;AR) モデル、移動平均 (Moving Average;MA) モデル、自己回帰移動平均 (Autoregressive Moving Average;ARMA) モデルがある。これらは株価の水準あるいは収益率のモデリングおよび予測に用いられる。AR モデルは過去の株価の線形結合で将来の株価を予測するモデルである。MA モデルは過去の株価の誤差項の線形結合で将来の株価を予測するモデルであ

る。ARMA モデルは AR と MA を組み合わせたモデルである [BJRL15]。これらは、時刻 t における予測対象を x_t として、p 次の自己回帰項 a_i と q 次の移動平均項 b_i を組合わせたモデル (ARMA(p,q)) として次式のように表現される。

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-i} + \varepsilon_t \tag{2.1}$$

ここで、ε_t は、独立同一分布に従う確率変数であり、平均 0 の正規分布に従う。パラメータの推定には、誤差項を最小化する最小二乗法を使う。ここで、L を $L^i x_t = x_{t-i}$ と定義されるラグ演算子として、x_t の d 階の階差 $(1-L)^d x_t$ が ARMA(p,q) モデルとして表現するとき、これを自己回帰和分移動平均 (ARIMA) モデルと呼び、ARIMA(p,d,q) モデルという。p,d,q の決定には AIC や BIC といった情報量基準を使用することが多い。

条件付き平均モデルは株価予測の様々な方法のベースラインとして使用される。また、株価以外の様々なデータと組み合わせ、当該データを投入することで予測精度の改善に役立つかどうかを示すために使われる。例えば、GDP 成長率や物価指数などの様々な経済指標を用いて株価を説明する研究 [CRR86, BJCS89] が多数ある。これらの研究によると、経済指標を投入することで、株価単体で予測するよりも予測精度が改善する。しかし、経済指標を用いても株価を説明することはなお難しいという指摘もある [KLW12]。

一方で、同業種同規模などの似通った銘柄間では価格差 (スプレッド) が平均回帰することが知られている。時系列解析の文脈では、Engle and Granger(1987)[EG87] によって、これを共和分性 (Co-integration) として特徴付け、様々な研究が行われている。共和分性は、簡便的にはランダムウォークのような非定常な 2 つの時系列データの線形結合が定常過程となる時系列的性質である。ペア・トレード戦略はこのような価格変動が似通った銘柄を見つけ、当該ペアのスプレッドが均衡水準の周りを推移すると仮定する。共和分性を満たすペアのスプレッドは定常過程となるため、平均が時点に依らず一定、すなわちある均衡水準への平均回帰性を持つ。そして、スプレッドが均衡水準から乖離したとき、将来その乖離が修正されるだろうという平均回帰に賭けて、相対的に割高な方を売り、割安な方を買うことで収益獲得を狙う戦略である。共和分性を利用したペアトレードの実証研究として、Gatev et,al(2006)[GGR06] の研究がある。Gatev et,al(2006)[GGR06] は 1962 年から 2002 年までの米国株式市場において共和分関係にあるペアに注目し、実際にペアトレード戦略を構築することでその収益性を検証している。その結果、 月次で 0.9% から

1.4% 程度の収益が得られており、コストを保守的に見積もっても収益機会が存在することを実証した。日本株式市場においても同様に佐藤 [佐藤 17] は、移動平均乖離率というテクニカル指標を使用したペアトレードを提案し、実証分析の結果、分析対象期間である 2002 年 1 月から 2016 年 6 月までで、Gatev et,al(2006)[GGR06] が示した方法によるペアトレードや配当込 TOPIX Core30 指数を上回る優れた収益率を確認した。

一方で、条件付き分散モデルとして分散不均一 (Autoregressive Conditional Heteroscedasticity;ARCH) モデル [Eng82] や、ARCH モデルを拡張した一般化分散不均一 (Genelarized Autoregressive Conditional Heteroscedasticity;GARCH) モデル [Bol86] が提案されている。GARCH モデルは、条件付き平均モデルを μ_t としたときに $x_t = \mu_t + u_t = \mu_t + \sigma_t \varepsilon_t$ と書けるとすると、次のように書ける。

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i u_{t-i} + \sum_{j=1}^{q} \beta_j \sigma_{t-i} \tag{2.2}$$

ただし、α_0 は定数、ε_t は平均が 0、分散が 1 の確率変数で、σ_t^2 はボラティリティである。日々のボラティリティは直接観測できないため、このようなモデルが非常に有用である。GARCH モデルには様々な拡張モデルが提案されているものの、実証的には GARCH モデルに対して有意な結果は得られず、GARCH モデルで十分であるとする研究もある [HL05]。ここまで述べてきた GARCH モデルはいずれも単変量の時系列データに対して適用されるものであったが、多変量の時系列に対して、その (動的な) 共分散ないし相関構造を記述できる GARCH モデルも存在する。例として BEKK モデル [EK95]、DCC-GARCH モデル [Eng02] などがある。

2.3.2 ノンパラメトリックな予測モデル

ボラティリティの予測は比較的良好な結果が得られるものの、株価予測のためにはパラメトリックなモデルよりもノンパラメトリックなモデルが近年多くの研究で使用されている。実際にこのアプローチにより高い予測精度を実現する手法も多くある [CBS+16]。しかしながら、ニューラルネットワークやサポートベクターマシーンのようなノンパラメトリックなモデルを用いて予測モデルを構築した場合は、目的変数の予測過程がブラックボックスとなっているため、各説明変数の目的変数に対する重要度・寄与度などが評価できないという欠点もある [FHT01]。

株価予測のためのノンパラメトリックな手法をインプットであるデータと使用されるアルゴリズムで整理する。データとしては、テクニカルとファンダメンタルの二つ、最近ではテキスト情報なども使われるようになった。

Hsu et,al(2009)[HHCH09] は、自己組織化マップ (Self-Organizing Map;SOM) とサポートベクター回帰 (Support Vector Regression;SVR) モデルを用いた 2 段階の予測アルゴリズムを提案した。入力データとしては日々の終値から作成した指数移動平均 (Exponential Moving Average;EMA) と n 日前からのリターンというテクニカル指標を使用した。彼らは、1997 年から 2002 年までの米国や日本などの 7 つの主要株式指数の日次データに対して実証分析を行い、提案した 2 段階のアルゴリズムの株価予測の精度が、SOM を用いない単一の SVR モデルと比較して、著しく向上することを示した。

Yang et,al(2009)[YHKL09] は、時系列データの局所的な特徴をとらえることのできる局所サポートベクター回帰 (Localized Support Vector Regression;LSVR) モデルを提案した。通常 SVM はすべての入力データに対して、固定されたマージンを学習する。これに対して LSVR は、ボラティリティの大きな領域には大きなマージンを、小さな領域では小さなマージンを適応的に学習する。入力データとしては n 日前からのリターンを使用し、米国株式指数であるダウ指数、SP500、NASDAQ 指数の 2004 年 4 月の日次データに対して実証分析を行い、LSVR の予測精度が SVR よりも改善することを示した。

Vanstone and Finnie(2010)[VF10] は、株価予測においては RMSE や MAE といった精度ではなく、トレーディング戦略としてのリスクやリターンなどを実際に計測する必要性を述べ、EMA などの代表的なテクニカル指標 13 個を入力としたニューラルネットワークによる個別銘柄のトレーディングシステムを構築した。2004 年から 2008 年までのオーストラリアの代表的な株式指数である ASX200 指数の構成銘柄を対象に分析を行い、ニューラルネットワークのどのようなパラメータでも、単純なバイアンドホールド戦略よりも優れた結果を報告している。

一方で、Teixeira and De Oliveira(2010)[TDO10] はニューラルネットワークのような複雑なモデルではなく、シンプルな k-NN 法に基づくトレーディング戦略を提案した。移動平均 (MA) やボリンジャーバンドなどの代表的なテクニカル指標を 22 個用いて、それらを入力とした k-NN 法により株価の予測を行う。2002 年から 2009 年までのブラジルのボベスパ指数内の 15 銘柄を用いて実証分析を行い、k-NN 法でも単純なバイアンドホールド戦略を大きく上回る収益を獲得できることを示した。

Huang(2012)[Hua12] は、生物界の進化の仕組みを模倣する最適解探索手法である遺伝的アルゴリズム (Genetic Algorithms;GA) を、SVR のパラメータおよび入力データの最適な組み合わせを探索するために使用し、これを株価予測に適用する手法を提案した。1996 年から 2010 年までの台湾株式市場の時価総額上位 200 銘柄を対象に、PBR などの代表的なファクター 14 個を入力データとし、提案手法を用いて株価予測を行うと、ベンチマークである 200 銘柄の単純平均を大幅に上回ることを実証した。

Vanstone et,al(2012)[VFH12] は PER、PBR、ROE、配当利回りという 4 つのファクターを用いて銘柄の割安度を判定するルールである Aby のフィルター [AJBEB01] を改良する手法を提案した。これは 4 つの入力をインプットにし、将来 200 日後のリターンをアウトプットとするニューラルネットワークを用いて銘柄の割安度を判定する。2004 年から 2008 年までのオーストラリアの代表的な株式指数である ASX200 指数の構成銘柄を対象に分析をおこない、単純なバイアンドホールド戦略と Aby フィルターをそのまま適用した結果を上回ることを報告した。

最近では、テクニカルやファンダメンタルといった数値データ以外のデータである、アナリストレポートやニュースといったテキストデータを入力として用いて株価を予測する試みが模索されている。テキスト情報としては主に企業が発信する情報、メディアが発信する情報、それら以外のインターネット上の情報の 3 つに分類できる。

Wang et,al(2012)[WHW12] は、企業のアニュアルレポートを用いて将来の ROE を予測する手法を提案した。この手法は、ROE をはじめに ARIMA モデルでモデル化し、その予測誤差をアニュアルレポートのテキストデータを入力にした SVR で学習し、ARIMA モデルに加える。中国と米国の代表的な企業に対して提案手法を検証すると、ベンチマークであるテキスト情報を加えない ARIMA モデルよりもテキストデータを加えた提案手法の予測精度が優れていることを確認した。

藏本ら (2013)[藏本 13] は日経新聞の記事を用いて、共起解析、主成分分析および回帰分析の 3 段階からなる CPR 法 [和泉 11] により、日本株式市場の長期的な動向を予測した。共起解析とは、単語ごとの出現頻度を単純にカウントするのではなく、単語間の隣接した共起関係をカウントする手法である。この共起関係をカウントした行列に対して主成分分析で次元を圧縮し、株価を説明変数とした回帰分析を行う。2001 年から 2010 年までの長期的な日本株式市場を対象に分析した結果、TOPIX や日経平均株価では月次予測で 60% 程度の高い予測力を記録した。

Bollen et,al(2012)[BMZ11] は、感情が個人の判断に大きく影響することから、Twitter 上の感情に関わる情報を引き出す事で、例えば株価の変動を予想出来るかどうかを検証した。具体的には、Twitter の投稿データを極性辞書を用いて、ポジティブかネガティブかと 6 種類の感情因子 (Calm, Alert, Sure, Vital, Kind, and Happy) に関する時系列データを作成する。そして、ダウ平均株価指数を対象に、過去 3 日間の株価に加え、この時系列データを入力とした自己組織化ニューラルネットワーク (Self-Organized Fuzzy Neural Network;SOFNN) を使った予想モデルを構築した。単純な過去の株価よりも感情因子 (特に Calm) を加えることで予測精度が大幅に向上することがわかった。以上のサマリーが表 2.2 である。

表 2.2　Related Works on Stock Price Prediction with Nonparametric Method

論文	データ		手法
Hsu et al. (2009)	テクニカル	株価リターン、EMA	SOM+SVR
Yang et al. (2009)	テクニカル	株価リターン	Localized SVR
Teixeira and Oliveira(2010)	テクニカル	MA など	kNN
Vanstone and Finnie (2010)	テクニカル	EMA など	ニューラルネットワーク
Huang (2012)	ファンダメンタル	PER など	GA + SVR
Vanstone et al. (2012)	ファンダメンタル	PER など	ニューラルネットワーク
Wang et al. (2012)	テキスト	アニュアルレポート	ARIMA+SVR
藏本ら (2013)	テキスト	日経新聞	CPR 法
Bollen et al. (2012)	テキスト	Twitter	SOFNN

2.4　本研究の位置づけ

以上の先行研究を踏まえて本研究の位置付けを行うと、時系列分析の観点から、データとして過去の価格変動パターンというテクニカルデータを用いたノンパラメトリックな手法を提案する。先行研究として、データに市場データである価格を用いた研究は多数あるものの、テクニカル分析のフォーメーション分析の観点から株価の変動パターンそのものに着目した研究は存在しない。実際に前節の研究においてテクニカルデータをインプットとして用いた予測は、すべて終値をベースに移動平均等で加工したデータを用いており、価格変動パターンそのものではない。一方で実務的なテクニカル分析には恣意性が入り、再現性が確保できないという欠点が挙げられるが、本研究では客観的、機械的に価格変動パターンを抽出し予測に活用する。そのため予測手法としてもノンパラメトリックな手法を用いるが、先行研究にあるニューラルネットワークやサポートベクターマシーンのような目的変数の予測過程が理解しづらく、各説明変数の目的変数に対する重要度・寄与度などが評価できないモデルではなく、シンプルな k-NN 法や決定木をベースとした可読性、解釈性の高いアルゴリズムを使用する。このようなアルゴリズムの可読性は、受託者責任に基づく説明責任が要求される機関投資家をはじめとした金融機関にとっては重要である。自らのポジションのリスクや運用結果を資金提供者である顧客に説明する必要があるからである。また金融業界以外においても、総務省は機械学習の利用の一層の増進とそれに伴うリスクの抑制のために「AI 開発ガイドライン案 [総務]」を 2017 年に策定した。このガイドライン案では、開発者は、AI システムの入出力の検証可能性および判断結果の説明可能性に留意するという透明性の原則と、開発者は、利用者を含むステークホルダに対しアカウンタビリティを果たすよう努めるというアカウンタビリティの原則が提唱されている。これらの原則は、機械学習のモデルをブラックボックス化して運用することおよびそのリスクに対して一定の歯止めをかけることを目的としていると考えられる。EU においても同様の内容が General Data Protection Regulation(GDPR[Boa]) として 2018 年より施行されている。以上から、ブラックボックスになりやすいノンパラメトリックなモデルの株価予測への適用にあたり、説明変数と予測結果の関係に関する定性的理解を与えるためのアルゴリズムである方が望ましい。本研究では、EMH とは異なり、価格変動にはある種のパターンが存在し、過去の価格変動と似た変動は繰り返し発生するであろう

と仮定する。本論文はこの仮説を基本として、その仮説に基づいた価格変動パターンを用いた投資手法の提案と各国の株式指数を用いた実証分析を行い、検証を行う。

第 3 章

価格変動パターンに基づく予測手法

3.1　はじめに

本章においては、単純な過去のパフォーマンス (一定期間のリターン) であるモメンタムとは異なる方法で、過去の価格変動が将来の株価予測に有効であることを確認する。具体的には、現在の価格変動パターンに類似した過去のパターンを探し、当該過去の変動パターンにおける将来の株価変動を用いた予測を行う (図 3.1)。

この方法はいわゆるテクニカル分析におけるフォーメーション分析として実務的によく行われる。テクニカル分析は株価のチャート上の過去パターンに基づいて取引を行う手法である。株価を予測する際に、投資家は現在の価格変動に似た過去の価格変動を参照する。例えば、予測対象の株価が上昇している場合、投資家は過去の同様の上昇局面を探す。そして、過去の上昇局面のその後の動きに基づいて、将来の株価を予測しようとする。これがフォーメーション分析の手順である。フォーメーション分析を含むテクニカル分析の欠点として、恣意性が高く主観的な分析になりがちな点である。本章で提案する手法はこの問題を改善し、システマティックに現在に類似した過去の価格パターンを抽出できる。

まず、現在の株価変動パターンに最も類似した過去の株価変動を探索する。そのために、特定期間の株価変動の間の類似度を測定する必要がある。株価の類似度を測定するために、主に音声認識の分野で使用される動的時間伸縮法 (Dynamic Time Warping;DTW[Ita75]) を採用する。DTW は長さの異なる時系列間の類似度が計測でき、さらに時系列の伸縮を許して類似度を計測するため、人が判断した時系列間の類似度と似通っているという性質を持つ。株価変動に DTW を適用するにあたり、株価の水準は時期により大きく異なるため、テクニカル分析の観点から、それを調整するための前処

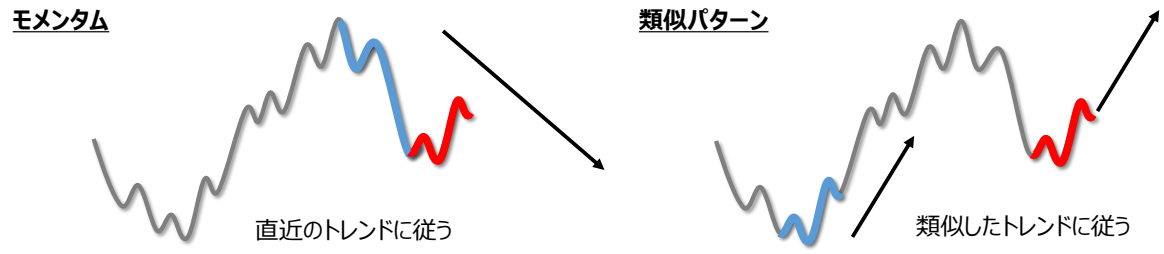

図 3.1　Difference of forecasting method between momentum and price fluctuation pattern.

理の方法を提案する (Indexing DTW;IDTW)。複数株価のチャートパターンを比較する際には指数化をするのが通常であるため、IDTW は月内の日次リターンを用いて指数化 (Indexing) した株価に対して DTW を適用し、類似度を計測する。

次に、IDTW によって計測した現在の株価変動と似た過去の株価変動に基づいて、将来の株価を予測する。シンプルに予測するには、類似した過去の株価変動の将来時点の株価は既知であるため、そのいくつかの平均値でもって予測すればよい。これは機械学習の代表なアルゴリズムである k-Nearest Neighbor(k-NN) アルゴリズムと同様の問題設定である。k-NN アルゴリズムは最もシンプルなノンパラメトリックの予測アルゴリズムの一つである。これは回帰や分類を行う際に、単純にデータと似たような過去のデータ (近傍) を k 個集め、それらの加重平均または多数決を予測値とする手法である。

しかしながら、k-NN アルゴリズムにはいくつの近傍を使用すればよいのか明確な基準がないという課題があった。この課題に対処するため、k^*-Nearest Neighbor(k^*-NN[AL16]) アルゴリズムを使用する。k^*-NN アルゴリズムは、k-NN アルゴリズムの改良手法である。最適な近傍の数である k をどのように設定するか、また回帰問題においては重み付き平均を取る際には、最適な重みを設定する方法は、長年重要な研究課題となっていた。Anava[AL16] らは、データごとに最適な重みを計算し、重みに基づき最適な近傍数 k を変えるアルゴリズムである k^*-NN 法を提案した。本章では、IDTW による類似度計測と組み合わせた k^*-NN アルゴリズム (IDTW+k^*NN) による株価変動パターンの予測力を検証する。具体的には、当月の株価変動パターンと過去の変動の類似度がデータとして、各過去の変動の翌月のリターンがラベルとして与えられる。株価変動パターンの類似度は前

述の IDTW によって計測される。次に、k^*-NN アルゴリズムにより、当月の株価変動に近い複数のラベルの加重平均を翌月の予測値とする。

提案手法の有効性を実証するため、各国の主要株価指数を用いてパフォーマンスの実証分析を行う。分析の結果は、提案方法が他の方法よりも月次の株価変動を予測するのに有効であることを示している。IDTW は、従来の DTW とその改良である Derivative DTW(DDTW[KP00]) よりも、精度および収益性の両方の点で優れていることが確認できた。さらに、Anava and Levy(2016)[AL16] らの検証と同じく、k^*-NN アルゴリズムが k-NN アルゴリズムよりも優れていることも示している。

3.2 株価変動パターンを用いた予測手法のレビュー

時系列予測の観点から株価を予測するため、様々な方法が研究されてきた。近年、機械学習やデータマイニングの手法を用いて、金融市場における予測問題や意思決定支援のためのアルゴリズムが多数提案されている。Cavalcante et,al(2016)[CBS+16] は、金融市場におけるこれらの応用についての包括的なレビューを、使用するアルゴリズムや問題設定など複数の切り口でまとめた。彼らのレビューにおいては、本章のように金融時系列を予測する方法として DTW と k-NN 法を使用した研究は存在しない。一方で、金融時系列予測以外の分野において、DTW と k-NN 法の組み合わせは単純ながら時系列を予測する手法として有効であることが知られている。例えば、時系列予測の代表的なデータセットであり、様々な分野の時系列予測の問題設定とデータが公開されている UCR[Keo] においても、ベンチマークとして DTW と k-NN 法は使用されている。また、Bagnall et,al(2017)[BLB+17] らは、実際に UCR のデータセットを用い、様々なアルゴリズムと比較した結果、DTW+k-NN を超える精度を出すアルゴリズムは少ないことを報告している。

株価予測の観点では、Coelho(2012)[Coe12] が DTW と k-NN アルゴリズムを用いて株価変動パターンを用いた株価予測を検証した。しかしながら、Coelho(2012)[Coe12] は株価変動パターンの期間や k-NN 法の近傍数を特定せず、予測に有効なパラメータが存在することを示したのみである。また、分析結果も、ある株価では有効なパラメータも他の株価では有効でなくなったりと結果が安定していない。一方で、Tsinaslanidis et,al(2014)[TK14] は時系列の重要なポイント (点) を抽出する Perceptually Important

Points (PIPs) と DTW を組み合わせた手法を提案し、本研究と同様に各国の株価指数で予測力を検証している。しかしながら、PIPs と DTW を組み合わせた手法では有効な予測ができないと結論付けており、EMH を支持する結果となっている。

本章では、k^*-NN 法によって近傍数を特定するとともに、各国の主要株価指数を複数用い、月次予測において収益獲得が可能であることを示す。

3.3 提案手法:k^*-Nearest Neighbors with Indexing Dynamic Time Warping

提案手法は、まず現在の株価変動に類似した過去の株価変動を探索する。そのためには株価変動の類似性を計測する必要があり、類似度計測のために IDTW を用いる。次に、過去の類似した株価変動に基づいて将来の株価を予測するため、k^*-NN 法を用いる。以下では、DTW とその改良である DDTW をレビューする。次に、株価変動に対して DTW を適用する際に、テクニカル分析の観点から、時系列データを指数化した IDTW を提案する。最後に k-NN アルゴリズムおよびその改良手法である k^*-NN アルゴリズムについて概観する。

3.3.1 Dynamic Time Warping

DTW は、時系列データ間の類似度を計測するための手法である。当然ながら DTW 以外にも時系列データ間の類似度を計測するための多くの方法が提案されている。例えば、相関係数やユークリッド距離は、単純な類似度計測手法としてよく使われる。しかし、金融時系列データ間の類似度を計測するにあたり次のような欠点がある。相関係数もユークリッド距離も時間軸の伸縮を考えず、特に、ユークリッド距離は、時間軸方向に少しでもズレが生じると類似度が大幅に低下する点がある [Ita75]。さらに、両者の最も重要な問題点として、2 つの時系列の長さが異なる場合には使用できない。

DTW 法は、これらの問題を解決できる類似度計測手法であり、時系列データにおける 1 点を、もう一方の時系列データにおける複数点のデータに対応付ける。つまり時間軸の非線形な伸縮を許容した類似度を計測する。DTW は時系列データを伸縮させてすべての点同士の距離を測ることで、最も距離が近い点同士が対応付けられる。従って、DTW は時間軸方向にズレがあっても問題なく距離を求めることができる。この特性により、

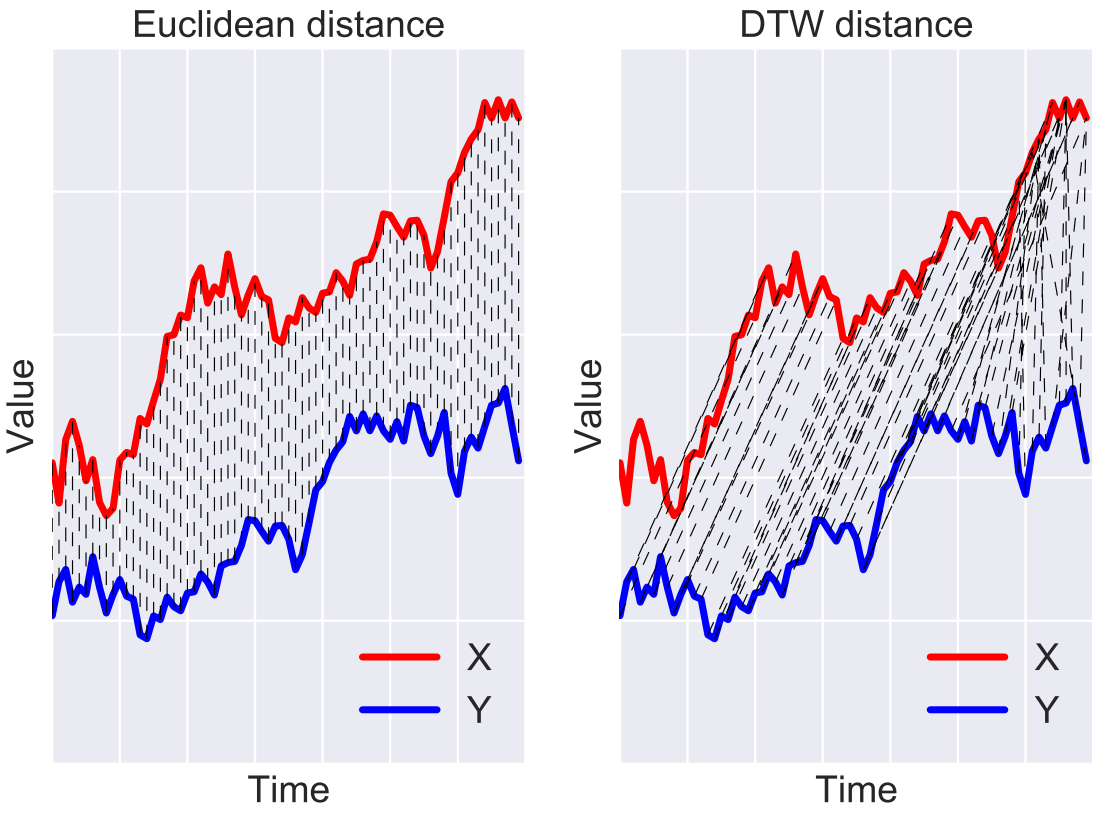

図 3.2　Correspondence of time-series data.

DTW により計測される類似度は人間の直感に合っているといわれ、時系列データの類似度計測において、周波数ではなく形状に着目する場合には様々な分野で DTW が適用されてきた。図 3.2 は、ユークリッド距離と DTW 法に基づく時系列データ類似度比較の例を示している。

ここからは具体的に DTW 法のアルゴリズムを概観する。DTW 法は、Algorithm1 によって時系列 x と y の類似度を計測する。長さが N, M の 2 つの時系列データ $x = (x[1], x[2], ..., x[i], ..., x[N])$ と $y = (y[1], y[2], \cdots, y[j], ..., y[M])$ に対して、2 つの点 $x[i], y[j]$ 間の距離 $d(x[i], y[j])$ を (i, j) 要素の値としてもつ $N \times M$ の距離行列を作成する。距離 $d(x_i, y_j)$ は通常、ユークリッド距離またはマンハッタン距離 (絶対値距離) を用いるが、非負であればどんな関数でもよい。本論文では、外れ値に対する耐性を考慮して、ユークリッド距離ではなく絶対値距離 $d(x[i], y[j]) = |x[i] - y[j]|$ を用いる。

次にワーピングパス $W = (w[1], w[2], ..., w[k], ..., w[K])$ を求める。ここで K は、$max(m, n) \leq K < m + n - 1$ を満たす整数であり、ワーピングパスの個数である。ワーピングパスとは 2 つの時系列データの対応付けのことであり、次の 3 つの条件を満た

す行列の「要素の順列」で表される。図 3.2 の右図にワーピングパスの例を点線にて示している。

- 境界条件: $w[1] = (1,1), w[K] = (N,M)$ これは、ワーピングパスを行列の左下の要素から開始し、右上で終了させることを意味する。
- 連続性: $w[k] = (a,b), w[k-1] = (c,d)$ とすると、$a - c \leq 1$ かつ $b - d \leq 1$ これにより、ワーピングパスの移動先が隣接する要素に制限される。
- 単調性: $w[k] = (a,b), w[k-1] = (c,d)$ とすると、$a - c \geq 0$ かつ $b - d \geq 0$ これにより、ワーピングパスは終点へと戻らずに単調に進む。

つまり、ワーピングパスは距離行列の要素 $(1,1)$ からはじめて、隣接する右・上・右上の要素を辿って終点の (N,M) へ到達する任意のルートを表したものであると解釈できる。

上記の 3 条件の下で得られるワーピングパスは非常に多く存在するものの、それぞれのワーピングパスの要素に対する距離行列の値であるワーピングコストと呼ばれるものが存在する。あるワーピングパス W のワーピングコスト $C(W)$ とはワーピングパスの要素 $w[k] = (a,b)$ と対応する距離行列の値 $c[k] = |x[a] - y[b]|$ の和である。つまり、

$$C(W) = \sum_{k=1}^{K} c[k] \tag{3.1}$$

この式 (3.1) から得られるワーピングコストの中で最小のものが DTW により計測した 2 つの時系列データの類似度となる。

$$DTW(x,y) := \min_{W} C(W) \tag{3.2}$$

ワーピングパスは非常に多く存在するため、全てのワーピングパスに対し、ワーピングコストを計算して最小値を求めることは計算量の観点から現実的ではない。そこで動的計画法 [Bel54] により以下の再帰関数を利用して DTW を計算する。この再帰関数を利用すると式 (3.1) は累積コスト $\gamma(i,j)$ によって以下のように計算することができる。ただし、$\gamma(0,0) = 0, \gamma(i,0) = \gamma(0,j) = \infty$ とする。

$$\gamma(i,j) = d(x[i], y[j]) + \min \begin{cases} \gamma(i-1,j) \\ \gamma(i,j-1) \\ \gamma(i-1,j-1) \end{cases} \tag{3.3}$$

Algorithm 1 DTW distance

procedure DTW(x, y)

▷ Initialize matrix D

  Var $D[N+1, M+1]$
  $D[1,1] = 0$
  for $i = 2$ to $N+1$ **do**
    for $j = 2$ to $M+1$ **do**
      $D[i,j] = \infty$
    end for
  end for

▷ Calculate DTW distance

  for $i = 2$ to $N+1$ **do**
    for $j = 2$ to $M+1$ **do**
      $D[i,j] = d(x[i-1], y_[j-1])$
        $+min(D[i,j-1], D[i-1,j], D[i-1,j-1])$
    end for
  end for
  return $D[N+1, M+1]$

end procedure

以上を踏まえ、最終的なアルゴリズムは Algorithm 1 の通りである。

ここで、次の 2 つの系列 a と b を用いて、Algorithm 1 に基づき類似度を計算する。

$$a = \{0, 59, 95, 95, 59, 0, -59, -95, -95, -59\} \tag{3.4}$$
$$b = \{59, 95, 95, 59, 0, -59, -95, -95, -59, 0\} \tag{3.5}$$

図 3.3 の通り、この 2 つの系列は位相が異なるのみであり、形状は似ている。まず単純な絶対値距離では類似度は、

$$\sum_{i=1}^{10} |a[i] - b[i]| = 380 \tag{3.6}$$

となる。次に Algorithm 1 に基づき DTW による類似度を計算する。DTW の計算に必要な Algorithm 1 にある行列 D が図 3.4 である。赤字が最小のワーピングパスを表している。

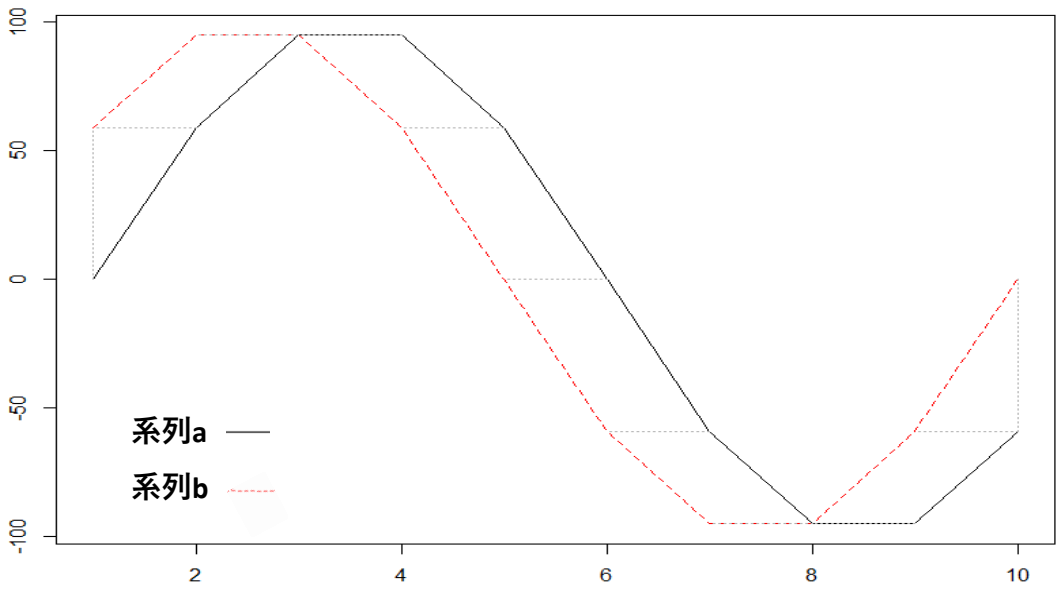

図 3.3　Example of DTW similarity.

∞	734	770	806	662	367	131	95	95	**59**	**118**
∞	616	652	688	544	308	131	59	**59**	95	190
∞	462	498	534	390	213	95	**59**	59	95	190
∞	308	344	344	236	118	**59**	95	131	131	190
∞	190	190	190	118	**59**	118	213	308	367	367
∞	131	95	95	**59**	118	236	390	544	662	721
∞	131	59	**59**	95	190	344	534	688	806	865
∞	95	**59**	59	95	190	344	498	652	770	770
∞	**59**	95	131	131	190	308	462	616	675	675
∞	**59**	154	249	308	308	367	462	557	616	616
0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞

系列a

系列b

図 3.4　Cost matrix of DTW.

例えば、2 行 2 列目の要素 $D[2,2]$ は、

$$d(a[1]-b[1])+\min(D[2,1],D[1,2],D[1,1])=|0-59|+\min(\infty,\infty,0)=59 \qquad (3.7)$$

のように計算でき、DTW による類似度は $D[11,11]=118$ となり、絶対値距離の 380 よりも小さい値を示している。この例のような位相が異なるのみで形状が類似している系列の類似度は、絶対値距離よりも DTW のほうが小さく、DTW はより形状に基づいて類似

Algorithm 2 DDTW distance

1: **procedure** DDTW(x, y)

▷ Initialize vectors

2: **Var** Dx, Dy

3: **for** $i = 2$ to $N - 1$ **do**

4: $Dx[i] = \frac{(x[i]-x[i-1])+((x[i+1]-x[i-1])/2)}{2}$

5: **end for**

6: **for** $j = 2$ to $M - 1$ **do**

7: $Dy[j] = \frac{(y[j]-y[j-1])+((y[j+1]-y[j-1])/2)}{2}$

8: **end for**

▷ Apply DTW

9: **return** $DTW(Dx, Dy)$

10: **end procedure**

度を計算しているといえる。

DTW 法の欠点は、その計算コストが $O(MN)$ のオーダーと、比較的大きいことである。様々な工夫により、効率的な計算方法が提案されている [KP00] ものの、本研究で扱う株価のデータ規模では、コンピュータの性能自体の向上により計算量は問題にならない場合が多い。

3.3.2 Derivative Dynamic Time Warping

データの個々の値を元に時系列間の距離を定義しており、距離が近い時系列データ同士でも、その形状はかならずしも類似しているわけではない。例えば、時系列データのある区間において値が上昇し、もう一方の区間では下落している場合でも、その区間において値に差がなければ、DTW では類似していると判断する。したがって、形状に着目するために値を何らかの形で正規化する必要がある。このような問題意識から、Keogh (2000)[KP00] は時系列データに DTW を直接適用するのではなく、時系列データの変分に対して DTW を適用する Derivative DTW(DDTW) を提案した。Algorithm2 は DDTW の概要を示す。ここで、D_x と D_y は元の時系列データの変分ないし微分である。

図 3.5 は、ある時点間の TOPIX 指数に対して DDTW を適用した場合の結果を示している。DDTW は DTW に比べ形状を意識した類似度を計測できるという利点があるが、株価のような金融時系列に対しては、より適切な基準化の方法が考えられる。

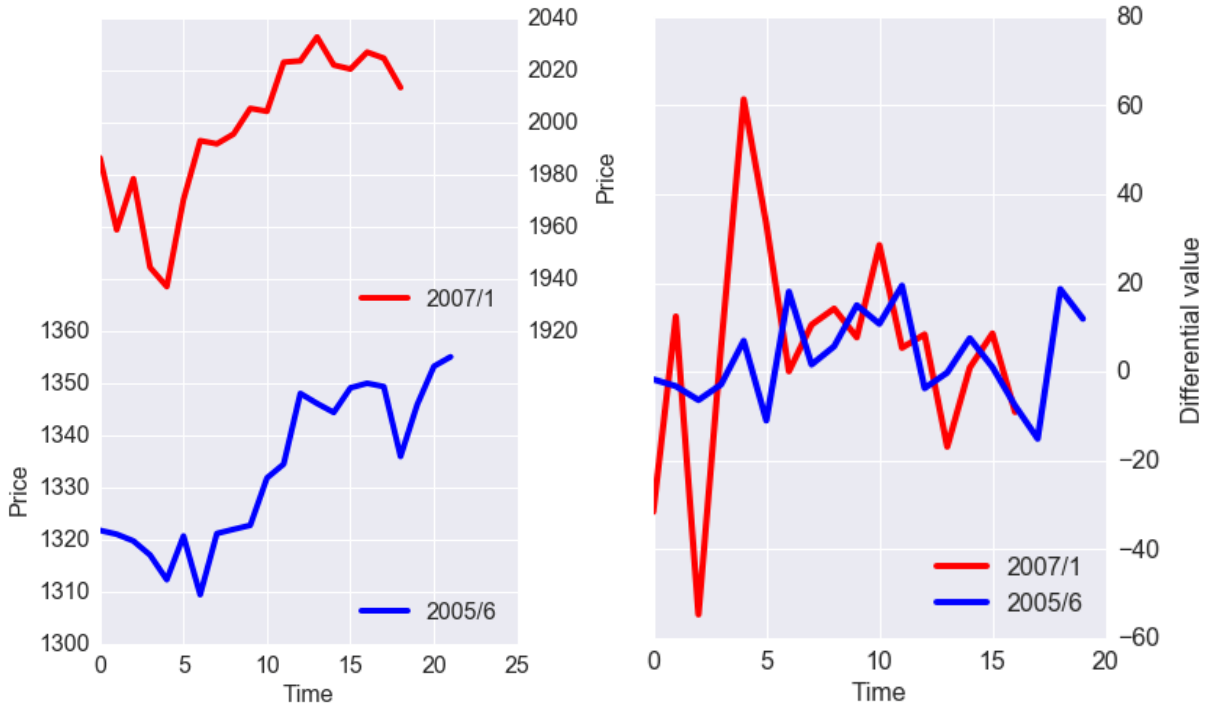

図 3.5　Conversion of DDTW for TOPIX data.

3.3.3　Indexing Dynamic Time Warping

株価変動に DTW を適用するにあたり、株価の水準は時期により大きく異なるため、テクニカル分析の観点から、それを調整するための前処理の方法を提案する。しかし基準化にあたって考慮すべきもう一つの重要な問題がある。それは株価の測定期間をどうするかという問題である。

この点に関連して、株価の季節性は投資家によって広く認識されている。また株式市場のアノマリーとして季節性の議論は多数の先行研究がある。とりわけ「月」に着目して数々提示されてきた。例えば、1 月の株高 (1 月効果) は最も有名なものの一つである。米国市場では第二次世界大戦以前から知られている。また 5 月から秋口までの株式低パフォーマンス (Sell in May、またはハロウィーン効果) も良く知られている。このような株価の季節性は [BJ02, ACF13] によって検証され、1964 年まで遡っても効果が確認でき、世界各国の市場でも同様の効果が確認できると指摘し普遍性を主張している。以上のように月という単位によって指摘されている季節性は様々だが、その発生要因については未解明なものが少なくない。また実務的にも、毎月のリターンは、ファンドや投資家にとって

Algorithm 3 IDTW distance

1: **procedure** IDTW(x, y)

▷ Scaling data

2: **Var** Ix, Iy ▷ Initialize Ix,Iy

3: $Ix[1] = 1, Iy[1] = 1$

4: **for** $i = 2$ to N **do**

5: $Ix[i] = Ix[i-1]\frac{x[i]}{x[i-1]}$

6: **end for**

7: **for** $j = 2$ to M **do**

8: $Iy[j] = Iy[j-1]\frac{y[j]}{y[j-1]}$

9: **end for**

▷ Apply DTW

10: **return** $DTW(Ix, Iy)$

11: **end procedure**

基本的な評価単位となっている。換言すれば、投資家は株価の変動を「月」単位で捉えることが多い。

したがって本研究では、日々の株価の終値で構成された毎月の株価変動を一つのパターンとして捉える。しかしながら毎月の営業日は、年や月によって異なるため、時系列データの長さは一定ではなくなる点に注意が必要である。

計測基準を確定させたところで議論を基準化に戻すと、投資家は株価の差分にほとんど注意を払わず、株価の形状がより重要であると考える。すなわち、実際に 2 つの株価変動を目視で比較するときには、2 つの株価の収益率の変動を比較するのではなく、始点における値を統一した指数化して比較するのが自然であると考えられる。一般的にテクニカル分析において、複数の株価系列を比較する際には指数化を行う。IDTW の背後には、投資家にとって重要な株価変動のパターンは指数化されたものであるという仮定がある。以上から、IDTW の Algorithm3 は、このようなパターンの変動を捉えるために前処理として前月末の値を基準に指数化を施す。ここで、I_x および I_y は元の時系列データの指数化した値である。

図 3.6 は、ある時点間の TOPIX 指数に対して IDTW を適用した場合の結果を示している。

図 3.6　Conversion of IDTW for TOPIX data.

3.3.4　k-Nearest Neighbors

k-NN アルゴリズムは、教師あり学習の手法のなかで最も単純なアルゴリズムの一つであり、分類問題、回帰問題の両方に対して、近傍数 k と各データ間の類似度を測るメジャーがあれば使用できる。k-NN アルゴリズムは、直感的には特徴空間中でテストデータに最も似ているデータを探し、それに基づいて判断する。まず、n 個のデータ $x_1, ..., x_n \in \mathbb{R}^d$ とそれらに対応するラベル $y_1, .., y_n \in \mathbb{R}$ とする回帰問題を考える。新しいデータ x_0 に対して、そのラベルを予想するため、x_0 に最も類似したデータを n 個のうちから k 個見つけ、これら k 個の平均あるいは重み付け平均 (分類問題の場合には多数決) を取る [CH+67]。その値が x_0 に対する予測値となる (Algorithm 4[BC94])。x_0 と類似したデータを探すためには一般には各データ間のユークリッド距離を用いられることが多い。

$$d_i = ||x_i - x_0|| \tag{3.8}$$

このように k-NN アルゴリズムは記憶に基づく方法であり、学習するモデルを必要とせず、そのため怠惰学習と呼ばれる。$k = 1$ とした場合の k-NN アルゴリズムは特に最近傍

Algorithm 4 k-NN

1: **procedure** k-NN(d, y, k)

▷ Sort the labels y by distance d

2: Sort y by the distances d in ascending order

▷ Calculate average value of y

3: **return** $\frac{1}{k}\sum_{i=1}^{k} y[i]$

4: **end procedure**

法 [FHJ51] と言われている。k-NN アルゴリズムは非常に単純な方法にもかかわらず、その予測精度は手書き文字認識や衛星画像、心電図パターンなどの多くの分類問題で非常に良好であると報告されている [MST+94, FHT01]。

一方で、k-NN アルゴリズムは、新しいデータが与えられたとき、近傍点を見つけるため、既存の全データとの類似度を計算する必要がある。そのため、大量のデータが存在する場合には、各データの類似度の計算とその保持に大量のメモリを消費し、計算時間も多くかかってしまうという欠点がある。データが多い場合には近似するアルゴリズムがいくつか提案されている [HW98, PQS04]。

3.3.5 k^*-Nearest Neighbors

k-NN アルゴリズムにおいて、最適な近傍の数である k をどのように設定するか、また回帰問題において最適な重みを設定する方法は、長年重要な研究課題となっていた。ここで、Anava and Levy(2016)[AL16] による k^*-NN アルゴリズムを概観する。先ほどの k-NN アルゴリズムと同様に、n 個のデータ $x_1, ..., x_n \in \mathbb{R}^d$ とそれらに対応するラベル $y_1, .., y_n \in \mathbb{R}$ を所与とする。そして、任意の $i \in \{1, ..., n\} = [n]$ について、$y_i = f(x_i) + \epsilon_i$ が成立し、$f(\cdot)$ および ϵ_i は以下の条件を満たすと仮定する。

(1) $f(\cdot)$ はリプシッツ連続である。すなわち任意の $x, y \in \mathbb{R}^d$ に対して、$|f(x) - f(y)| \leq L \times d(x, y)$ が成立する。ここで、$d(\cdot, \cdot)$ は適当な非負関数で、本研究では DTW、DDTW または IDTW である。また、$L \in \mathbb{R}$ はリプシッツ定数と呼ばれる実数である。

(2) ϵ_i のノイズは有界である。すなわち、任意の $i \in [n]$ に対して、$E[\epsilon_i | x_i] = 0$ であり、ある定数 $b > 0$ で $|\epsilon_i| \leq b$ とすることができる。加えて、任意のデータ x_i とノイズ

項 ϵ_i は独立であると仮定する。

以上の仮定のもと、新たなデータ x_0 が与えられたとき、関数 $f(x_0)$ を推定することが目標である。ここで、k^*-NN 法の推定量 $\hat{f}(x_0)$ は、$\hat{f}(x_0) = \sum_{i=1}^{k} \alpha_i y_i$ である、つまりノイズのある複数ラベルの加重平均の形に限定する。すると、目標は次の通り、予測値と真の値 $f(x_0)$ との絶対値を最小にすることとなる。

$$\arg\min_{\boldsymbol{\alpha}\in\Delta_n} |\sum_{i=1}^{n} \alpha_i y_i - f(x_0)| \tag{P1}$$

最小化は単体 $\Delta_n = \{\boldsymbol{\alpha} \in \mathbb{R}^n | \alpha_i > 0, \sum_{i=1}^{n} \alpha_i = 1\}$ 上で行われる。ここで、(P1) 式を緩和し、上から抑えた

$$|\sum_{i=1}^{n} \alpha_i y_i - f(x_0)| \leq |\sum_{i=1}^{n} \alpha_i \epsilon_i| + L \sum_{i=1}^{n} \alpha_i d(x_i, x_0).$$

を考える。

ここで、$|\sum_{i=1}^{n} \alpha_i \epsilon_i|$ に対して、次の Hoeffding の不等式から、

$$P(|\sum_{i=1}^{n} \epsilon_i - \sum_{i=1}^{n} E[\epsilon_i]| > \varepsilon) \leq 2e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(U_i - L_i)^2}}$$

$\epsilon_i \in [U_i, L_i]$ であり、$\sum_{i=1}^{n}(U_i - L_i) = b$ とおくと、$C = b\sqrt{\frac{1}{2}\log(\frac{2}{\delta})}$ に対して、$|\sum_{i=1}^{n} \alpha_i \epsilon_i| < C\|\boldsymbol{\alpha}\|_2$ が確率 $1-\delta$ で成立する[*1]。ゆえに、(P1) 式を高い確率で上から抑えることが保証される新しい最適化問題 (P2) を得る。

$$\arg\min_{\boldsymbol{\alpha}\in\Delta_n} |C\|\boldsymbol{\alpha}\|_2 + L\sum_{i=1}^{n} \alpha_i d(x_i, x_0)| \tag{P2}$$

一般性を失うことなく、データ x_n が x_0 からの距離に従って昇順に、すなわち $d(x_1, x_0) < d(x_2, x_0) < ... < d(x_n, x_0)$ と並べられているとする。また $\boldsymbol{\beta} \in \mathbb{R}^n$ は $\beta_i = Ld(x_i, x_0)/C$ のベクトルである。以上の条件のもと Karush-Kuhn-Tucker(KKT) 条件によれば、(P2) の最適解 α^* は次のように求めることができる。

$$\alpha_i^* = \frac{(\lambda - \beta_i) \times \mathbf{1}\{\beta_i < \lambda\}}{\sum_{i=0}^{n}(\lambda - \beta_i) \times \mathbf{1}\{\beta_i < \lambda\}}$$

[*1] 証明は Hoeffding の不等式において、ε に C を代入することで得られる。

Algorithm 5 k^*-NN

procedure k^*-NN$(\beta, y, L/C)$

▷ Initialize λ and k

  Var $\lambda[N]$

  $k = 0, \lambda[0] = 0$

  $\beta = L/C \times \beta$

  while $\lambda[k] > \beta[k+1]$ And $k \leq N-1$ **do**

    $k = k + 1$

    $\lambda[k] = \frac{1}{k}(\sum_{i=0}^{k} \beta[i] + \sqrt{k + (\sum_{i=0}^{k} \beta[i])^2 - k \sum_{i=0}^{k} \beta[i]^2})$

  end while

▷ Calculate weight parameter α

  Var $\alpha[N]$

  for $i = 1$ to N **do**

    $\alpha[i] = \frac{(\lambda[k]-\beta[i]) \times \mathbf{1}\{\beta[i]<\lambda[k]\}}{\sum_{i=0}^{N}(\lambda[k]-\beta[i]) \times \mathbf{1}\{\beta[i]<\lambda[k]\}}$

  end for

  return $\sum_{i=0}^{N} \alpha[i] \times y[i]$

end procedure

ここで、$\mathbf{1}\{A\}$ は指示関数であり、A が真の場合に 1、偽の場合に 0 をとる関数である。また λ は Algorithm 5 で定義される数である。Anava and Levy(2016)[AL16] は、Algorithm 5 が (P2) の解を求めることに対応することを証明した。また、k^*-NN アルゴリズムでは、k 近傍以外の α^* はゼロであり、逆に非ゼロの α^* の個数が k 近傍である。ここで、最適な重みは、単一のパラメータ L/C、Lipschitz 連続性の定数対雑音比に依存する。ノイズが低下する、すなわち L/C が増加すると、必要な近傍数 k^* は減少する。逆にノイズが多い場合、すなわち L/C が低下すると、必要な近傍数 k^* は増えるという直感に合った結果となっている。

実際には L/C はクロスバリデーション (CV) によって決定する。CV[Sto74] は、分類と回帰スキームのアルゴリズムの汎化性能を評価するための最も広く使用されている方法の 1 つである。しかし、今回のような時系列予測に関しては、非定常性およびデータの系列相関の可能性が考えられる場合、CV がそのような問題を考慮していないため、問題になることを Bergmeir and Benitez(2012)[BB12] は示している。また、直感的には過去の時系列データを予測するために将来の時系列データを使用すべきでない。以上の観点から

表 3.1　Statistics of each indices in all periods

	CAC	DAX	FTSE	SPX	TPX
年率リターン [%]	6.39	10.12	6.41	9.09	0.87
年率リスク [%]	21.71	22.43	17.36	17.47	21.12
年率シャープレシオ	0.29	0.45	0.37	0.52	0.04
歪度	0.06	-0.08	-0.01	-0.09	-0.01
尖度	7.85	8.56	9.09	12.09	8.96
Jarque-Beta 統計量	7,125.47	9,349.49	11,197.89	24,925.44	10,444.31
p 値	0.000	0.000	0.000	0.000	0.000

本研究では、k-NN および k^*-NN のパラメータ k と L/C を決定するために、CV の代わりにアウトオブサンプル評価により決定する。

3.4　実証分析

3.4.1　分析手順

提案手法である IDTW によって計測した過去の株価変動パターンを用いた k^*-NN による予測の有効性を示すため、DTW、DDTW、IDTW を k-NN と k^*-NN を組み合わせた 6 つの手法の比較を行う。それぞれ DTW+kNN、DDTW+kNN、IDTW+kNN、DTW+k^*NN、DDTW+k^*NN および IDTW+k^*NN と表記する。DTW+kNN は金融市場において先行研究である Coelho(2012)[Coe12] で検証されている。

予測対象は TOPIX(TPX)、S&P500(SPX)、FTSE100(FTSE)、DAX30(DAX)、CAC40(CAC) の月次の騰落とする。これらは日、米、独、英、仏それぞれの国における代表的な株価指数であり、多くの投資家に利用されている。価格データは金融情報端末である Bloomberg[*2]から取得した。すべて指数のデータ期間は、1989 年 1 月初から 2017 年 8 月末までとした。各指数の全期間における統計量は表 3.1 の通りである。表 3.1 から、当該データ期間においては SPX がシャープレシオで測って最もパフォーマンスが良い。また、CAC が金融資産としては珍しく正の歪度を持ち、全ての指数について Jarque-Bera 検定によって 1% 水準以下で正規性は棄却されている。

データ期間のうち、1989 年 1 月から 2005 年 12 月までのデータを k-NN および k^*-NN

[*2] Bloomberg のティッカーコードはそれぞれ TPX、SPX 、UKX、DAX、CAC Index。

の参照期間として利用し、リーマンショックを含めた過去 10 年間の長期間を検証するためにテスト期間を 2006 年 1 月から 2017 年 8 月までとする。

検証にあたり、はじめに DTW、DDTW、および IDTW を使用して、参照期間のデータから 1 ヶ月毎に株価変動の類似度を計測する。次に、計測した類似度を用いて k-NN または k^*-NN により翌月の騰落を予測し、上昇予測ならば月末の価格で買い、下落予測ならば月末の価格で売りというシミュレーションを毎月繰り返した。なお、今回は参照期間は毎月追加していき、古い期間の削除は行なわなかった。翌月の株価の予測に使用する k-NN と k^*-NN のパラメータを決定するために、前月までの過去 36 ヶ月の予測結果から、最も正答率の高いパラメータを毎月選択した[*3]。具体的な手順は以下のとおりである。

Step 1. 類似度ベクトル β を DTW、DDTW、IDTW それぞれを用いて計測する。β は当月 t と当月以前の過去すべての月間の価格変動との類似度を昇順に並べたものである (図 3.7 上段)。例えば、t 月が 2005 年 12 月とすると、2005 年 12 月の日々の株価と 2005 年 12 月以前の全ての月内の日々の株価との間の類似度を計算する。

Step 2. Step 1 で計算した類似度と、各月の翌月のリターン (各月のその翌月の株価/各月の株価-1) をラベル y として、k-NN と k^*-NN によって翌月 $t+1$ の騰落を予測する (図 3.7 中段)。ここでハイパーパラメータである k と L/C の範囲は先行研究である Anava ら [AL16] と同じ範囲で、過去 36 ヶ月で最も正答率の高いパラメータを選択し、使用する。

Step 3. 翌月の予測リターンが正の場合は、月末の価格で各指数を 1 単位を買う。負の場合は、1 単位売り、翌月の収益を計算する (図 3.7 下段)。なお、ポジションは持ち越さずに翌月末時点で解消する。

Step 4. 月を進めて $t = t+1$ とし、Step 1. に戻る。

[*3] パラメータ k は 1 から 10 までの範囲で、パラメータ L/C は $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ の範囲とした。これらは先行研究である Anava ら [AL16] と同じ設定である。また 36 ヵ月という期間は分析開始前に決定している。なお、本章にて用いた k^*-NN の R 言語および C++ 言語での実装を CRAN にて ksNN パッケージ [KS19] として公開している。

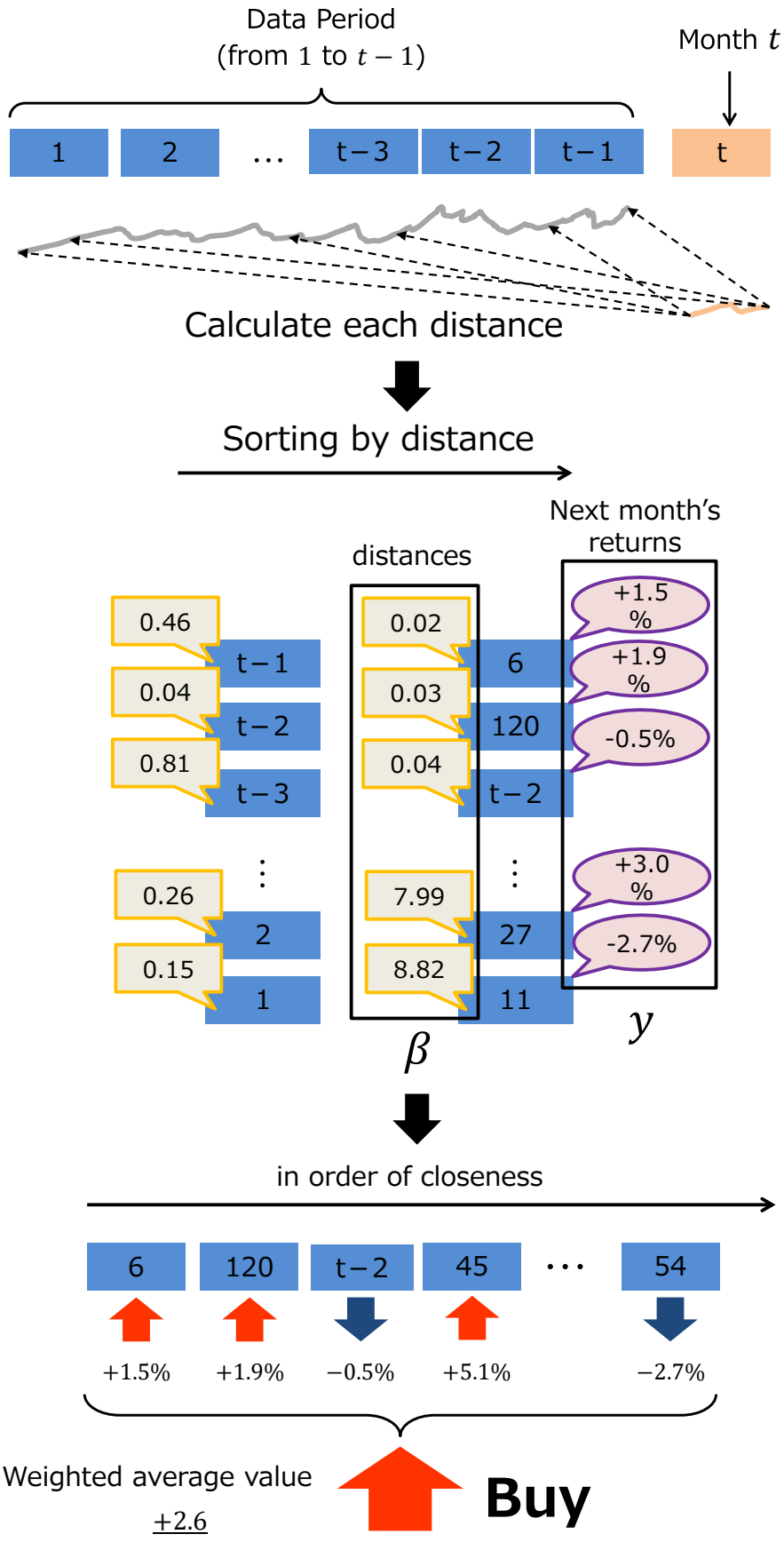

図 3.7　Stock price prediction framework.

3.4.2 分析結果

分析の結果は各月の騰落の予測精度および合計リターン (収益性) の両方で評価する。予測精度は平均絶対誤差 (Mean Absolute Error;MAE)、平均平方二乗誤差 (Root Mean Square Error;RMSE)、正答率の 3 つの指標で計測する。

時点 $t(1 \leq t \leq T)$ における予測値を $f(t)$、実際の値を y_t としたとき、MAE と RMSE はそれぞれ次のように定義される。

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - f(t)| \tag{3.9}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - f(t))^2} \tag{3.10}$$

正答率は単純に、テスト期間のデータが n 個のうち、予測リターンと実際のリターンの符号が同一であったデータを a として、a/n で定義される。

まず予測精度について、表 3.2、3.3、3.4 はそれぞれ 6 つの手法のすべての月の平均の MAE、RMSE および正答率を表す。右端の列は、すべての株式指数の正答率の平均値である。太字は 6 つの方法のうち最も良い手法を示している。この表からわかるのは提案手法である IDTW+k^*-NN がすべての指数において最も予測精度が高いことである。MAE と RMSE で測って最も良い予測精度であり、またすべての指数の平均値も 60% 以上と高い正答率となっている。

この予測力は、予測期間は異なるものの本研究と同じく TOPIX を対象に 10 年程度の月次の騰落を株価と新聞記事などのテキスト情報を使用し予測した、蔵元ら (2013)[藏本 13] の報告と同程度である。蔵元ら (2013)[藏本 13] は日経新聞などのテキスト情報が追加で必要となるものの、提案方法は株価情報のみで予測できる点が重要な差異である。

各手法について個別にみていくと、各株式指数についてはばらつきがあるものの、全指数の平均について、IDTW+k-NN は DTW+k-NN と DDTW+k-NN より予測精度が高い。同様に、IDTW+k^*-NN は、DTW+k^*NN と DDTW+k^*-NN よりも予測精度が高い。従って、IDTW は DTW と DDTW よりも予測精度の点で上回っているといえる。同様に、k^*-NN は k-NN を予測精度の点で上回る。また、金融時系列予測以外の分野において、DTW と k-NN 法の組み合わせは単純ながら時系列を予測する手法として有効である

表 3.2　The average MAEs of all years for each method. The out-of-sample period is from January 2006 to August 2017. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the six methods.

		CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
DTW	k-NN	9.014	8.975	5.574	7.705	8.873	8.028
	k^*-NN	5.220	5.636	3.824	3.646	5.123	4.690
DDTW	k-NN	11.008	10.729	7.649	8.731	10.058	9.635
	k^*-NN	4.501	5.068	3.570	3.252	4.694	4.217
IDTW	k-NN	8.204	12.822	7.183	9.296	10.761	9.653
	k^*-NN	**3.727**	**4.005**	**2.905**	**2.948**	**4.036**	**3.524**

表 3.3　The average RMSEs of all years for each method. The out-of-sample period is from January 2006 to August 2017. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the six methods.

		CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
DTW	k-NN	11.810	11.735	7.242	10.028	11.472	10.458
	k^*-NN	6.697	7.273	5.057	4.807	6.554	6.078
DDTW	k-NN	14.049	13.740	10.105	11.078	12.245	12.243
	k^*-NN	5.942	6.811	4.703	4.552	6.049	5.611
IDTW	k-NN	11.265	16.313	9.681	11.725	14.240	12.645
	k^*-NN	**4.984**	**5.387**	**3.843**	**3.985**	**5.292**	**4.698**

ことが知られているが [BLB+17]、そのままでは予測精度が低い結果となった。

一方で、収益性について表 3.5 は各 6 つの手法の合計収益率を表す。合計収益率は当然予測精度に比例するが、変動が大きい時に予測が外れると差が生じる。収益性についても、提案手法である IDTW+k^*-NN は FTSE100 を除いて最も高い。各手法について個別にみていくと、正答率と同様に全指数の平均については、IDTW は DTW と DDTW を上回っているといえる。k^*-NN が k-NN を上回っている点も同様である。

以上から予測精度、収益性の両面で IDTW は DTW と DDTW を上回り、そして k^*-NN は k-NN を上回り、提案手法の有効性が確認できた。

図 3.8 から 3.12 は、6 つの手法を用いた各インデックスの予測に基づく投資の累積リターンの推移を示している。青線は各指数をそのまま買い続けるという単純な戦略の累積

表 3.4 The average accuracy of all years for each method. The out-of-sample period is from January 2006 to August 2017. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the six methods.

		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%
DDTW	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%**

表 3.5 The total returns of each method. The out-of-sample period is from January 2006 to August 2017. The rightmost column is the total mean for each method. The bold values are the highest cumulative returns of the six methods.

		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%
	k^*-NN	93.53%	154.90%	114.29%	139.60%	156.98%	131.86%
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%**

リターンを表している。IDTW+k^*-NN はすべての指数で、上昇トレンドを描き、最終的に投資対象指数を上回っている。

3.4.3 モメンタム・リバーサル効果との比較

ここでは、前節で最も結果のよかった IDTW+k^*-NN と、単純なモメンタム戦略との比較を行う。モメンタム戦略は、過去の先行研究において有効であるとされる、直近 1 ヵ月のリターンの正負でポジションを変える 1Mom と直近 1 ヵ月を除く、過去 11 ヵ月のリターンでポジションの正負を変える 12-1Mom の両方を使用する。いずれも直近 1 ヵ月または過去 11 ヵ月のリターンが正ならば、指数を 1 単位買い、負ならば指数を 1 単位売りという戦略を毎月繰り返す。この戦略の結果が表 3.6 である。表 3.6 から、モメンタム

表 3.6　The total returns of each method. The out-of-sample period is from January 2006 to August 2017. The rightmost column is the total mean for each method. The bold values are the highest cumulative returns of the three methods.

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%

表 3.7　The correlations between IDTW+k^*-NN and momentum strategy. The rightmost column is the total mean for each method.

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

戦略は 1Mom と 12-1Mom ともに全ての指数において正の収益が獲得できているものの、IDTW+k^*-NN の方が全ての指数においてモメンタム戦略よりも多くの収益を獲得していることがわかる。また、図 3.13 から 3.17 は、IDTW+k^*-NN と 1Mom と 12-1Mom による各インデックスの予測に基づく投資の累積リターンの推移を示している。IDTW+k^*-NN はすべての指数で、最終的にモメンタム戦略を上回っていることが確認できる。

また、表 3.7 は、IDTW+k^*-NN の収益率と 2 つのモメンタム戦略の月次の収益率間の相関係数を表している。いずれも低い相関係数となっていることから、価格変動パターンに着目した IDTW+k^*-NN による投資手法は単純なモメンタム戦略とはリターンの出方が異なる戦略であることが確認できる。

3.5　まとめ

本章における貢献は以下の 2 点である。現在の価格変動に類似した過去の価格変動パターンを探し、それをもとに予測を行うという観点から、IDTW を用いた k^*-NN 法を提案した。これはテクニカル分析におけるフォーメーション分析から恣意性を除いたアルゴリズムであるといえる。代表的な各国の株価指数を用いた実証分析を行い、先行研究により提案されていた方法 (DTW+k-NN、DDTW+k-NN) を含むいくつかの予測方法のパフォーマンスを比較し、次の結果が得られた。

- IDTW は、予測精度と収益性の両面で DTW と DDTW よりも優れている。
- 同様に、k^*-NN は予測精度と収益性の両面で k-NN よりも優れている。
- 提案手法である IDTW-k^*-NN は 6 つの手法の中で予測精度と収益性ともに最良の結果であった。

提案手法は過去の日次の株価と 1 つのパラメータ L/C のみを必要とする。これは、提案方法の予測精度以外の顕著な利点である。

一方で、本章では月内の日次の価格変動パターンが月次の株価予測に有効であることを示したが、参照期間および予測期間の選択がパフォーマンスに及ぼす影響を確認する必要がある。これについては、今村ら (2018)[今村 18b] によって検証されている。今村ら (2018)[今村 18b] によると、株式、為替予測において、1 ヶ月の予測には 1 ヶ月の参照期間、2 ヵ月の予測には 2 ヵ月の参照期間が最も予測精度が良いという直感的に合致した結果が得られている。また 1 ヵ月という月の単位による参照期間を用いた予測が、20 日といった日数ベースの参照期間による予測よりも予測精度が良いことを検証しており、本章における月間の日次リターンを使うことをサポートしている。

本章では、価格変動パターンが似ているというシンプルな仮定で予測を行ったが、どのような価格変動パターンが予測に有効であったかは不明である。そこで次章では、本章で有効であると確認できた過去の価格変動パターンをクラスタリング手法を用いて可視化し、具体的にどのようなパターンが有効であったかを検証する。

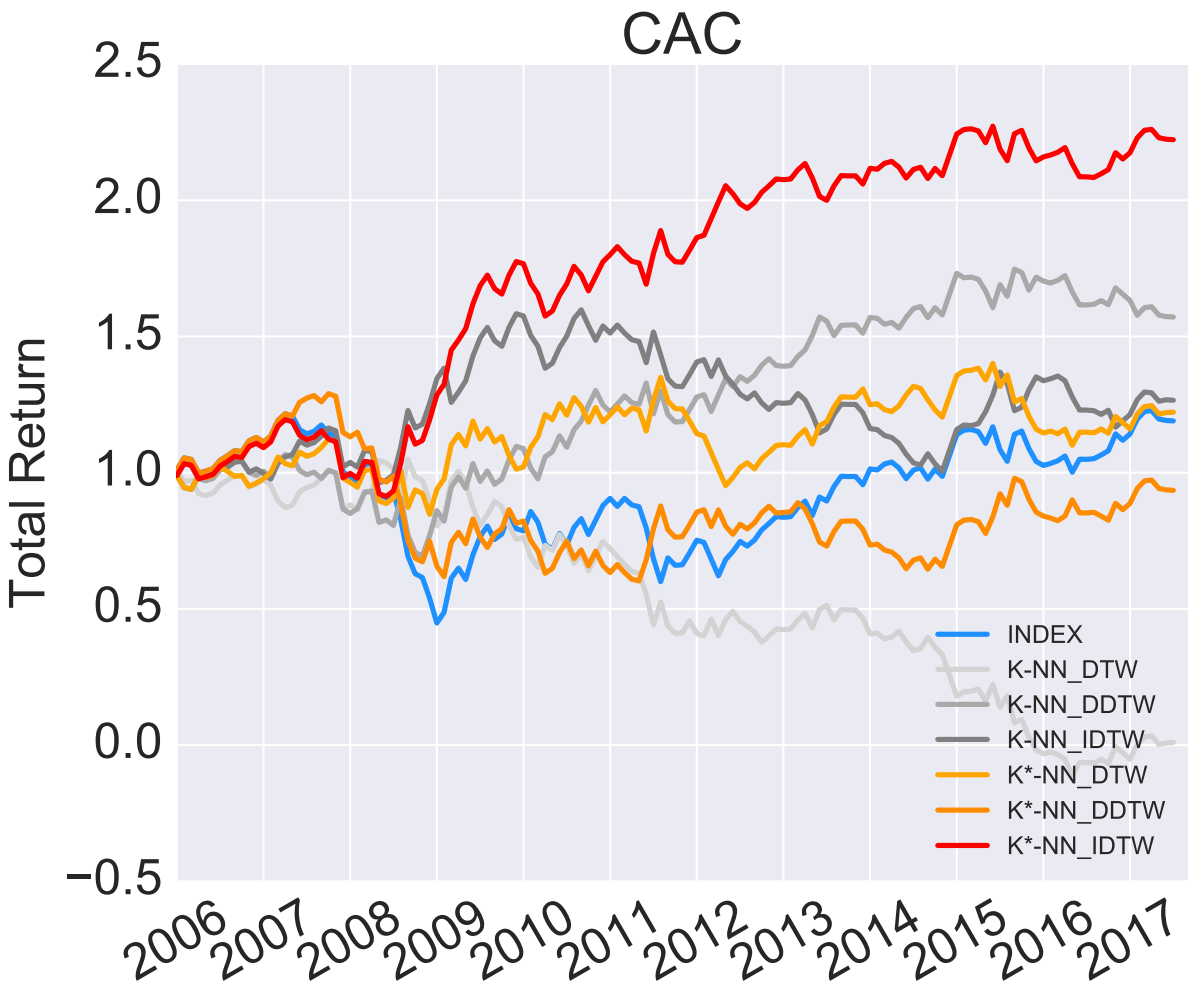

図 3.8　Change in cumulative returns of CAC index and for the six methods. The out-of-sample period is from January 2006 to August 2017.

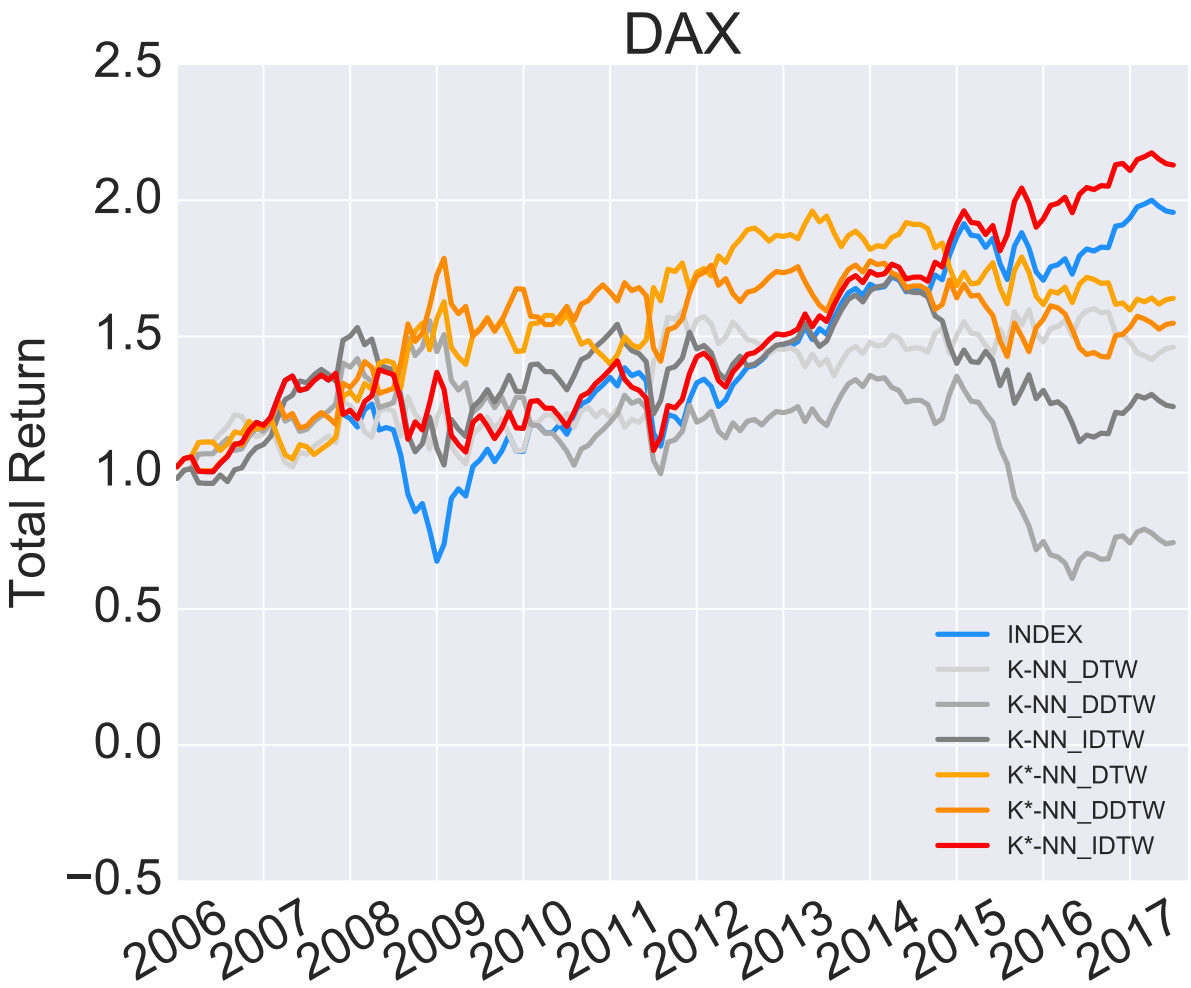

図 3.9　Change in cumulative returns of DAX index and for the six methods. The out-of-sample period is from January 2006 to August 2017.

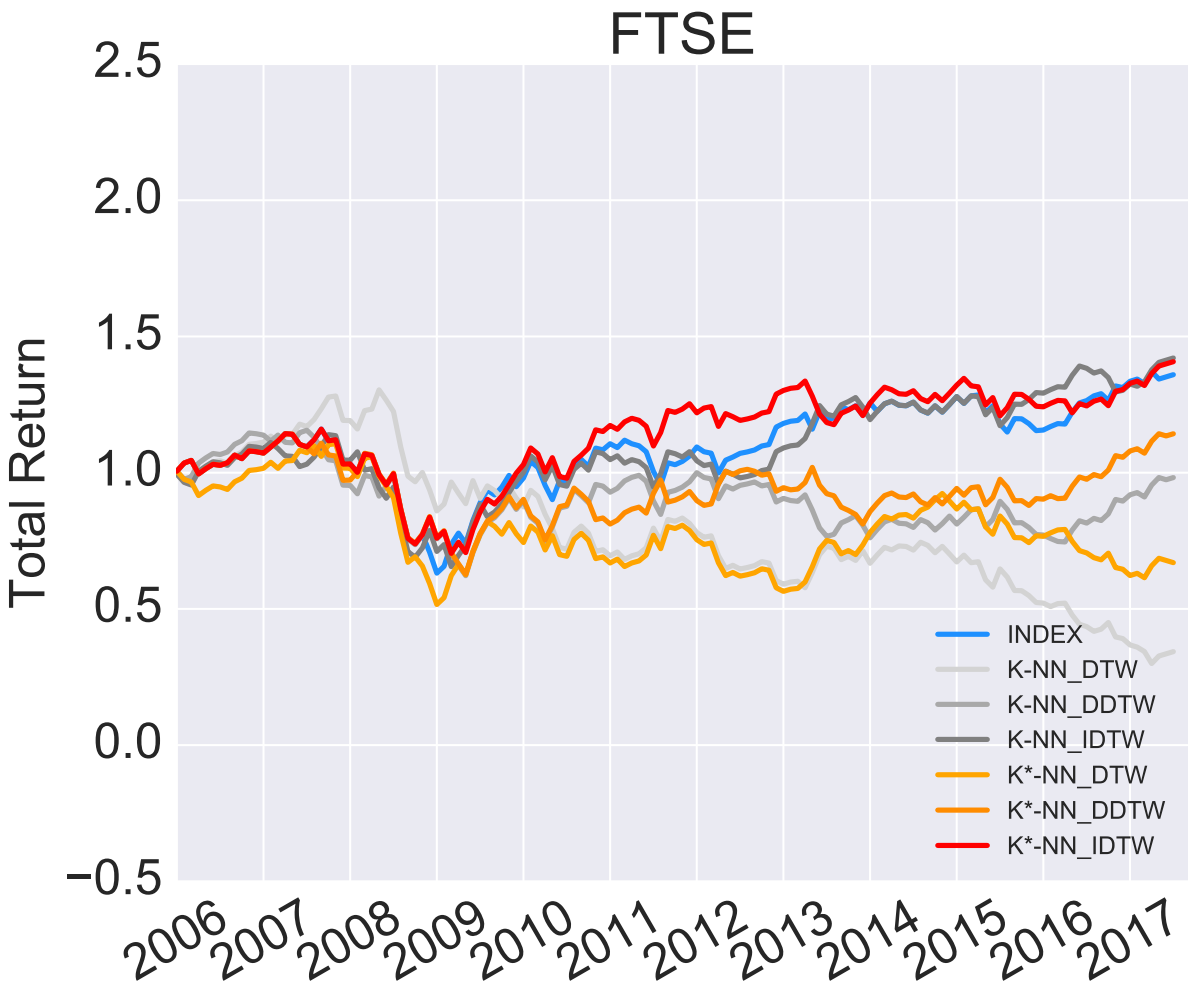

図 3.10 Change in cumulative returns of FTSE index and for the six methods. The out-of-sample period is from January 2006 to August 2017.

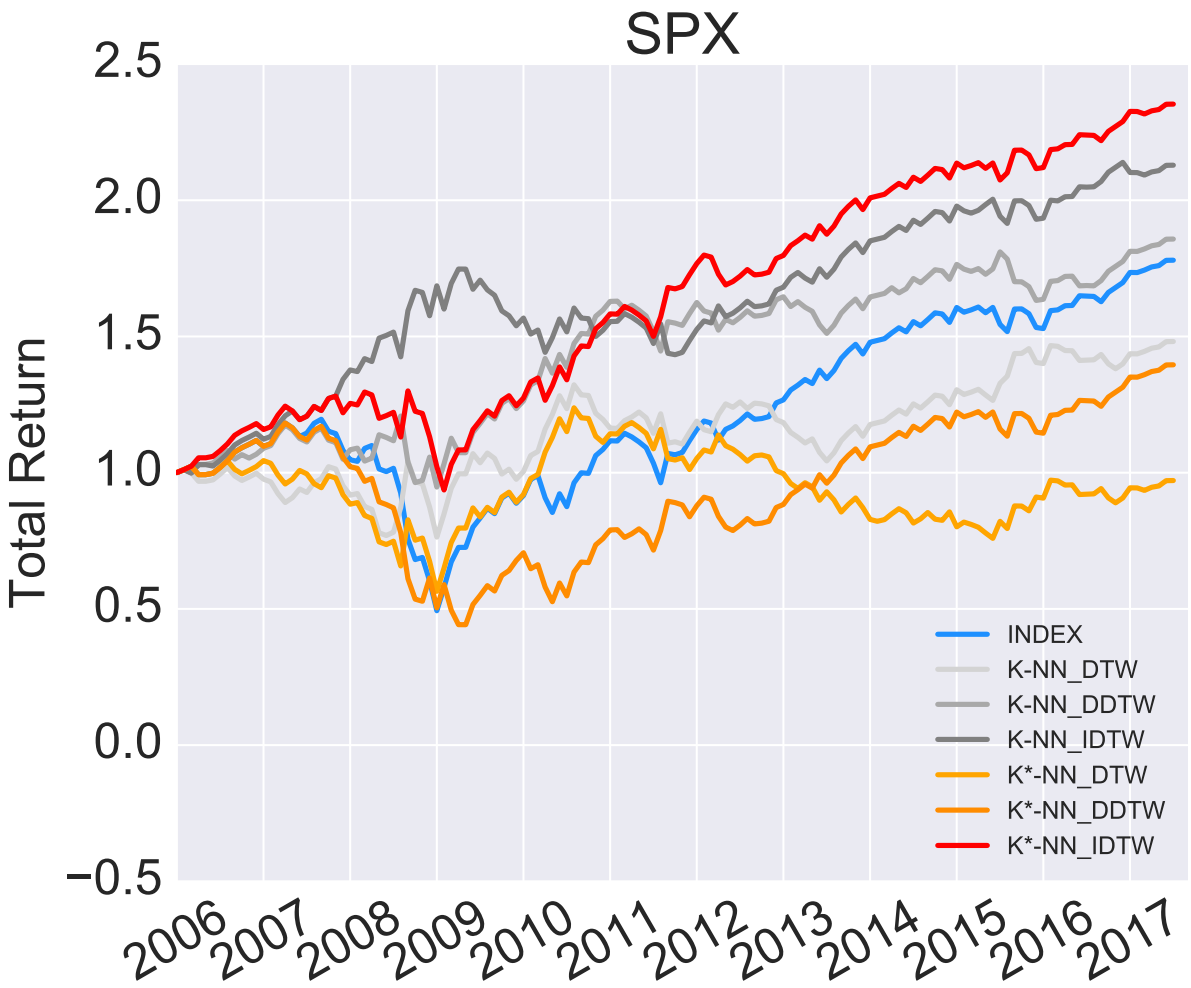

図 3.11　Change in cumulative returns of SPX index and for the six methods. The out-of-sample period is from January 2006 to August 2017.

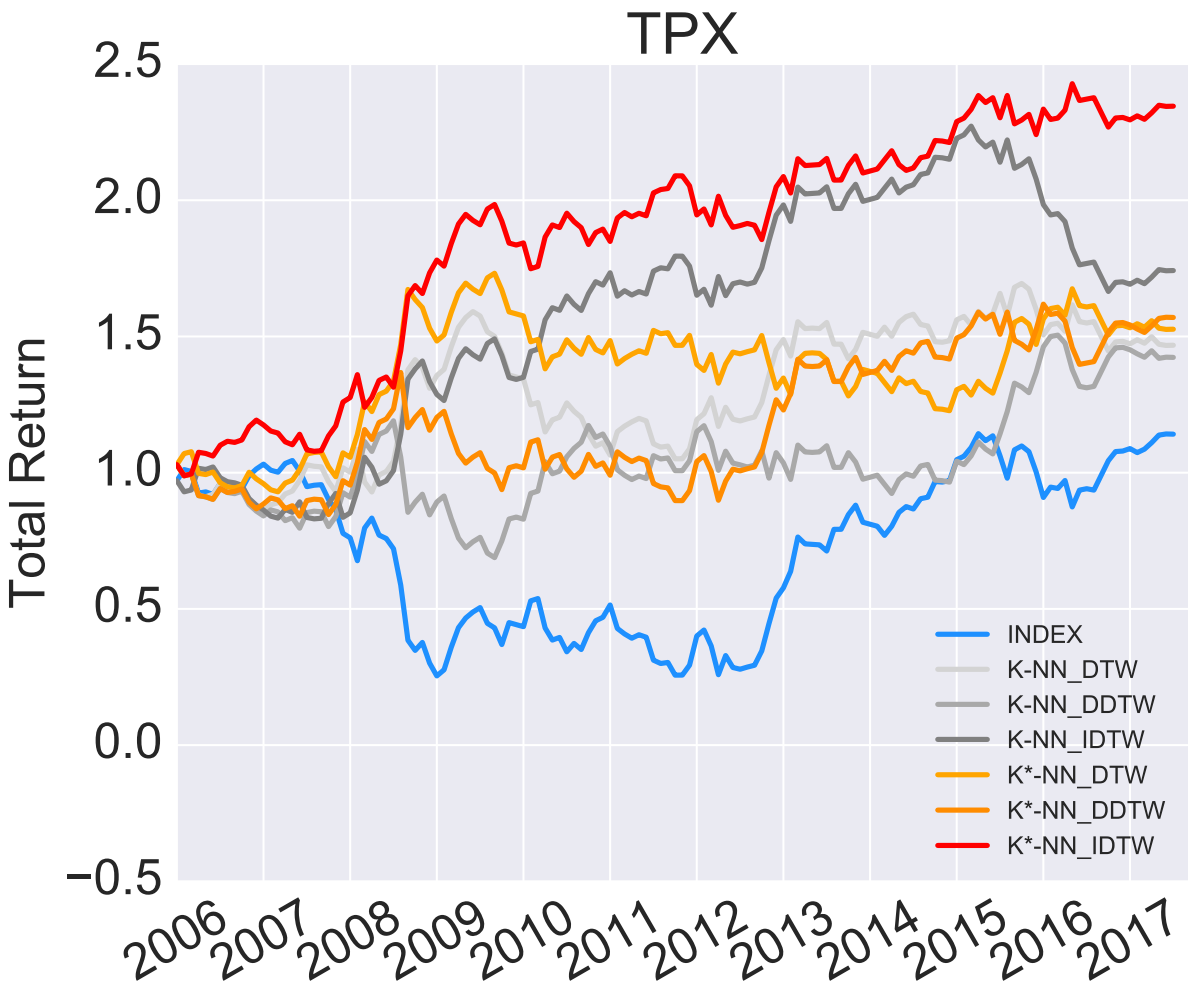

図 3.12　Change in cumulative returns of TPX index and for the six methods. The out-of-sample period is from January 2006 to August 2017.

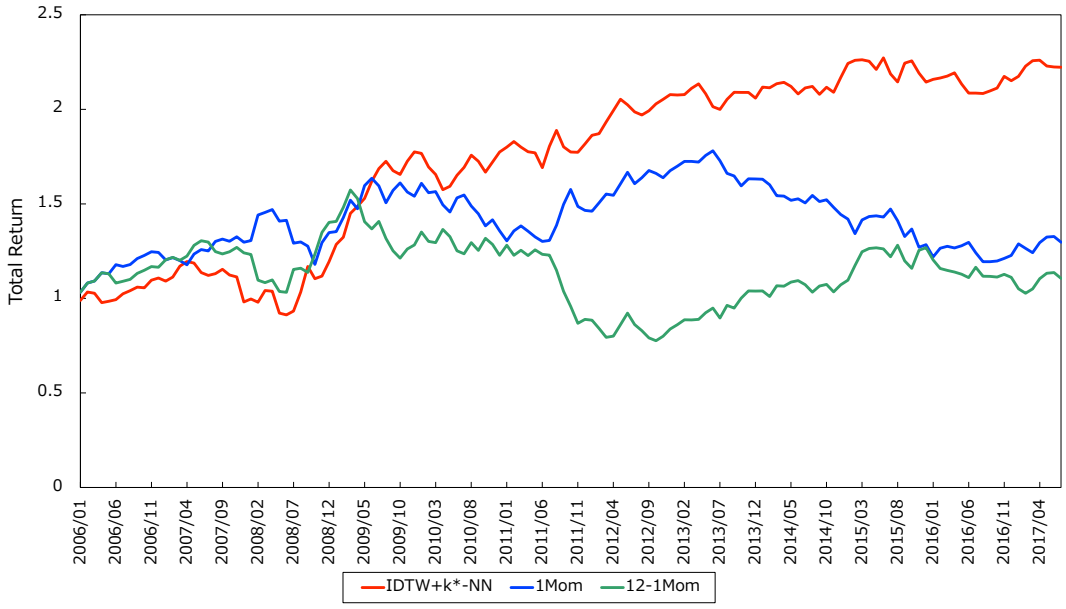

図 3.13 Change in cumulative returns of CAC index and for the IDTW+k^*NN,1Mom and 12-1Mom. The out-of-sample period is from January 2006 to August 2017.

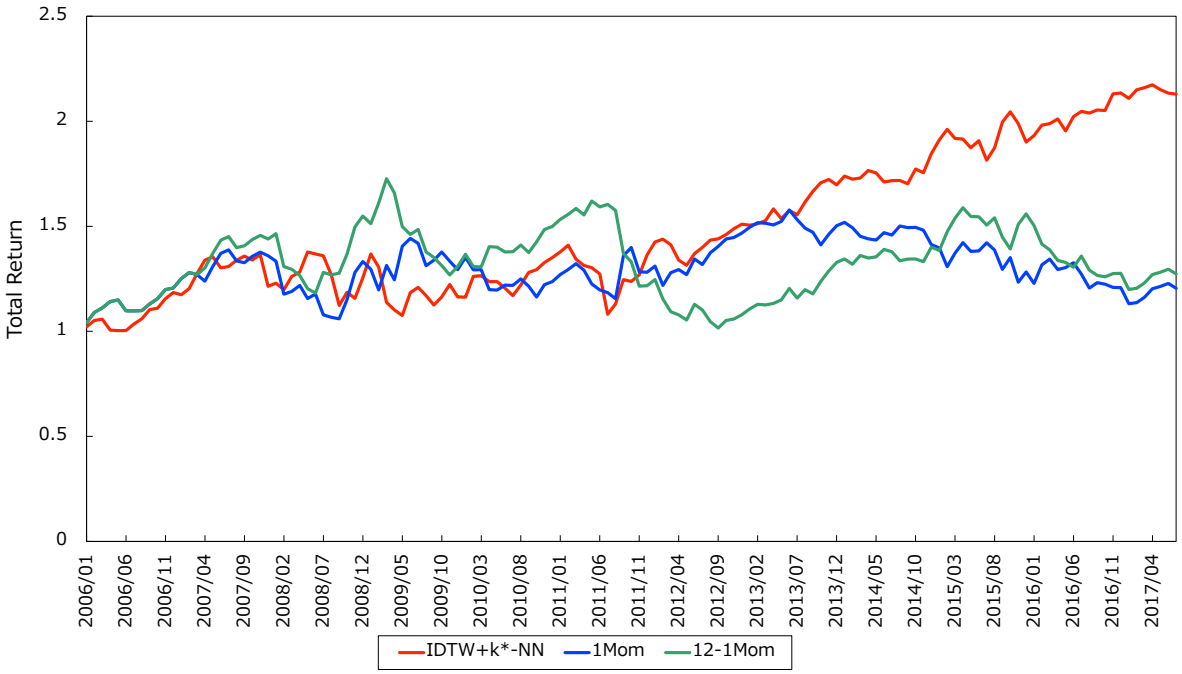

図 3.14 Change in cumulative returns of DAX index and for the IDTW+k^*NN,1Mom and 12-1Mom. The out-of-sample period is from January 2006 to August 2017.

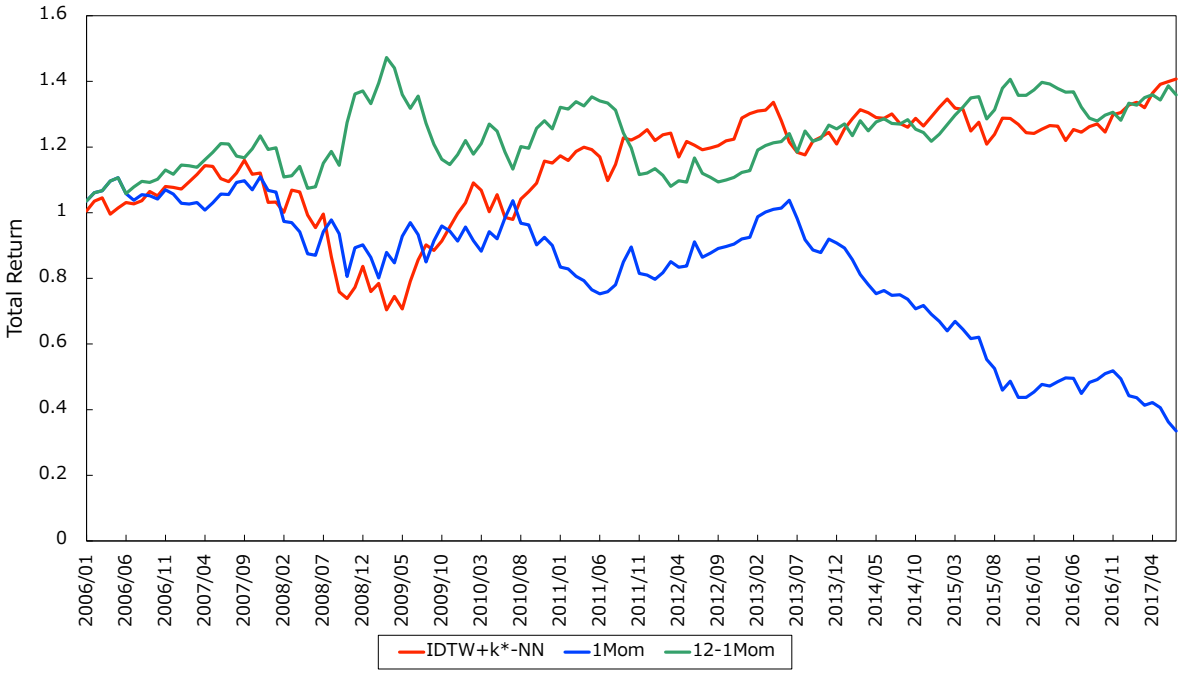

図 3.15 Change in cumulative returns of FTSE index and for the IDTW+k^*NN,1Mom and 12-1Mom. The out-of-sample period is from January 2006 to August 2017.

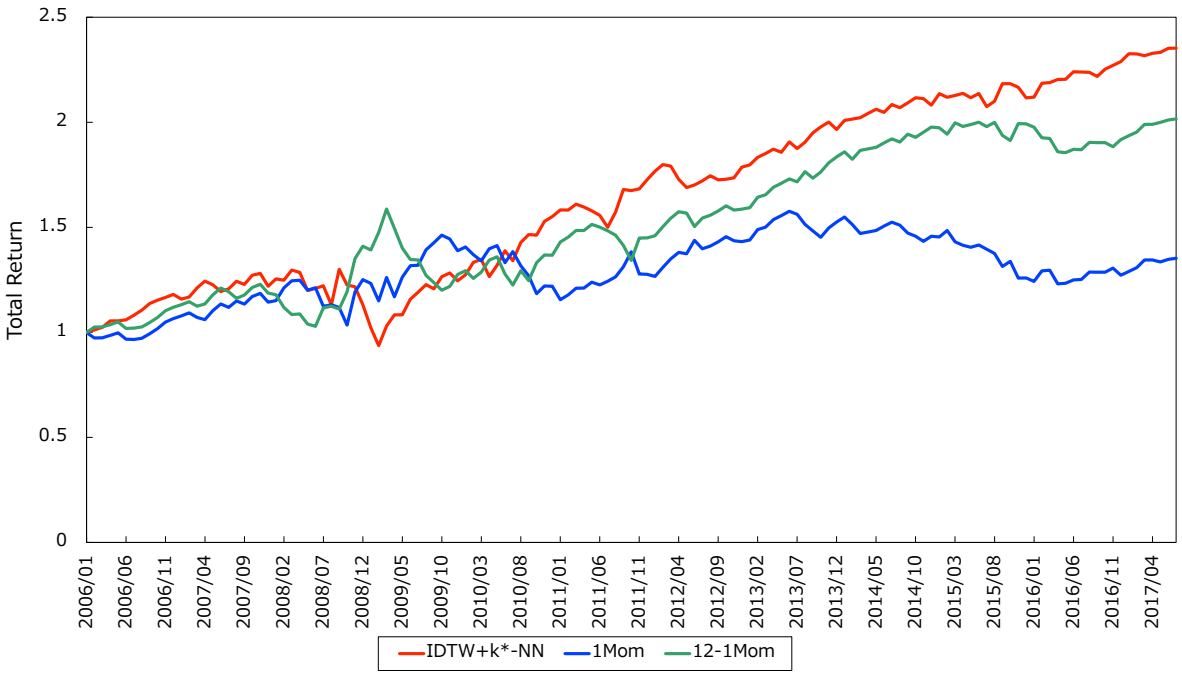

図 3.16 Change in cumulative returns of SPX index and for the IDTW+k^*NN,1Mom and 12-1Mom. The out-of-sample period is from January 2006 to August 2017.

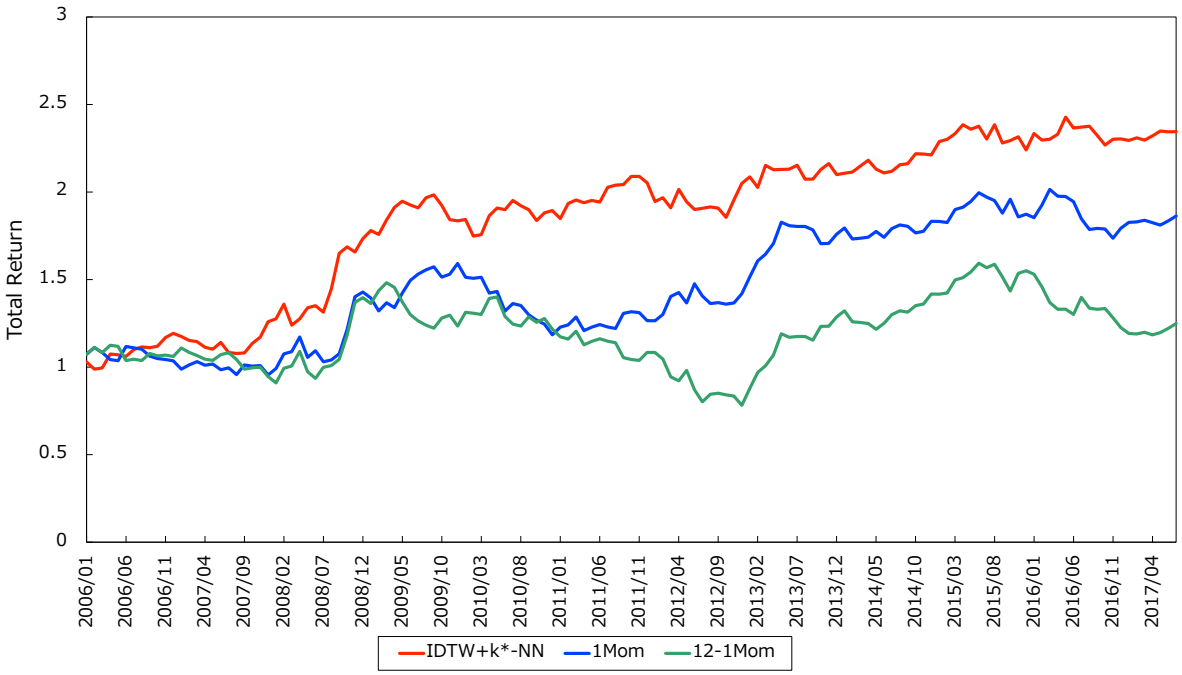

図 3.17 Change in cumulative returns of TPX index and for the IDTW+k^*NN,1Mom and 12-1Mom. The out-of-sample period is from January 2006 to August 2017.

3.6 補論

ここでは、Anava に基づき、k^*-NN のアルゴリズムを導出する。まず、解くべき問題は次式である。

$$\arg\min_{\boldsymbol{\alpha}\in\Delta_n} |\sum_{i=1}^{n}\alpha_i y_i - f(x_0)| \tag{P1}$$

この式を変形し、

$$|\sum_{i=1}^{n}\alpha_i y_i - f(x_0)| = |\sum_{i=1}^{n}\alpha_i\varepsilon_i - \sum_{i=1}^{n}\alpha_i(f(x_i)-f(x_0))| \tag{3.11}$$

$$\leq |\sum_{i=1}^{n}\alpha_i\varepsilon_i| + L\sum_{i=1}^{n}\alpha_i d(x_i,x_0) \tag{3.12}$$

$$\because |f(y)-f(x)| < L\times d(x,y) \tag{3.13}$$

$$\leq C||\boldsymbol{\alpha}||_2 + L\sum_{i=1}^{n}\alpha_i d(x_i,x_0) \tag{3.14}$$

$$\because |\sum_{i=1}^{n}\alpha_i\varepsilon_i| \leq C||\boldsymbol{\alpha}||_2 \tag{3.15}$$

を得る。ここで、$\beta_i = L/C\times d(x_i,x_0)$ とおくと、

$$\arg\min_{\boldsymbol{\alpha}} ||\boldsymbol{\alpha}||_2 + \boldsymbol{\alpha}^T\boldsymbol{\beta} \tag{3.16}$$

$$s.t. \sum_{i=1}^{n}\alpha_i = 1 \tag{3.17}$$

$$\alpha_i \geq 0 \tag{3.18}$$

と問題を書き直せる。

ここでラグランジュ関数 $L(\boldsymbol{\alpha}, \lambda, \boldsymbol{\theta})$ を用いて、KKT 条件より

$$L(\boldsymbol{\alpha}, \lambda, \boldsymbol{\theta}) = (||\boldsymbol{\alpha}||_2 + \boldsymbol{\alpha}^T \boldsymbol{\beta}) + \lambda(1 - \sum_{i=1}^{n} \alpha_i) - \sum_{i=1}^{n} \theta_i \alpha_i \tag{3.19}$$

$$\frac{\partial L}{\partial \boldsymbol{\alpha}} = 0 \tag{3.20}$$

$$(1 - \sum_{i=1}^{n} \alpha_i) = 0 \tag{3.21}$$

$$\lambda \in R \tag{3.22}$$

$$\theta_i \geq 0 \tag{3.23}$$

$$\sum_{i=1}^{n} \theta_i \alpha_i = 0 \tag{3.24}$$

$$\alpha_i \geq 0 \tag{3.25}$$

が得られる。これらの最適性条件を満たす α^* を求める。

$$\frac{\alpha_i}{||\boldsymbol{\alpha}||_2} = \lambda - \beta_i + \theta_i \tag{3.26}$$

また、$\alpha > 0$ と $\alpha = 0$ の 2 つの場合があるが、今興味があるのは $\alpha > 0$ であるので、$\theta = 0$ である。ゆえに、

$$\frac{\alpha_i^*}{||\boldsymbol{\alpha^*}||_2} = \lambda - \beta_i \tag{3.27}$$

が得られる。ここで、両辺を 2 乗し、$\alpha^* > 0$ の和をとると、

$$1 = \sum_{\alpha_i^* > 0} \frac{(\alpha_i^*)^2}{||\boldsymbol{\alpha^*}||_2^2} = \sum_{\alpha_i^* > 0} (\lambda - \beta_i)^2 \tag{3.28}$$

また、$\alpha_i^* > 0$ であるような α の個数を k^* とおくと、

$$k^* \lambda^2 - 2\lambda \sum_{i=1}^{k^*} \beta_i + (\sum_{i=1}^{k^*} \beta_i^2 - 1) = 0 \tag{3.29}$$

が成り立つので、

$$\alpha_i^* = \frac{\lambda - \beta_i}{\sum_{i=1}^{k^*} (\lambda - \beta_i)} \tag{3.30}$$

また二次方程式の解から、

$$\lambda = \frac{1}{k^*} \left\{ \sum_{i=1}^{k^*} \beta_i + \sqrt{k^* + (\sum_{i=1}^{k^*} \beta_i)^2 - k^* \sum_{i=1}^{k^*} \beta_i^2} \right\} \tag{3.31}$$

が得られる。これは Anava のアルゴリズムと一致する。

第 4 章

価格変動パターンのクラスタリング手法

4.1 はじめに

前章において、IDTW を用いて適切にスケーリングされた月内の日次株価変動パターンが、将来の株価予測に有効な特徴量であることが確認できた。本章では、背後にある価格変動パターンを時系列クラスタリングによって具体的に抽出、可視化し、当該現象の意味付けを行う。特に予測という観点から有効な価格変動パターンのクラスタを特定する。しかしながら前章までと同様に、月内の日次株価変動はデータ数 (営業日) が月毎に異なるため、単純なベクトル空間上のユークリッド距離を用いたクラスタリング手法 (例えば、k-means 法) は適用できない。よって、株価を対象に異なるデータ数においても、欠損値として値の挿入および削除を行うことなく、より自然にデータ間の類似性を測定し、そして対応するクラスタリング方法を適切に組み合わせる必要がある。本章では以上を考慮し、前章の IDTW を用いて作成された類似度行列に対して、k-medoids クラスタリング [KR87] を行うことを提案する。これにより月内の日次株価のような長さの異なる金融時系列データに対して、自然なクラスタリングが可能となることが本手法の利点である。

本章では、IDTW を用いた k-medoids クラスタリングを TOPIX 指数に対して適用し、予測に有効な価格変動パターンのクラスタを特定するとともに、各価格変動パターンのクラスタの特徴を考察する。

4.2 時系列クラスタリングの先行研究

クラスタリングとは、何らかの外的な要素なしに与えられたデータの集合を内的結合と外的分離が達成されるようなクラスタと呼ばれる部分集合に分割する教師なし学習の一種である [ELL93]。内的結合とは同じクラスタ内のデータ同士は似ており、外的分離とは違うクラスタにあるデータ同士は似ていないことをいう。クラスタリングは教師データを必要としないため、簡便であり、自然言語処理・データマイニング・パターン認識・イメージ解析・バイオインフォマティクスといった幅広い分野で用いられている [GMW07]。一般にクラスタリングの結果は、似たデータを集めた結果として出来るグループであり、その分類結果は分析目的などに応じて妥当性を検証する必要がある。クラスタリングには階層的クラスタリング手法と非階層的クラスタリング手法の 2 種類がある。階層的クラスタリング手法は、ただ 1 つのデータを含む N 個のクラスタから、クラスタ間の非類似度関数に基づいて、最も距離の近い 2 つのクラスタ同士を逐次的に合併していく。この合併を、全てのデータが 1 つのクラスタに合併されるまで繰り返すことで階層的な構造を獲得する。2 つのクラスタ間 (C_1 および C_2) の非類似度関数 $D(C_1, C_2)$ の違いによって次のような手法がある [神蔦 03]。

- 最短距離法:$D(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} d(x_1, x_2)$
- 最長距離法:$D(C_1, C_2) = \max_{x_1 \in C_1, x_2 \in C_2} d(x_1, x_2)$
- 群平均法:$D(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} d(x_1, x_2)$
- ウォード法:$D(C_1, C_2) = E(C_1 \cap C_2) - E(C_1) - E(C_2)$

ここで、$d(x, c_i)$ は距離関数、$E(C_i) = \sum_{x_i \in C_i} (d(x_i, c_i))^2$ であり、c_i はクラスタ C_i の重心 (平均) である。

一方で、非階層的手法はいくつのクラスターに分類するかをあらかじめ決めておき、決められたクラスター数にデータが分割できるようにまとめる手法であり、階層的な構造が必要ない場合に使用される。非階層的クラスタリング手法である、k-means 法 [For65] は実装が容易であり計算効率にも優れているため、クラスタリングにおける定番アルゴリズムとなっている。k-means 法は、クラスタの重心点であるセントロイド c_i をクラスタの代表点とし、入力データを x とすると $\sum_{i=1}^{k} \sum_{x \in c_i} (d(x, c_i))^2$ の評価関数を最小化するよう

Algorithm 6 k-means clustering

```
1: procedure K-MEANS CLUSTERING(\{x_1, \cdots, x_N\}, k)
2:     Randomize m_1, \cdots, m_k
3:     while stopping criterion has not been met do          ▷ Cluster Assignment
4:         for i = 1 to N do
5:             c_i := \arg\min_j d(x_i, m_j)
6:             c = c \cup c_i
7:         end for                                           ▷ Update Centroids
8:         for j = 1 to k do
9:             m_j := mean(x|c = j)
10:        end for
11:    end while
                                                             ▷ Return the medoids
12:    return m_1, \cdots, m_k
13: end procedure
```

に k 個のクラスタを分割する。具体的なアルゴリズムは次の通り (Algorithm 6)、クラスタへの割り当てと代表点の再計算を交互に繰り返して行う。はじめに 2 行目で k 個のクラスターの初期値 $m_1, ..., m_k$ を割り当てる。次に、各データ x_i がどのクラスタに属するかの判定（5 行目）と、それをリストとして保持 (6 行目) する。そして、クラスタ $m_1, ..., m_k$ を、各 c_j に該当するデータ x の平均に更新 (9 行目) する。

k-means クラスタリングは座標をクラスタの重心 (セントロイド) とするため、外れ値の影響を受けやすい。また、距離関数を用いるため、データが距離の定義できるベクトル空間に存在する必要がある。これらの問題を解決するために、k-medoids クラスタリング [KR87] が提案された。

一方で、株価のような時系列データをクラスタリングする際に、k-means 法を用いると意味のあるクラスタリングができないことがあることを Keogh and Lin(2005)[KL05] は議論した。彼らは、1 本の時系列データを 1 時点毎にずらして時系列データを作成していくこと (スライディングウィンドウ方式) により得られた部分時系列をクラスタリングする問題が、ランダムなクラスタリングとなって意味のないものであることを実験的に示した。n 個の部分時系列 $[t, t+w-1], [t+1, t+(w-1)+1]...[t+n, t+(w-1)+n]$ を 1 本の全体の時系列から取り出す。これらの部分時系列を長さ w のベクトルと見なしてクラス

タリングすることを部分時系列クラスタリング (Subsequence Time-Series clustering;STS クラスタリング) と呼ぶ。このような STS クラスタリングを行うと、クラスタリングの結果はノイズやクラスタリングの初期値に依存して決まるランダムな分割しか導かないことを実証的に示した。また、k-means クラスタリングを適用するとその重心は、元のパターンとは全く異なるサイン曲線のようなものになることを例示している。この原因として直観的な説明を与えており、それを自明な一致 (trivial match) としている。これは、$[t, t+w-1]$ とその隣の $[t+1, t+(w-1)+1]$ は重なりが多く、類似度が非常に高くなる。そのため、時間的に近い系列が同じクラスタに判定されやすくなり、結果として意図したクラスタリングが得られない。Keogh and Lin(2005)[KL05] はこの自明な一致の単純な回避法として、1 ずつずらすのではなく、一度にある程度大きなステップ幅でずらすことを挙げている。

本章では、各月内の日次株価を部分時系列として使用するため、時系列データの重なりは存在しない。また、k-means ではなく、k-medoids を用いることで元の時系列と全く異なるパターンが現れることを回避している。

以上を踏まえ、IDTW による類似度行列を使用した k-medoids 法を提案する。

4.3 提案手法

4.3.1 k-medoids clustering with Indexing Dynamic Time Warping

k-medoids 法とは、k-means 法と類似した非階層的クラスタリング手法である。k-means 法との具体的な相違点は、クラスターは centroid(重心) ではなく medoid で代表される点である。medoid とは、クラスター内のデータ点で、その点以外のクラスター内の点までの類似度の総和が最小になる点である。直感的には、最もクラスターの中心付近に位置する点を表す。従って、medoid は必ずクラスタリングの対象となるデータ内に存在する。そのため、k-medoids 法は分類するデータ間の類似度行列を与えれば適用でき、任意の類似度尺度 (距離を含む) に対してクラスタリングを実行できる。すなわち、クラスタリングの対象が長さの異なる時系列やグラフのような、ベクトル空間で表現されたデータ以外であっても類似度 (非負の関数) を定義できれば適用できる。また、k-means がエラーを距離の 2 乗で評価するのに対し、k-mediods は類似度そのもので評価するので、ノイズや外れ値の影響を受けにくい。金融時系列は分布の裾が厚く、外れ値の影響が大きく

なることが想定されるため、この点も金融時系列を解析するにあたり k-mediods 法は好ましい。

そこで本章では、月間の株価のような長さの異なる金融時系列データに対して、IDTW を非類似度尺度に使用した k-medoids 法を適用することを提案する。類似度を計測する尺度に IDTW を用いることで、金融時系列データに対して自然なクラスタリングが可能となる。

本章と同様に株価の価格変動パターンのクラスタリングを行った研究として、白浜 (2011)[白浜 11] や Aghabozorgi et,al.(2014)[AT14] がある。白浜 (2011)[白浜 11] と Aghabozorgi et,al(2014)[AT14] では、本章と同様に DTW を類似度として株価のクラスタリングを行っている。しかし、白浜 (2011)[白浜 11] は階層的クラスタリング手法を用いており、Aghabozorgi et,al(2014)[AT14] は 3 段階のステップを必要とする階層的なクラスタリング手法を提案している。両者ともに階層的なクラスタリング手法であるためいずれも計算量が大きく、後者は 3 段階を踏むため考慮すべきパラメータが多い。一方で、本研究の手法のパラメータはクラスタ数 k のみであり、k-medoids 法による非階層的なクラスタリングであり計算量は比較的小さい。また分析の目的としても、白浜 (2011)[白浜 11] と Aghabozorgi et,al(2014)[AT14] ともにある時点における断面での複数の株価間の株価変動パターンのクラスタリングを行ったのに対して、本章ではある株価の時系列における株価変動パターンのクラスタリングを目的とする点が異なる。

具体的なアルゴリズムは以下の通り (Algorithm 7)。$x_1, \cdots, x_N$ はそれぞれ長さの異なる時系列データ、k はクラスタ数である。停止条件 (stopping criterion) はクラスタの割り当てが変化しなかった場合、または繰返し回数が上限 (100 回) に達した場合である。はじめに 2 行目で k 個のクラスタの初期値 $m_1, ..., m_k$ を割り当てる。次に、各データ x_i が、IDTW を用いてどのクラスタに属するかの判定（5 行目）と、それをリストとして保持 (6 行目) する。そして、クラスタ j に含まれるデータ x を D_j と定義 (9 行目) し、クラスタ m_j を、各 D_j に含まれるデータ点 x_i のうち、その点以外のクラスター内の点 x_l までの IDTW で計測した類似度の総和が最小になる点に更新 (10 行目) する。

Algorithm 7 k-medoids clustering with IDTW

procedure K-MEDOIDS CLUSTERING WITH IDTW($\{x_1, \cdots, x_N\}, k$)
  Randomize $m_1, \cdots, m_k$
  while stopping criterion has not been met **do** ▷ Cluster Assignment
    for $i = 1$ to N **do**
      $c_i := \arg\min_j IDTW(x_i, m_j)$
      $c = c \cup c_i$
    end for ▷ Update Medoids
    for $j = 1$ to k **do**
      $D_j := \{x | c = j\}$
      $m_j := \arg\min_{x_i \in D_j} \sum_{x_l \in D_j} IDTW(x_i, x_l)$
    end for
  end while
  ▷ Return the medoids
  return $m_1, \cdots, m_k$
end procedure

4.4 実証分析

4.4.1 分析手順

本章では、TOPIX 指数 (配当込) を用いた実証分析により、類似した価格変動パターンのクラスタリングと、抽出したクラスタが特徴量として有効かどうかを確認する。特徴量としての有効性を確認するために、提案手法を含む下記の手法を用いて TOPIX 指数の予測精度を比較する。指数データは情報端末の Bloomberg から取得した。クラスタリング手法は、提案手法の有効性を確認するため、1. 最もシンプルな時系列モデルである、AR モデルによる予測 (AR)、2. 先行研究である Niennattrakul and Ratanamahatana(2007)[NR07] を参考に時系列データのクラスタリングのベンチマークとしてふさわしいとされる k-medoids clustering with DTW(DTW)、3.IDTW と k-NN の改善手法である Anava and Levy(2016)[AL16] の k^*-NN を組み合わせた前章の IDTW + k^*-NN(k^*-NN) および、4. 提案手法の k-medoids clustering with IDTW(IDTW) をそれぞれ比較する。

なお、本手法のような非階層のクラスタリングの際にはクラスタ数 k を決める必要があ

るが、ここでは、$2 \leq k \leq 12$ まで動かし、予測精度が最も良くなる、すなわち最も価格変動を説明するクラスタ数を k として抽出する。これは予測において有効な価格変動パターンのクラスタを抽出するためである。予測精度は収益率と正答率の両面で評価する。具体的な分析手順は下記の通り (図 4.1 も併せて参照)。AR については、毎月利用できるすべてのデータの月次リターンを用いて、パラメータの再推定を行った。またラグ次数については 10 までの最適なラグを AIC 基準で毎月選択した。k^*-NN については先行研究である Anava and Levy(2016)[AL16] と同様の条件で予測を行った。

データ期間は 1989 年 1 月から 2017 年 3 月までとし、検証期間は 2007 年 1 月から 2017 年 3 月までの 10 年間とした。

Step1

$t-1$ 月までの月間の日次価格変動をもとに、クラスター数 k を固定し、クラスタリングを行う。各月は翌月のリターンの上昇、下落をラベルとして保持する。

Step2

t 月の月間の価格変動が属するクラスターを特定する。

Step3

t 月が属するクラスタを求め、当該クラスタにおいて、上昇 (下落) のラベルが多い場合は、月末の価格で買い (売り)、収益を計算する。同数の場合には、上昇と下落のラベルのそれぞれのリターンの平均値を取り、大きい方を採用する。

Step4

$t = t + 1$ に進めて Step 1 に戻る。

4.4.2 分析結果

DTW、および IDTW を用いた各クラスタ数ごとの結果のサマリーが表 4.1 である。

収益率、正答率ともに IDTW が DTW を上回る傾向がみられる。IDTW のクラスタ数については、$k = 4, 5, 6$ 辺りに収益率と正答率のピークがあり、それ以降については収益率、正答率ともに減衰している。そのため、予測という観点からは TOPIX 指数の価格変動パターンのクラスタ数は 5 程度であると言える。このとき、累積収益率は 162% で、正答率も 64% と良好な結果であった。これは分析期間は異なるものの、本研究と同じく 10 年間の TOPIX の月次予測を、テキスト情報をもとにおこなった蔵本ら (2013)[藏本 13] の

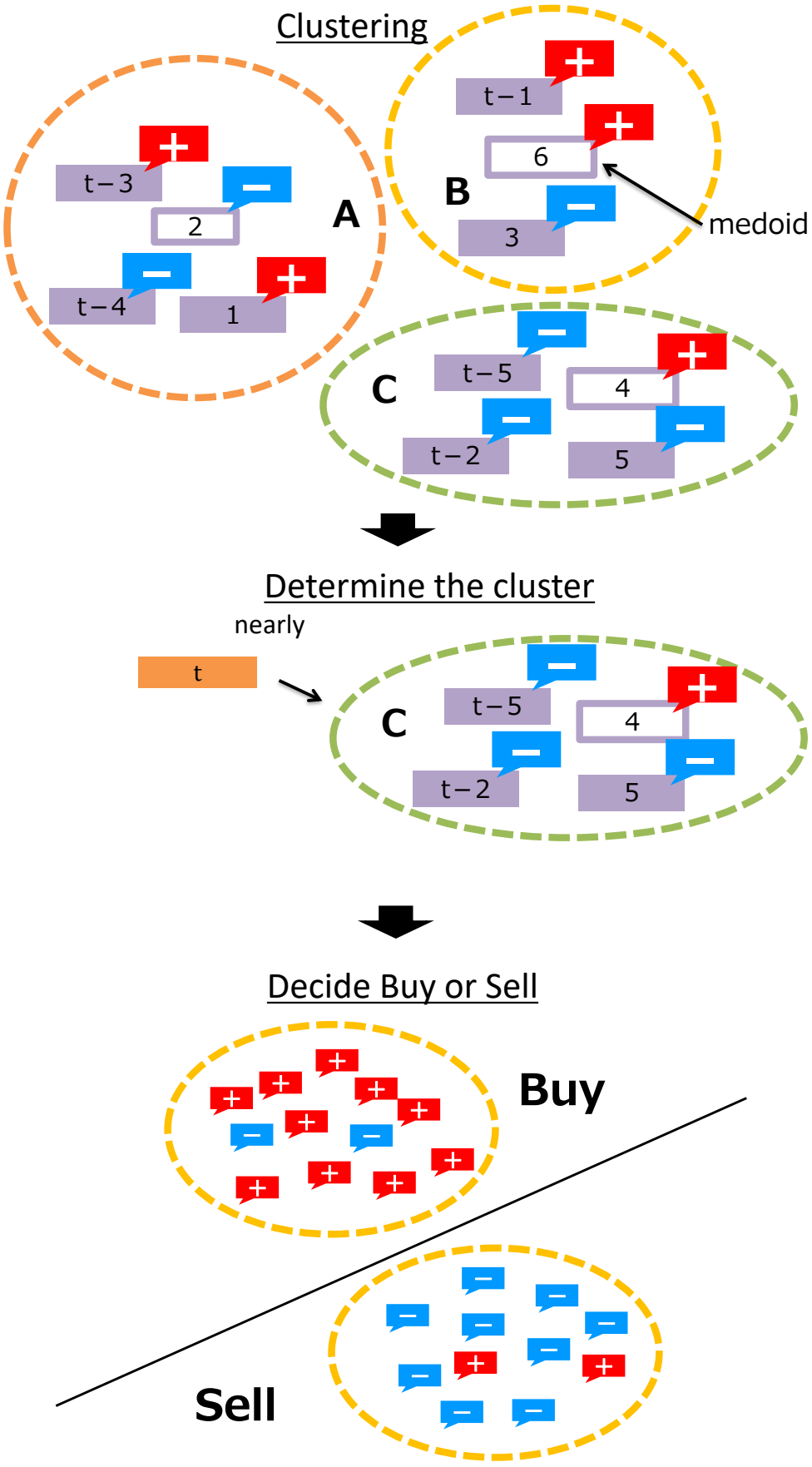

図 4.1　Stock price prediction framework.

表 4.1 The average accuracy of all years and total retrun for DTW and IDTW with k-medoids clusturing. The out-of-sample period is from January 2007 to March 2017. The bold values are the best measurements of each column.

	Total Return[%]		Accuracy[%]	
	IDTW	DTW	IDTW	DTW
$k = 2$	113	23	58.87	53.23
$k = 3$	98	-17	57.26	51.61
$k = 4$	148	66	61.29	53.23
$k = 5$	**162**	64	63.71	54.03
$k = 6$	151	50	**64.52**	54.03
$k = 7$	131	5	61.29	51.61
$k = 8$	132	6	62.1	53.23
$k = 9$	147	45	62.9	54.84
$k = 10$	114	41	58.06	55.65
$k = 11$	104	**79**	58.87	**59.68**
$k = 12$	90	56	59.68	58.06

表 4.2 The average accuracy of all years and total retrun for each method. The out-of-sample period is from January 2007 to March 2017.The bold values are the best measurements of each row.

	AR	k^*-NN	DTW($k = 11$)	IDTW($k = 5$)
Accuracy[%]	56.45	58.87	59.68	**64.52**
Total Return[%]	65	101	79	**162**

手法と同程度の予測力である。一方で DTW によるクラスタリングでは、$k = 11, 12, 13$ 辺りに収益率と正答率のピークがあり、IDTW と比べ、クラスタ数が増加していることが確認できる。

表 4.2 は、ベンチマークである AR、k^*-NN と最も精度の良かった $k = 11$ の DTW と $k = 5$ の IDTW の正答率と収益率を比較した。表から、IDTW によるクラスタリングは、AR や k^*-NN を正答率、収益率の両方で上回ったことが確認できる。

また、図 4.2 は TOPIX と DTW、IDTW の収益率が最も良い結果の収益率推移を示した図である。IDTW は期間に依らず右肩上がりの推移であり、TOPIX と DTW を上回る結果となった。したがって、提案手法により抽出された代表的な価格変動パターンは予測

に有効であることが確認できた。

次に抽出された各クラスタの解釈をおこなう。図 4.3 から 4.7 は、2017 年 3 月の分析手順における Step 1 の IDTW を用いた、$k = 5$ のクラスタリング結果を示した図である。黒線が中心点 (medoids) として選ばれた価格変動であり、赤線が当該クラスタに属する価格変動を表す。また、表題のサンプルはクラスタ内に含まれるサンプル数、かっこ内は翌月のリターンの上昇確率を表す。各クラスタの解釈を順に述べる。図 4.3 は、黒線の中心点は変動を伴わない横ばいの変動であるが、カッコ内に着目すると翌月に下落する可能性が高い。図 4.4 は、価格変動が下落傾向に関わらず、翌月の上昇確率が高く、価格の方向性が反転しているため、下落から上昇のリバーサルを表している。図 4.5 は、強い価格上昇を表しており、翌月の上昇確率も高く、価格の上昇傾向が継続しているため、上昇モメンタムを表している。図 4.6 は、ほぼ横ばいの変動を示し、翌月も横ばいの変動で動きがない相場である。図 4.7 は、強い価格下落を表しており、翌月の下落確率も高く、価格の下落傾向が継続しているため、下落モメンタムを表している。

一般に日本株式市場においては、ファクターとしてモメンタムの有効性が確認できないことが多い [AMP13]。しかし、本節で確認したように価格変動の類似度に基づきクラスタを適切に定義することで、当月の価格変動が大きい場合に、上昇、下落ともに強いモメンタム効果が確認できた。一方で、リバーサルの効果については、下落から上昇方向の効果しか顕著に確認できいことがわかった。

以上のような価格変動パターンによるクラスタ毎の解釈が可能であることが、実証ファイナンスのモメンタム、リバーサルの考え方である、比較月の価格に対してどれだけ上がったか、下がったかという見方から、価格変動パターンに拡張した見方をすることの利点である。

4.5 まとめ

本章では、株価の価格変動パターンのクラスタリングを行い、代表的な価格変動パターンを抽出し、予測のための特徴量として使用した。抽出にあたっては、指数化した価格変動に対して DTW 距離を計測する IDTW を用いた非類似度行列に対して k-medoids クラスタリング (k-medoids clustering with IDTW) を行う手法を提案した。当該手法による時系列クラスタリングにより、予測に有効な価格変動パターンを可視化して把握すること

が可能となる。TOPIX 指数を用いた実証分析とクラスタを可視化した結果、以下の点が確認できた。

- 収益率、正答率ともに k-medoids clustering with IDTW が TOPIX 指数、ベンチマーク手法の両者を上回り、高い予測精度となった。
- k-medoids clustering with IDTW を用いると、予測精度という点で価格変動パターンのクラスタ数は 5 程度であることがわかった。
- 日本市場では有効性が確認できないモメンタム効果が、当月の価格変動が大きい場合に、上昇、下落ともに強く確認できた。

本章までで、価格変動パターンが似ているというシンプルな仮定で予測を行い、具体的にどのようなパターンが有効であったかを検証した。しかしながら、予測に有効な特徴量として株価変動パターン以外の情報も組み込むことで更なる予測精度の改善が見込むことができる。そこで次章では、価格変動パターンとその他の情報を組み合わせた予測手法の提案と検証を行う。

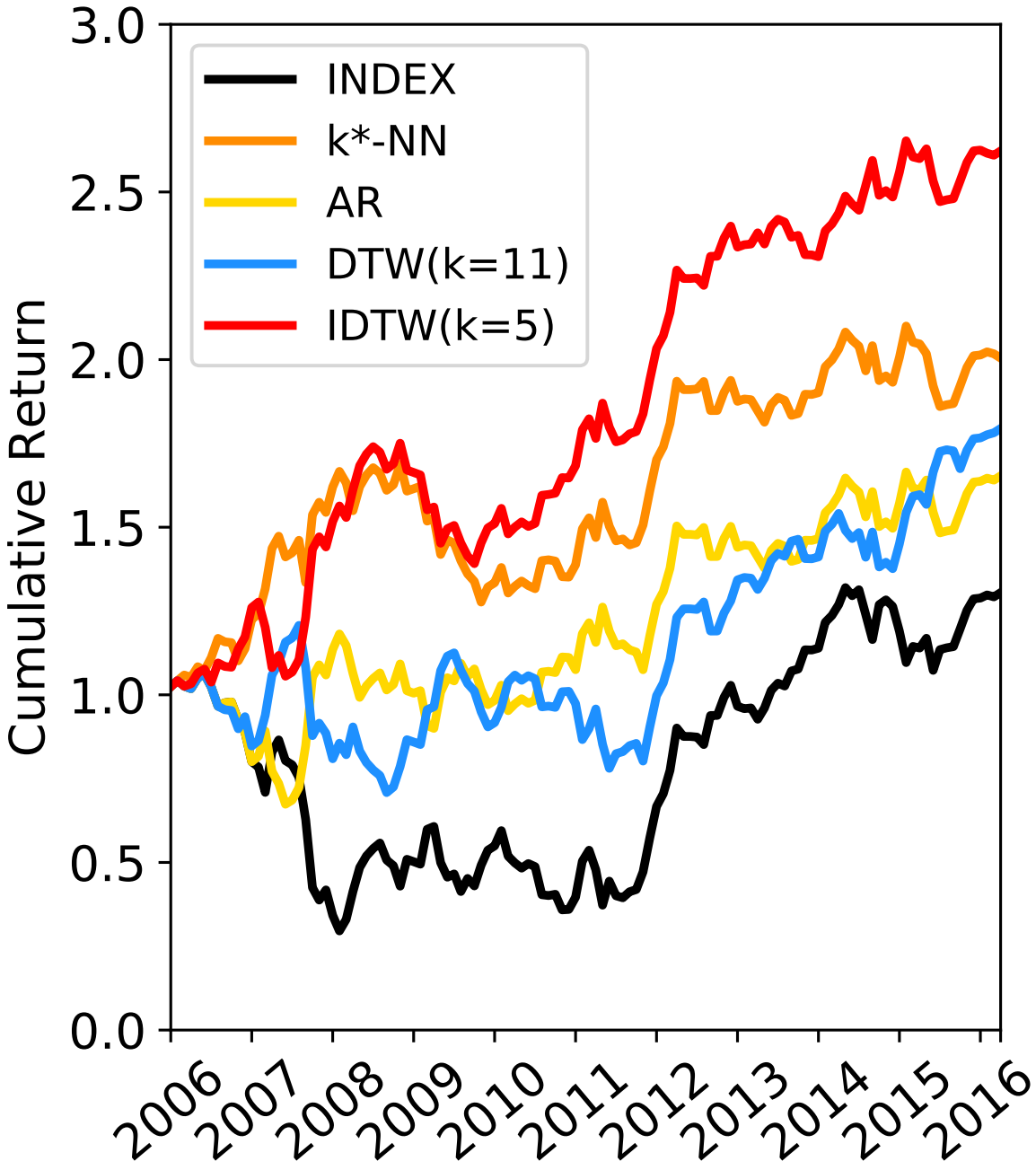

図 4.2　Cumulative return for each prediction method.

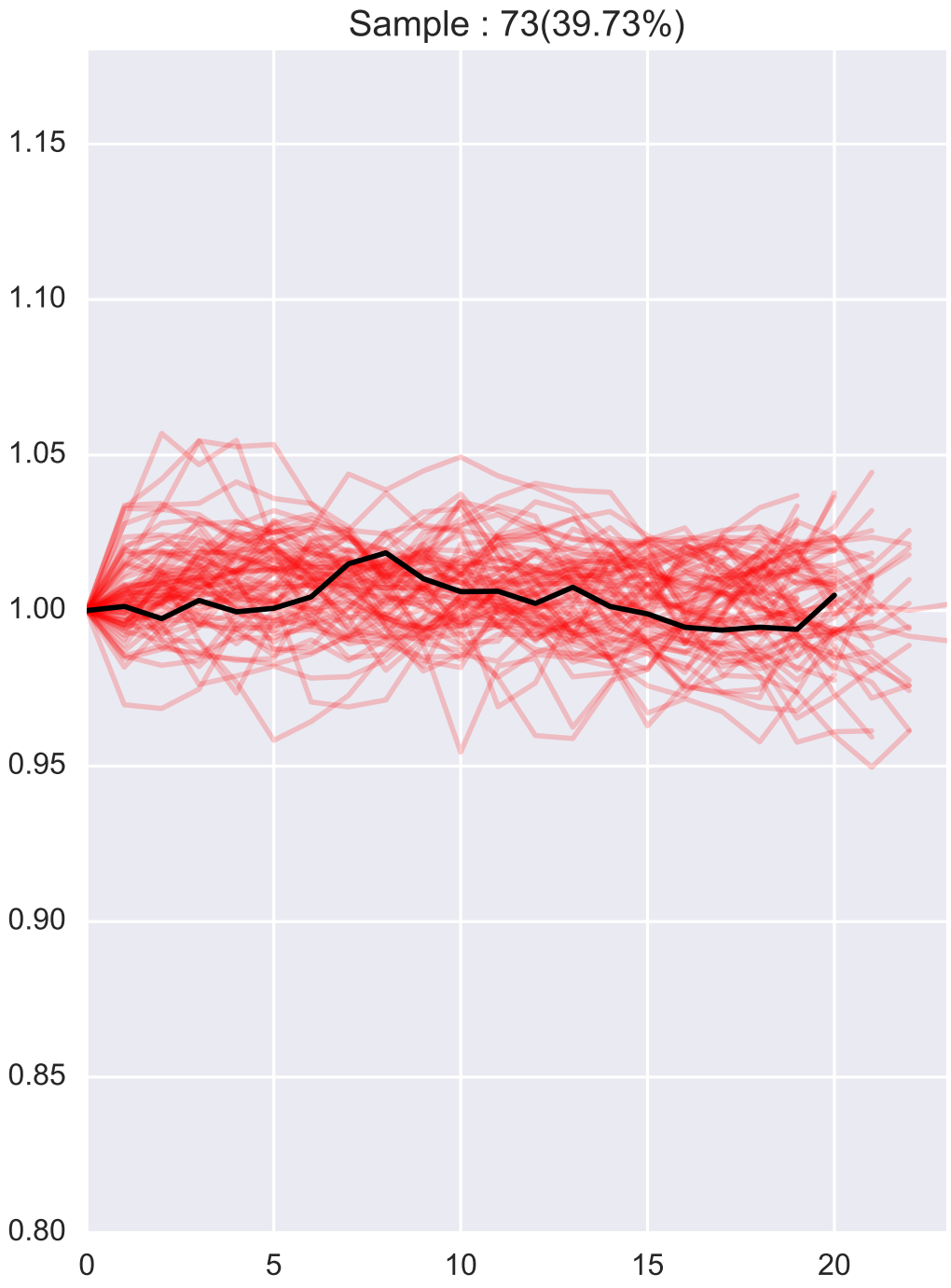

図 4.3　IDTW Based k-medoids clustering as of 2017/3

図 4.4　IDTW Based k-medoids clustering as of 2017/3

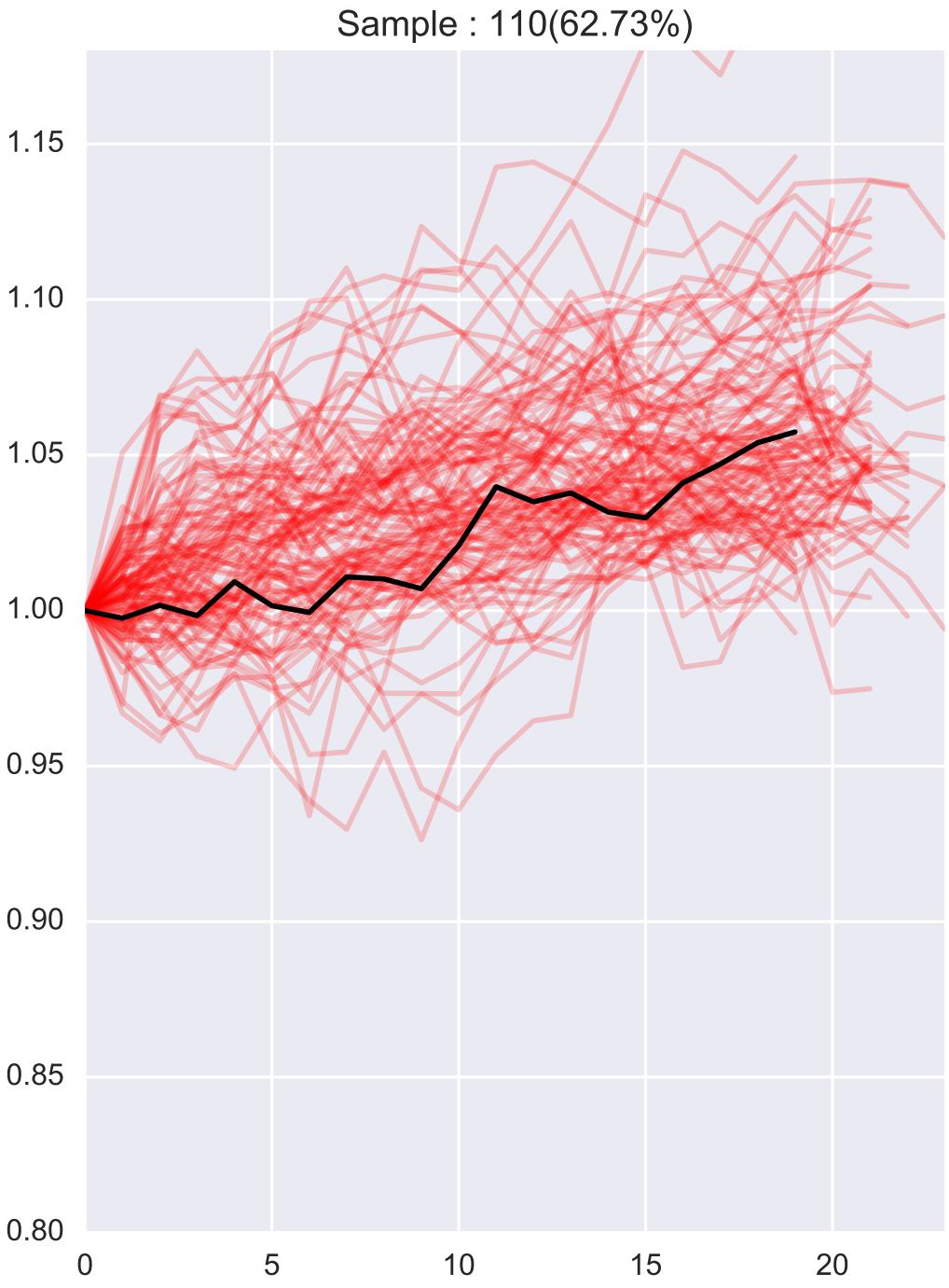

図 4.5　IDTW Based k-medoids clustering as of 2017/3

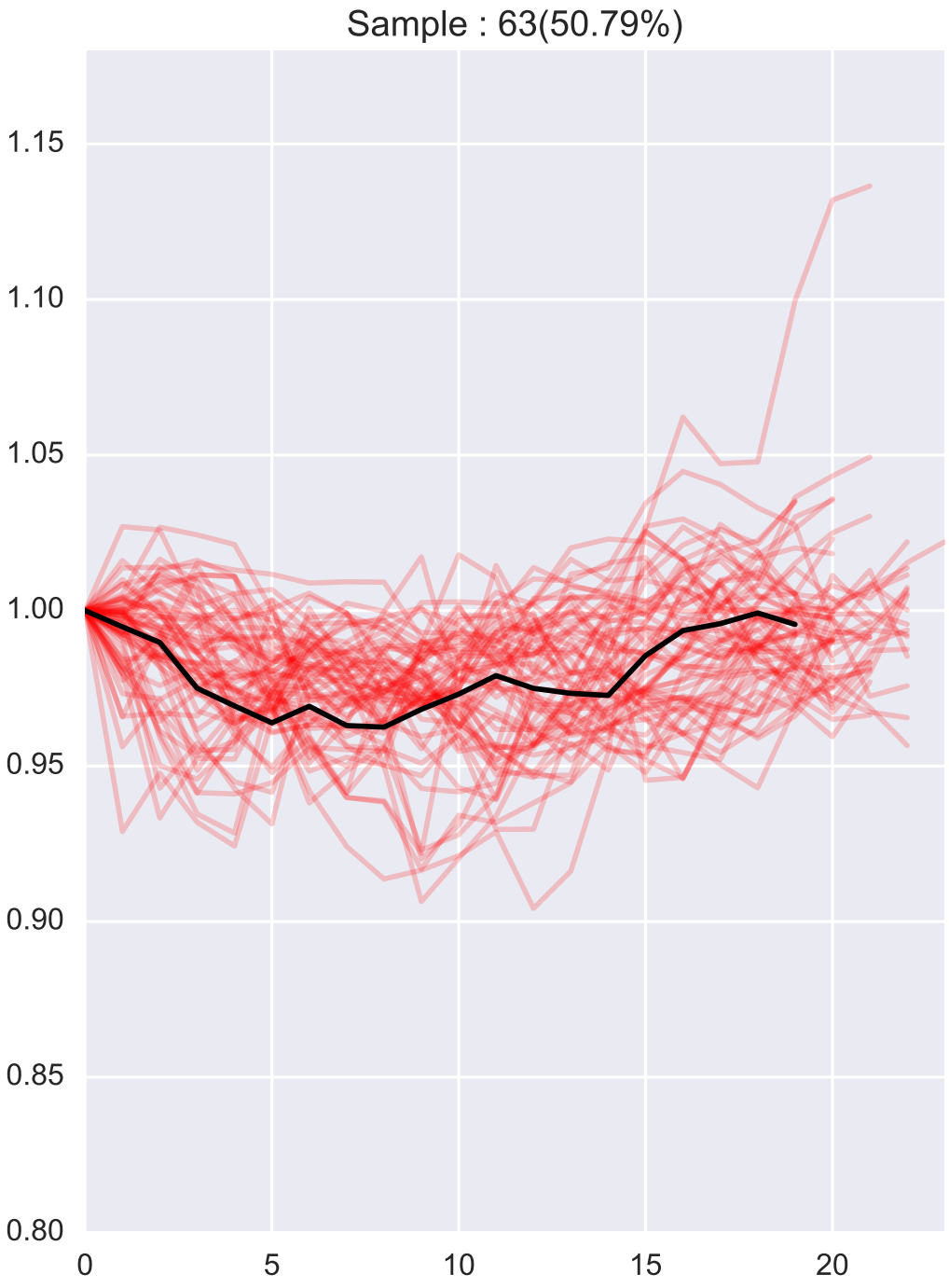

図 4.6　IDTW Based k-medoids clustering as of 2017/3

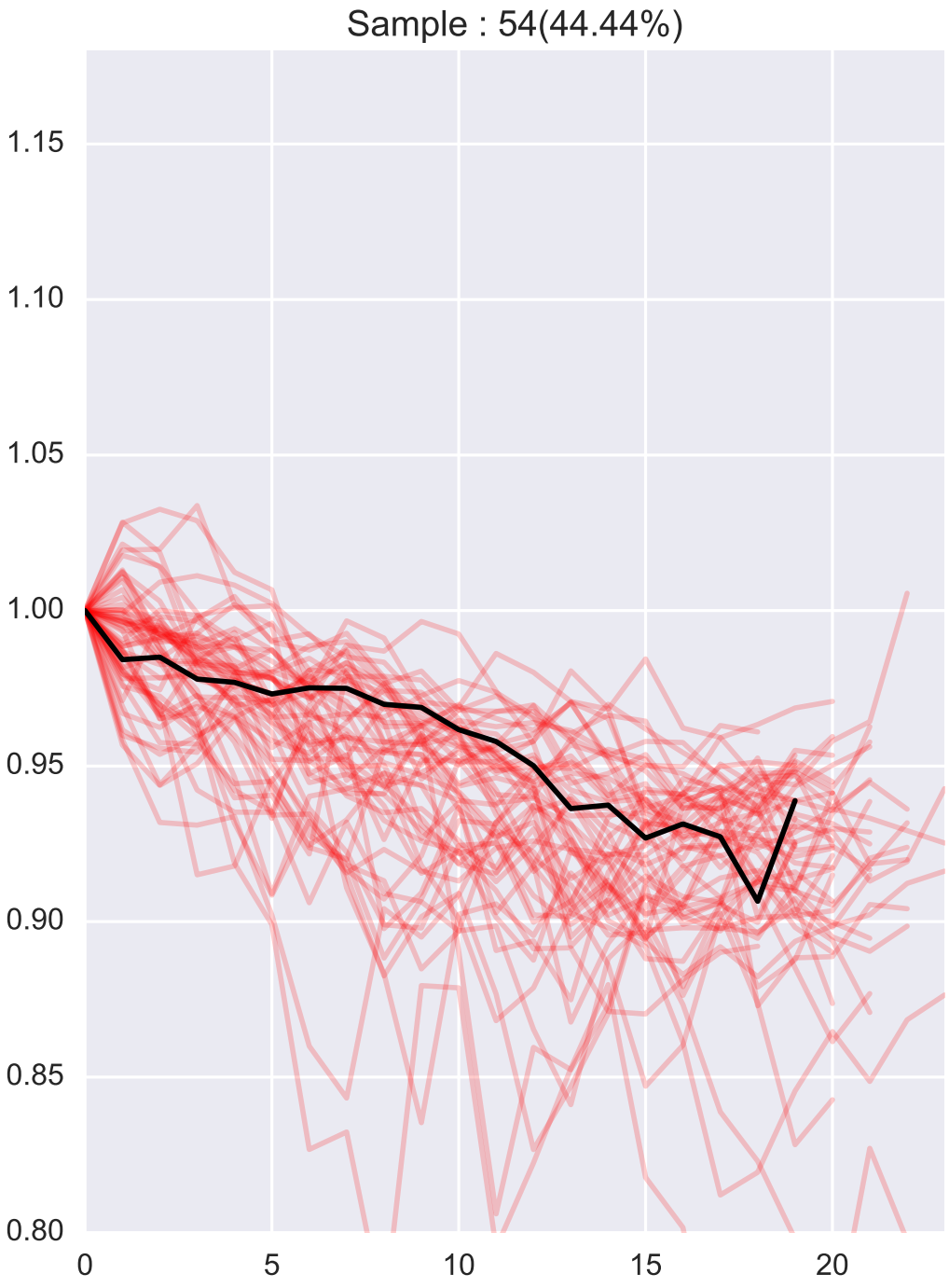

図 4.7　IDTW Based k-medoids clustering as of 2017/3

第 5 章

価格変動パターンとクロスセクションデータを組み合わせた予測手法

5.1　はじめに

前章までに、価格変動パターンは有効な特徴量であって、単純なモメンタムとは異なる変数であることを述べた。本章では、価格変動パターンを明示的に扱った予測モデルを構築し、さらに価格変動パターン以外の属性を予測に用いるための手法を検討する。具体的には、時系列およびクロスセクションの属性を持つデータセットに対する時系列勾配ブースティング決定木を提案する。また、弱学習器として、Yamada et,al(2003)[YSYT03] の時系列決定木に、時系列以外の通常の特徴量も追加した決定木を使用することを提案する。勾配ブースティングは、ブースティング・アルゴリズムの一種である。ブースティングとは、集団学習の枠組みの一つで、複数の弱学習器を統合して全体の学習器を構成する手法である。弱学習器としては、決定木が使われることが多い。ブースティングにおける弱学習器として決定木を用いることは、データの外れ値に強い、数値変数と離散変数や欠損値などを扱いやすいなどの利点がある。

勾配ブースティング木を含むツリーモデルは一般に可読性に優れており、説明責任という観点から実務への適用がしやすい。勾配ブースティング木は決定木を多段に組み合わせることで予測値を導出する手法であり、非常に高い精度を実現できる。その結果として、Kaggle や KDD Cup といった様々なデータ解析コンテストにおいて、最も使用されているアルゴリズムの一つである [CG16]。しかしながら従来の決定木学習手法は時系列属性を想定していないため、株価のような時系列データを含むデータ集合に適用する場合、

データの前処理が必要となる。最も単純な前処理の方法として、時系列データを計測値の平均値や標準偏差などの時系列データのモーメント情報で置き換える方法が考えられる。ただし、この方法は時系列データの構造、すなわち形状を無視しており、例えば形が大きく異なる時系列データを同一視してしまう欠点がある。そこで、勾配ブースティング・アルゴリズムに用いる弱学習器として、以上の問題点を克服した時系列データの形を陽に扱い時系列データ全体を対象とする Yamada et,al(2003)[YSYT03] の基準例分割テストによる分割を用いた時系列決定木を使用する。さらに時系列属性以外の通常の属性も取り込んだ時系列決定木を弱学習器として使用することを提案する。Yamada et,al(2003)[YSYT03] の時系列決定木は相違度基準として DTW を用いているが、金融時系列の分割の際には、時系列データの類似度を計測する手法として、前章までの IDTW を続けて使用する。時系列ペアに対し距離が定義されていることから、前章のように決定木学習手法ではなく、最近傍法 (k-NN や k^*-NN[AL16]) を用いることも考えられる。ただし、最近傍法は怠惰学習であるために学習モデルが存在しないという欠点がある。以上の時系列決定木を弱学習器として用いた時系列勾配ブースティング木を用いて、世界各国の複数の株式指数を対象に実証分析を行う。次章では先行研究である時系列決定木についてレビューする。

5.2 先行研究 (時系列決定木)

5.2.1 決定木

決定木とは分類問題にも回帰問題にも使用可能な教師ありの機械学習手法の一つである。決定木は、データの属性に対して分割を繰り返し適用していく貪欲法に基づくアルゴリズムであって、結果として得られた木は人間がみて妥当性を検証できるため、ノンパラメトリックな手法の中では非常に可読性が高い手法であるといわれている。また説明変数には数値データ、カテゴリデータ等の様々なデータ型が利用可能で、かつ必要な前処理が少ないという利点もある。

決定木は、「根ノード」と呼ばれるデータの集合に対して、「葉ノード」と呼ばれる 2 つの集合にデータを分割し、葉ノードに属するデータのばらつきが少なくなるように分割を行う。学習された結果は、ツリー構造で可視化できるためモデルを直感的に理解しやすい反面、モデルの精度は他のノンパラメトリックな手法と比べて低くなってしまうことが多い。特に回帰問題を扱う決定木を回帰木という。本章では回帰問題を扱うため、回帰木

[LFOS84] について概観する。回帰木は、ある区間内では次のような平均値 $\hat{y}$ を取る回帰関数である。

$$\hat{y} = \frac{1}{|N|} \sum_{i=1}^{N} y_i \tag{5.1}$$

学習は根ノードから、2 つずつ葉ノードを追加していき、木を成長させる。葉ノードを追加する基準としては、ある閾値 θ でサンプルを分割したときの評価関数 H で判断する。この評価関数 H は情報利得とも呼ばれ、直感的にはデータ分割の前後を比較してどれだけ綺麗にデータが分割できたかを表す。

ここで、データセット D は、n 個の説明変数 $\{x_i, i = 1, ..., n\}$ を含み、各サンプル x_i はそれぞれ m 個の属性 $\{a_j, j = 1, ..., m\}$ とクラスラベルまたは目的変数 $\{y_i, i = 1, ..., n\}$ を持つとする。

ここで、最も情報利得の高い属性で葉ノードを分割することを考える。データセット D をある属性の特定の閾値 θ 未満であるデータ $L(a_j, \theta) = \{(y, x)|e_k(a_j) < \theta\}$ と以上であるデータ $R(a_j, \theta) = \{(y, x)|e_k(a_j) \geq \theta\}$ に分割する。そして、次の情報利得 $H(L, R, \theta)$ を計算する。

$$H(L, R, \theta) = I(D) - \frac{1}{|L|} I(L) - \frac{1}{|R|} I(R) \tag{5.2}$$

ここで、$I(\cdot)$ は不純度を定量化した関数であり、分類問題であれば、クラスラベルを c_i としたジニ係数 $I(D) = 1 - \sum_{c_i} \frac{|c_i|}{|D|}^2$ やエントロピー $I(D) = -\sum_{c_i} \frac{|c_i|}{|D|} \log_2 \frac{|c_i|}{|D|}$ を用い、回帰問題であれば標準偏差 $I(D) = \frac{1}{|D|} \sum (y_i - \hat{y})^2$ を用いる。そして、属性と閾値とのすべての組み合わせにおいて、情報利得 H の値を計算し、それが最も小さくなる属性と閾値の組み合わせを葉ノードとして追加する (Algorithm 8)。

木の成長を止める基準としては、

1. 葉ノードに含まれサンプル数が少なくなる。
2. 分割前と分割後で情報利得が減少しなくなる。
3. 特定の深さまで到達する。

のいずれかあるいは複数の基準が用いられる。

Algorithm 8 Split

1: **procedure** SPLIT($\{x_1, ..., x_n\}, \{y_1, ..., y_n\}$)
2: **for each** samples e_k **do**
3: **for each** attribute a_j **do**
4: $L(a_j, \theta) = \{(y, x) | e_k(a_j) < \theta\}$
5: $R(a_j, \theta) = \{(y, x) | e_k(a_j) \geq \theta\}$
6: **end for**
7: **end for**
8: $\theta^* = \arg\min_{\theta} H(L, R, \theta)$
9: **return** best split θ^* and a_j
10: **end procedure**

5.2.2 時系列決定木

時系列決定木は、内部ノードに基準となる属性時系列を持つ決定木であり、基準例分割テストまたはクラスター分割テストによってサンプルを分割していく [YSYT03]。山田ら (2003)[YSYT03, 山田 03] の実験からクラスター分割テストより基準例分割テストの結果が良好であることから、本稿では基準例分割テストを分割手法として用いる (Algorithm 9)。

先ほどの決定木と同様にデータセット D は、n 個の説明変数 $\{x_i, i = 1, ..., n\}$ を含み、各サンプル x_i はそれぞれ m 個の属性 $\{a_j, j = 1, ..., m\}$ とクラスラベルまたは目的変数 $\{y_i, i = 1, ..., n\}$ を持つ。ただし各サンプルの属性 a_j は、それぞれ時間順に値を並べた時系列データである。ここで $ex_i(a_j)$ はデータセットの中の一つのサンプルであり、基準例と呼ぶ。基準例分割テストは、各サンプル x_i の属性 a_j に関する時系列データを $x_i(a_j)$ で表すと、データセット全体を、$DTW(x(a_j), ex_i(a_j)) < \theta_i$ を満たすデータセットから構成される集合 $L(x, a_j, \theta_i) = \{(y, x) | DTW(x(a_j), e_i(a_j)) < \theta_i\}$ とそれ以外の集合 $R(x, a_j, \theta_i) = \{(y, x) | DTW(x(a_j), e_i(a_j)) \geq \theta_i\}$ に分割する。ここで $DTW(x, y)$ は、動的時間伸縮法 (DTW) に基づいた類似度を表す。

基準例分割テストは、すべてのサンプルを基準例 $e_i(a_j)$ と DTW で計測して、ある閾値 θ で類似しているか、あるいは類似していないかという基準を置く。そして閾値 θ を変えながらサンプルの分割を行う。当該分割の評価基準 $H(\cdot)$ としては、通常の決定木で用い

Algorithm 9 StandardExSplit

1: **procedure** STANDARD EXAMPLE SPLIT($\{x_1, ..., x_n\}, \{y_1, ..., y_n\}$)
2: **for each** samples e_k **do**
3: **for each** attribute a_j **do**
4: **for each** samples x_i **do**
5: $L(a_j, \theta) = \{(y, x) | DTW(x_i(a_j), e_k(a_j)) < \theta\}$
6: $R(a_j, \theta) = \{(y, x) | DTW(x_i(a_j), e_k(a_j)) \geq \theta\}$
7: **end for**
8: **end for**
9: **end for**
10: $\theta^* = \arg\min_{\theta} H(L, R, \theta)$
11: **return** best split θ^* and a_j
12: **end procedure**

られている尺度がそのまま使える。

基準例分割テスト (Algorithm 9) は時系列属性のみを考慮するが、時系列属性以外の連続値、名目値を認めるように拡張を行う。すなわち、各属性 a_j は、時間順に値を並べた時系列データ又は系列を持たない通常の連続値、名目属性を持つとする。そして、基準例分割テスト (Algorithm 9) に通常の決定木で使う分割基準を追加する。すると、属性が時系列データの場合、基準例分割テストで分割され、時系列属性以外の場合は通常の閾値で分割する。このアルゴリズムを、クロスセクションデータ込みの基準例分割テストと呼ぶ (Algorithm 10)。

Algorithm 10 の分割方法を下記の表 5.1 を例に説明する。基本方針としては、すべての属性および時系列属性についてラベルを分割して、最も良く分割できる属性をノードに加えていく。はじめに属性 1(a_1) と属性 2(a_2) は通常の決定木と同様に、各属性の値でソートし、情報利得 H と閾値 θ を計算する。次に、時系列属性 (a_3) は各日付ごとに、例えば、1985/1 の時系列属性データを基準例 $e_1(a_3)$ として選択し、1985/2 以降の全ての時系列属性データとの類似度を計算する。そして、類似度の値でソートし、情報利得 H と閾値 θ を計算する。この手順を、1985/2 を基準例 $e_2(a_3)$ として情報利得と閾値を計算し、すべての日付を基準例として同様に情報利得と閾値を計算する。最後に属性、時系列属性の中で最も良く分割 (最大の情報利得 H^*) できる属性をノードに加える。

クロスセクションデータ込みの標準例分割テストによって構築された時系列決定木の例

表 5.1　The sample dateset of the time-series decision tree with cross-section data.

日付	ラベル	属性 1	属性 2
1985/1	Up	0.23	0.30
1985/2	Down	0.73	0.53
1985/3	Up	0.58	0.85
1985/4	Up	0.38	0.91
1985/5	Down	0.86	0.40
1985/6	Up	0.84	0.53

日付	時系列属性
1985/1	0.29
1985/1	0.70
1985/1	0.46
1985/1	0.66
1985/2	0.14
1985/2	0.56
1985/2	0.05
1985/2	0.20
...	...

が図 5.1 である。時系列に対して、通常の特徴量による分類だけでなく、時系列が自身の過去の形状と類似しているか、類似していないか、という形状に基づいた分類が行われている。そのため、どのような判断基準で予測を行ったかについて、直感的に理解が可能である。

5.3　提案手法 - 時系列勾配ブースティング木

ブースティングはアンサンブル学習と呼ばれる手法の一つである。アンサンブル学習とは、弱学習器 (weak learner) と呼ばれるモデルを複数使用することにより強い (精度の良い) モデルを構築する手法をいう。バギングとブースティングという 2 つがアンサンブル学習の代表的な手法である。バギング (Bootstrap Aggregating) とは、ブートストラップサンプリングを繰り返してデータセットを複数生成し、それらを弱学習で学習した結果を合成してより精度の良いモデルを作る学習法である [Bre96]。各弱学習が並列的に学習できる一方で、精度は後述のブースティングには劣るという特徴がある。

一方で、ブースティングは、訓練データで弱分類器を学習し、それを最終的なモデルの一部として逐次的に追加することを繰り返す手法である。バギングとは対照的に、並列処理ができないが精度は優れている。ブースティングの方法の一つとして、Friedman(2001)[Fri01] が勾配ブースディング (Gradient Boosting) を提案した。これは、勾配降下法とブースティング法から構成されている。

先ほどまでと同様に説明変数として $X = \{x_i\}_{i=1}^N$、目的変数として $Y = \{y_i\}_{i=1}^N$ を与え

Algorithm 10 StandardExSplit with Cross-Sectional Data

procedure STANDARD EXAMPLE SPLIT WITH CROSS-SECTIONAL DATA($\{x_1, ..., x_n\}, \{y_1, ..., y_n\}$)
  for each samples e_k **do**
    for each attribute a_j **do**
      if attribute a_j is time-series **then**
        for each samples x_i **do**
          $L(a_j, \theta) = \{(y, x) | DTW(x_i(a_j), e_k(a_j)) < \theta\}$
          $R(a_j, \theta) = \{(y, x) | DTW(x_i(a_j), e_k(a_j)) \geq \theta\}$
        end for
      else
        $L(a_j, \theta) = \{(y, x) | e_k(a_j) < \theta\}$
        $R(a_j, \theta) = \{(y, x) | e_k(a_j) \geq \theta\}$
      end if
    end for
  end for
  $\theta^* = \arg\min_{\theta} H(L, R, \theta)$
  return best split θ^* and a_j
end procedure

たとき、勾配ブースティングは、真の関数 $F(x_i)$ に対する損失関数 $L(\cdot)$ を最小にする F^* を学習する。

$$F^* = \arg\min_{F} \sum_{i} L(y_i, F(x_i)) \tag{5.3}$$

ここでブースティングにより、パラメータ α_i を持つ弱学習器 $f_i(X; \alpha_i)$ の重み付き和を用いて真の関数 $F(x_i)$ を近似する。

$$F(X) = \beta_0 f_0(X; \alpha_0) + \beta_1 f_1(X; \alpha_1) + ... + \beta_M f_M(X; \alpha_M) \tag{5.4}$$

ただし、β_m は弱学習器の重み、M は弱学習器の数である。以上から、次のように学習器を構成する。

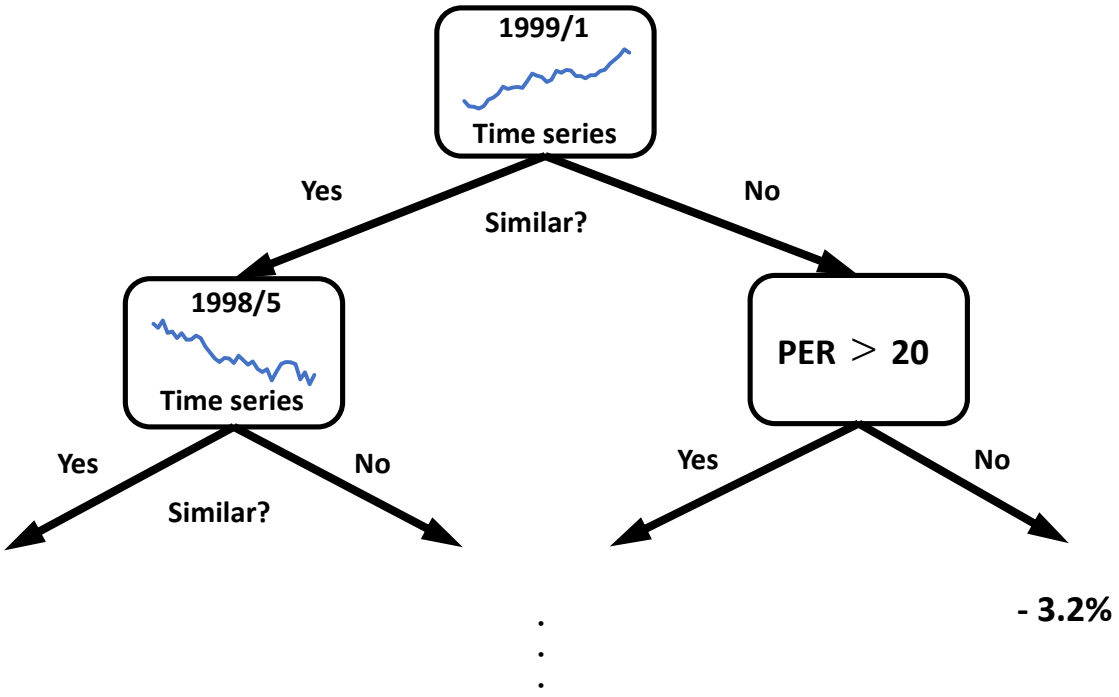

図 5.1　The illustration of the time-series decision tree with cross-section data.

$$\arg\min_{\alpha,\beta} \sum_i L(y_i, F_m(x_i)) \tag{5.5}$$

$$s.t. \qquad F_m(x_i) = F_{m-1}(x_i) + \beta_m f(x_i; \alpha_m) \tag{5.6}$$

勾配ブースティングでは、任意の損失関数 $L(\cdot)$ に対して、以下の 2 段階で近似的に最小化する。

- 擬似残差と呼ばれる $\hat{y}_i = [\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$ を教師データとして、弱学習器 $f(X; \alpha_m)$ のパラメータ α_m を推定する。
- 求めた弱学習器を用いて、重み β_i の最適化を $\arg\min_{\beta} \sum_i L(y_i, F_m(x_i))$ として実行する。

任意の損失関数 $L(\cdot)$ について、勾配ブースティングアルゴリズムを導出することができるが、特に損失関数を二乗誤差関数とした場合、すなわち $L(y_i, F(x_i)) = \frac{1}{2}(y_i - F(x_i))^2$ の場合は、疑似残差は $\hat{y}_i = y_i - F_{m-1}(x_i)$ となり、$m-1$ 番目までの弱学習器の予測と、

Algorithm 11 Gradient Boosting Tree

procedure GRADIENT BOOSTING TREE(y, x)

▷ Initialize F_0 with a time-series decision tree

$F_0(x) = \text{FitTimeSeriesTree}(y, x)$

for $m = 1$ to M **do**

$r_{im} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$

▷ Fit a time-series decision tree to predict targets r_{im}

$f(x) = \text{FitTimeSeriesTree}(r_{im}, x)$

$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} L(y_i, F_{m-1}(x) + \beta f(x_i))$

$F_m(x) = F_{m-1}(x) + \beta_m f(x)$

end for

return $F(x) = \sum_{m=1}^{M} F_m(x)$

end procedure

ラベル y_i との差になる。つまり、損失関数 $L(\cdot)$ が減少する方向 (勾配) を次の弱学習器のラベルとして与え、損失関数を減少させていく。

勾配ブースティングのための弱学習器としては決定木 (回帰木) が用いられることが多く、これを特に勾配ブースティング木と呼ぶ。ここでは、勾配ブースティング木を構成する弱学習器として前節の時系列決定木を使用する。これを時系列勾配ブースティング木 (図 5.2 および Algorithm 11) と呼ぶ。

5.4 実証分析

ここでは、提案手法である時系列勾配ブースティング木の有効性を、各国の株式指数を対象に行う。使用する株式指数は TOPIX、S&P500、FTSE100、DAX30、CAC40 である。比較のため、時系列決定木 [YSYT03](TSDT)、クロスセクションデータ込みの時系列決定木 (TSDT+CS)、時系列勾配ブースティング木 (TSGBT)、クロスセクションデータ込みの時系列勾配ブースティング木 (TSGBT+CS) の 4 つの手法を用いる。

時系列決定木を構築するための時系列データ間の類似度としては、前章までと同様に毎日の株価で構成された毎月の株価変動パターン間の IDTW で計測する。TSDT + CS および TSGBT + CS で使用する時系列データ以外のクロスセクションデータは、表 5.2 の通りである。これらのデータは、実務および様々な研究においてよく使用される

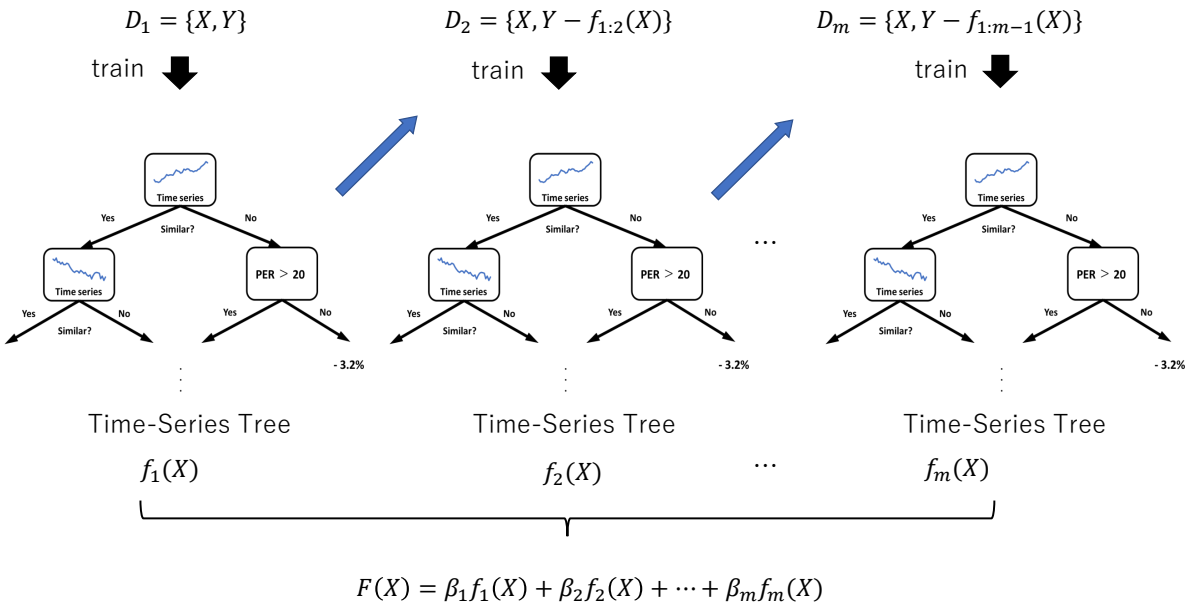

図 5.2　The illustration of the time-series gradient boosting decision tree with cross-section data.

表 5.2　Cross-section data used in TSDT+CS and TSGBT+CS.

Data	Description
PER	Net Income / Market Value
PBR	Net Assets / Market Value
MOM	Stock returns over the past 12 months except for past month
DIV	Dividends paid in a year / Market Value
ROE	Net income/Net Assets

[VFH12]。

価格データや表 5.2 のデータはすべて Bloomberg[*1]より取得した。各指数のデータ期間はデータの取得可能な時点が異なるため、1993/8(TOPIX)、1990/2(SP500)、2001/7(DAX)、2001/6(CAC) および 2000/1(FTSE) から全指数共通で 2018/6 までである。最初の学習に過去 10 年分 (120 サンプル) を使用する。テスト期間は、2003/8(TOPIX)、2000/2(SP500)、2011/7(DAX)、2011/6(CAC) および 2010/1(FTSE) から、全指数共通で 2018/6 までである。R の gbm([RR04]) と rpart([TAR10]) パッケージ

[*1] ティッカーはそれぞれ TPX Index、SPX Index、UKX Index、DAX Index、CAC Index である。

を参照し、4 つの手法を実装した。gbm パッケージにおける木の深さ (interaction.depth) を 3 に変更するが、その他のすべてのパラメータはパッケージのデフォルト値を用いた。具体的な分析手順は次の通りである。

Step1

$t-1$ 月までのデータを使いて、各モデルの学習を行う。例えば、t 月が 2003 年 7 月の場合、2003 年 8 月 ($t+1$ 月) の TOPIX 指数を予測するために、2003 年 6 月 ($t-1$ 月) までの過去 120 ヶ月のデータを用いて、4 つのモデルの学習を行う。

Step2

t 月のデータを与え、翌月のリターンを予測する。例として、Step 1 で学習した図 5.1 のクロスセクションデータを含む時系列決定木 (TSDT+CS) を用いて TOPIX 指数の 2003 年 7 月のデータで、2003 年 8 月の株価変動を予測する。つまり、2003 年 7 月の日々の株価を時系列属性として持ち、通常の属性として、2003 年 7 月末時点の表 5.2 の属性を持つデータを用いて予測する。最初のノードでは、2003 年 7 月と 1999 年 1 月の時系列データ (日々の株価) 間の類似性に基づいて、分岐を進める。2003 年 7 月の時系列がノードの 1999 年 1 月の株価変動と類似していないとすると、右に分岐し、次に 2003 年 7 月の PER が 20 より大きいかどうかで分岐を進める。このように分岐を進め、予測リターンを求める。

Step3

翌月の予測リターンが正ならば、各指数を 1 単位買い、負ならば 1 単位売りとして収益計算する。

Step4

$t = t+1$ に進めて Step 1 に戻る。

5.4.1 結果と考察

各モデルの評価として、MAE と RMSE を用いて精度を評価する。また、収益性の観点からトータルの収益率を用いて評価する。

表 5.3 と表 5.4 はそれぞれ MAE と RMSE を表す。右端が各手法の指数毎の平均値を示している。各テーブル内の太字は 4 つの手法のうち最も精度の良いものを示している。

表 5.3 と表 5.4 の両方から、全ての指数の平均では TSGBT + CS が 4 つの手法の中で

表 5.3　The average MAEs of all years for each method. The out-of-sample period is to June 2018. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the four methods.

MAE	TOPIX	SP500	DAX	CAC	FTSE	avg
TSDT	5.038	3.695	4.852	4.960	3.754	4.460
TSDT+CS	4.399	3.522	4.129	3.300	2.872	3.644
TSGBT	3.752	3.056	3.673	3.360	2.650	3.298
TSGBT+CS	**3.495**	**2.861**	**3.098**	**2.766**	**2.188**	**2.882**

表 5.4　The average RMSEs of all years for each method. The out-of-sample period is to June 2018. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the four methods.

RMSE	TOPIX	SP500	DAX	CAC	FTSE	avg
TSDT	6.195	4.899	6.417	5.921	4.666	5.619
TSDT+CS	5.554	4.474	5.243	4.085	3.677	4.607
TSGBT	**4.955**	4.146	4.819	4.202	3.324	4.289
TSGBT+CS	4.961	**4.036**	**4.304**	**3.524**	**2.974**	**3.960**

MAE と RMSE で測って最も精度が良いことがわかる。時系列決定木に対して勾配ブースティング法を用いることで、MAE で計測して 1.162 ポイント、RMSE で計測して 1.33 ポイント全ての指数の平均で改善した。同様に、勾配ブースティング木にクロスセクションデータを加えることで、平均して MAE で計測して 0.416 ポイント、RMSE で計測して 0.329 ポイント改善した。従って、勾配ブースティング法とクロスセクションデータの導入の両方が精度改善に寄与するといえる。表 5.6 は、4 つの手法の合計収益率を表している。収益性の面でも、全ての指数の平均では TSGBT + CS が 4 つの手法の中で最も収益率が高いことがわかる。時系列決定木に対して勾配ブースティング法を用いることで、126%、同様に、勾配ブースティング木にクロスセクションデータを加えることで、69%、収益率は全ての指数の平均で上昇した。

以上から、勾配ブースティングによる効果とクロスセクションデータを入れる効果の両方において精度、収益性の改善が確認できた。

次に、重要度比 (Importance Rate;IR) を使用して、モデルがどういう変数を重要視しているか可視化し確認する。各分岐においてこれが最大となる変数とその閾値を見つけるのが決定木のアルゴリズムであるが、最終的な学習モデルにおいて、ある変数 x_i が平均的

表 5.5　The average Accuracys of all years for each method. The out-of-sample period is to June 2018. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the four methods.

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%**

表 5.6　The total returns[%] for each method. The out-of-sample period is to June 2018. The rightmost column is the total mean for each method. The bold values are the most accurate measurements of the four methods.

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**

にどれだけ gain を与えたのかという指標が IR である。IR とは、ある変数 x_i が平均的にどれぐらい評価基準を改善しているのかの全分岐点についての平均であり、これによって変数の重要性を評価する指標である。IR は、合計が 100% になるように標準化している。表 5.7 は、2018 年 6 月予測時点における TSGBT + CS の IR 値のトップ 20 を示している。YYYY/MM は、標準例分割テストに使用された価格変動パターンを表している。すべての株式指数の予測において、クロスセクションデータが上位 20 に現れることが確認できる。

5.4.2　まとめ

本論文では、時系列およびクロスセクションの属性を持つデータセットに対する時系列勾配ブースティング決定木を提案した。勾配ブースティング・アルゴリズムに用いる弱学習器として、時系列データの形を陽に扱い時系列データ全体を対象とする [YSYT03] の基準例分割テストによる分割を用いた時系列決定木を使用した。また弱学習器として、時系列決定木を用いるが、先行研究における時系列決定木 [YSYT03] は時系列属性以外のデー

表 5.7　Feature and importance rate (IR) of TSGBT+CS on 2018/5.

TOPIX		SP500		DAX		CAC		FTSE	
Feature	IR	Feature	IR	Feature	IR	Feature	IR	Feature	IR
MOM	6.96	DIV	6.50	PBR	11.58	MOM	7.84	2008/12	9.91
DIV	4.99	1998/10	5.29	2001/08	8.95	2008/09	7.66	PER	7.96
1998/02	4.20	ROE	5.25	2002/08	7.14	PBR	6.49	2010/09	5.21
2005/03	3.65	MOM	5.12	PER	5.88	2008/06	4.79	DIV	5.09
PBR	3.09	PER	4.73	2009/02	4.85	2009/01	4.64	2004/03	3.95
2004/02	3.09	2000/02	4.37	2006/09	4.36	DIV	4.63	2003/11	3.64
1995/01	2.91	1997/11	3.59	2007/06	3.57	2010/04	4.06	2009/01	2.55
2003/04	2.81	2002/06	3.56	2010/05	2.46	2006/07	2.68	MOM	2.15
2002/03	2.76	2003/01	3.09	2007/10	2.23	2009/03	2.66	2006/02	2.12
2000/01	2.52	1992/06	2.55	2005/10	2.21	2006/09	2.43	2008/08	1.99
1996/01	2.48	1994/11	2.33	2002/02	2.19	2003/05	2.37	2007/11	1.99
1997/10	2.43	1997/10	2.18	2003/03	2.19	2008/11	2.05	2009/06	1.97
1994/07	2.23	1999/08	2.13	MOM	2.13	2004/09	2.02	2008/05	1.97
2006/01	2.01	1999/07	2.12	2002/06	2.11	2001/12	2.00	2009/10	1.96
1996/02	1.80	1991/01	2.00	2002/10	1.85	2010/02	1.85	2008/06	1.88
1993/08	1.77	2000/07	1.84	2003/09	1.83	PER	1.84	2008/10	1.84
1997/02	1.76	1999/10	1.76	2003/11	1.73	2011/09	1.81	2011/03	1.71
1997/05	1.67	1999/04	1.74	DIV	1.69	2007/09	1.80	2003/03	1.70
1995/08	1.48	2000/03	1.71	2008/01	1.65	2003/09	1.61	ROE	1.61
2004/04	1.44	1996/06	1.69	ROE	1.61	ROE	1.57	2003/08	1.56

タを考慮できないため、時系列属性以外のデータを組み込んだ時系列決定木を提案した。代表的な各国の株式指数を対象とした実証分析により次のような結果が得られた。

- 時系列決定木に対して勾配ブースティングを用いることで予測精度と収益率が各指数の平均で改善。
- 同様に時系列決定木に対してクロスセクションのデータを組み込むことで予測精度と収益率が各指数の平均で改善。
- TSGBT+CS は予測精度、収益性ともに各指数の平均で最も良い。

当然ながら、株式指数に対して説明力を持つであろう特徴量は本章で検討したもの以外に多数あると考えられ、それらを組み込むことでさらなる予測精度改善が期待できる。また、株式指数以外の個別銘柄、あるいは債券や為替などの他のアセットへ本手法を適用し、

その予測精度を検証することが今後の課題である。金融時系列以外の時系列データセットについての実験評価については [今村 18a] で議論し、データセットが数百以上のケースでは時系列勾配ブースティング木は、単純な時系列決定木と DTW+k-NN をデータの種類を問わずに上回る結果となった。

第 6 章

総括と今後の研究展望

株価の予測可能性は、学術的にも実務的にも重要な研究テーマである。実務的には、株価の予測はファンダメンタル分析とテクニカル分析に基づき行われてきた。テクニカル分析はその有効性は実務的にも学術的にも認められつつも、主観性、恣意性が強いものであるとの指摘を多く受けている。本研究は、テクニカル分析の哲学・論理である価格変動パターンという観点に着目しながらも、客観的、機械的な予測手法を開発することを目的とする。すなわち、現在の価格変動パターンが過去のある時点と似ていれば、そのパターンを用いて将来の価格の予測ができるだろうという仮説を基本とし、このような幾何学的な価格変動パターンを機械的に抽出し、予測へ活用し、その有効性を検証した。

はじめに第 2 章において、実務的なファンダメンタル分析、テクニカル分析という観点ではなく、株価予測に関する学術的な先行研究を次の観点から整理した。株価予測の先行研究はクロスセクション分析と時系列分析という方法論の観点、テクニカルとファンダメンタルというデータの観点、そしてパラメトリックとノンパラメトリックという手法の観点がある。本研究の位置付けは、時系列分析の観点から、データに過去の価格変動パターンというテクニカルデータを用いた、ノンパラメトリックな手法の提案である。先行研究として、データに市場データである価格を用いた研究は多数あるものの、価格変動パターンそのものを扱った先行研究は少ない。また、予測手法としてもノンパラメトリックな手法を用いるが、NN や SVM のような目的変数の予測過程が理解し辛く、各説明変数の目的変数に対する重要度・寄与度などが評価できないモデルではなく、シンプルな k-Nearest Neighbor 法や決定木をベースとした可読性、解釈性の高いアルゴリズムをベースとした。

第 3 章では、各国の代表的な株式指数に対して、アルゴリズムにより機械的に抽出され

た過去の価格変動パターンが、将来の株価予測に有効であるかをシンプルな手法を用いて検証した。音声認識の分野で時系列データの類似度計測に用いてきた DTW を、テクニカル分析の観点から IDTW として改良し、価格変動パターンの抽出に使用した。DTW は時系列間の長さが異なる場合でも使用でき、かつ時間方向のずれを許容するため、人間の直感にあった時系列間の類似度を計測できるといわれている。そしてノンパラメトリックな予測に手法の中で最もシンプルな k-NN 法の改良である、k^*-NN 法と IDTW を組み合わせた株価予測手法を提案した。提案手法を用いて各国の代表的な株価指数を予測したところ、過去の価格変動パターンは有効な特徴量であることが確認できた。また株価変動パターンは単純なモメンタム、リバーサル戦略以上の収益機会が存在することを合わせて確認した。一方で、DTW と k-NN の組み合わせは時系列データマイニングにおいて非常に有効な手法であるが、株価に対してそのまま適用しても予測精度は低いことが確認できた。株価のような金融時系列に対して、DTW を用いた価格変動パターンを用いて予測に有効なアルゴリズムを作成したこと、単純なモメンタム戦略とは相関が低く、より多くの収益が獲得できることを実証したことが本章の貢献である。

第 3 章では、価格変動パターンが似ているというシンプルな仮定で予測を行ったが、どのような価格変動パターンが予測に有効であったかは不明である。そこで第 4 章では、3 章で有効であると確認できた過去の価格変動パターンを、クラスタリング手法を用いて可視化をこころみた。日本の株価指数 TOPIX を対象に可視化を行い、予測力という観点から、どのような価格変動パターンがクラスタリングによって抽出されたかを確認した。前章までと同様に、月間の日次株価変動はデータ数 (営業日) が月毎に異なるため、単純なベクトル空間上のユークリッド距離を用いたクラスタリング手法 (例えば、k-means 法) は適用できない。よって、株価を対象に異なるデータ数においても、欠損値として値の挿入および削除を行うことなく、より自然にデータ間の類似性を測定し、そして対応するクラスタリング方法を適切に組み合わせる必要がある。そこで価格変動パターンの可視化のため、IDTW を類似度として用いた k-medoids 法によるクラスタリング手法を提案し、最も収益性や予測精度の高いクラスタ数 (5 つ) でクラスタリングを行った。株価変動パターンのクラスタを抽出したところ、当月の株価変動が大きい場合に、上昇、下落ともに強いモメンタム効果が可視化により確認できた。

第 4 章までは、価格変動パターンの類似性に着目した予測を行い、具体的にどのようなパターンが有効であったかを検証した。しかしながら、予測に有効な特徴量として株価変

動パターン以外の情報も組み込むことで更なる予測精度の改善を見込むことができると考えられる。第 5 章では、前章までに確認した予測に有効な株価変動パターンと、その他のテクニカル、ファンダメンタルデータを組み合わせ、可読性の高いモデルを構築した。具体的には、木構造に基づく予測手法である時系列勾配ブースティング木を提案した。時系列勾配ブースティング木を構成する弱学習器として、時系列決定木を用いるが、先行研究における時系列決定木は時系列属性以外のデータを考慮できないため、時系列属性以外のデータを組み込んだ時系列決定木を提案した。各国の株式指数を用いた実証分析の結果、単純な時系列決定木に対して勾配ブースティングを用いることで予測精度が改善し、また、時系列決定木に対してクロスセクションのデータを組み込むことでも予測精度が改善することが確認できた。

今後の研究課題については以下のとおりである。

1. 第 3 章の方法を個別銘柄に適用し、過去と最も類似した局面からの予測リターンを、単純なモメンタム戦略と同様にクロスセクションで有効なファクターであるかどうか検証すること。
2. 第 3 章、第 5 章の方法を株式指数以外の個別銘柄や為替、債券などその他の金融時系列に対しても同様のアプローチで予測が可能かどうかを検証すること。また第 3 章、第 5 章のどちらがより精度、収益性が優れているのかを比較すること。
3. 本研究では、株価変動パターンを単一の指数で判断したが、これを複数の指数や銘柄を用いて多変量での株価変動パターンを定義し、その効果を検証すること。
4. 本研究における手法はすべて DTW に基づいているが、これは 2 つの時系列データのデータ数を M, N とすると $O(MN)$ の計算量である。そのため、データ数が大きくなると計算負荷が非常に高くなるため、効率的計算法を開発すること。

謝辞

本論文の作成にあたり終始適切な助言を賜り、また丁寧に指導して下さった吉田健一教授に感謝します。また、研究のアプローチや考察の方法などについて適切な助言をいただいた倉橋節也教授、山田雄二教授にも心より感謝申し上げます。最後に、これまで私をあたたかく応援してくれた両親、明るく励まし続けてくれた妻の奈津美、息子の瑞貴、娘の美咲に心から感謝します．本当にありがとうございました。

参考文献

[ACF13] Sandro C Andrade, Vidhi Chhaochharia, and Michael E Fuerst. "sell in may and go away" just won' t go away. *Financial Analysts Journal*, Vol. 69, No. 4, pp. 94–105, 2013.

[AHXZ06] Andrew Ang, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang. The cross-section of volatility and expected returns. *The Journal of Finance*, Vol. 61, No. 1, pp. 259–299, 2006.

[AHXZ09] Andrew Ang, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang. High idiosyncratic volatility and low returns: International and further us evidence. *Journal of Financial Economics*, Vol. 91, No. 1, pp. 1–23, 2009.

[AJBEB01] Carroll D Aby Jr, Nat R Briscoe, R Stephen Elliott, and Andrew Bacadayan. Value stocks: A look at benchmark fundamentals and company priorities. *Journal of Deferred Compensation*, Vol. 7, No. 1, pp. 20–20, 2001.

[AL16] Oren Anava and Kfir Levy. k*-nearest neighbors: From global to local. In *Advances in Neural Information Processing Systems*, pp. 4916–4924, 2016.

[AMP13] Clifford S Asness, Tobias J Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. *The Journal of Finance*, Vol. 68, No. 3, pp. 929–985, 2013.

[AN18] Masaya Abe and Hideki Nakayama. Deep learning for forecasting stock returns in the cross-section. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 273–284. Springer, 2018.

[AT14] Saeed Aghabozorgi and Ying Wah Teh. Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applica-*

tions, Vol. 41, No. 4, pp. 1301–1314, 2014.

[AV09] George S Atsalakis and Kimon P Valavanis. Surveying stock market forecasting techniques–part ii: Soft computing methods. *Expert Systems with Applications*, Vol. 36, No. 3, pp. 5932–5941, 2009.

[Ban81] Rolf W Banz. The relationship between return and market value of common stocks. *Journal of financial economics*, Vol. 9, No. 1, pp. 3–18, 1981.

[Bas77] Sanjoy 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.

[Bas83] Sanjoy Basu. The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. *Journal of financial economics*, Vol. 12, No. 1, pp. 129–156, 1983.

[BB12] Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, Vol. 191, pp. 192–213, 2012.

[BC94] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, Vol. 10, pp. 359–370. Seattle, WA, 1994.

[BC96] Randall S Billingsley and Don M Chance. Benefits and limitations of diversification among commodity trading advisors. *Journal of Portfolio Management*, Vol. 23, No. 1, p. 65, 1996.

[Bel54] Richard Bellman. The theory of dynamic programming. Technical report, RAND Corp Santa Monica CA, 1954.

[Bha88] Laxmi Chand Bhandari. Debt/equity ratio and expected common stock returns: Empirical evidence. *The journal of finance*, Vol. 43, No. 2, pp. 507–528, 1988.

[BJ02] Sven Bouman and Ben Jacobsen. The halloween indicator," sell in may and go away": Another puzzle. *The American Economic Review*, Vol. 92, No. 5, pp. 1618–1635, 2002.

[BJCS89] James N Bodurtha Jr, D Chinhyung Cho, and Lemma W Senbet. Economic

forces and the stock market: An international perspective. 1989.

[BJRL15] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[BLB+17] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, Vol. 31, No. 3, pp. 606–660, 2017.

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, Vol. 2, No. 1, pp. 1–8, 2011.

[Boa] European Data Protection Board. General data protection regulation. `https://gdpr-info.eu/`.

[Bol86] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, Vol. 31, No. 3, pp. 307–327, 1986.

[Bre96] Leo Breiman. Bagging predictors. *Machine learning*, Vol. 24, No. 2, pp. 123–140, 1996.

[Car97] Mark M Carhart. On persistence in mutual fund performance. *The Journal of finance*, Vol. 52, No. 1, pp. 57–82, 1997.

[CBS+16] Rodolfo C Cavalcante, Rodrigo C Brasileiro, Victor LF Souza, Jarley P Nobrega, and Adriano LI Oliveira. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, Vol. 55, pp. 194–211, 2016.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM, 2016.

[CH+67] Thomas M Cover, Peter E Hart, et al. Nearest neighbor pattern classification. *IEEE transactions on information theory*, Vol. 13, No. 1, pp. 21–27, 1967.

[CLM+97] John Y Campbell, Andrew W Lo, Archie Craig MacKinlay, et al. *The econo-*

metrics of financial markets, Vol. 2. princeton University press Princeton, NJ, 1997.

[Coe12] Mariana Sátiro Coelho. *Patterns in financial markets: Dynamic time warping*. PhD thesis, NSBE-UNL, 2012.

[Cr33] Alfred Cowles 3rd. Can stock market forecasters forecast? *Econometrica: Journal of the Econometric Society*, pp. 309–324, 1933.

[CRR86] Nai-Fu Chen, Richard Roll, and Stephen A Ross. Economic forces and the stock market. *Journal of business*, pp. 383–403, 1986.

[DM16] Kent Daniel and Tobias J Moskowitz. Momentum crashes. *Journal of Financial Economics*, Vol. 122, No. 2, pp. 221–247, 2016.

[EG87] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pp. 251–276, 1987.

[EK95] Robert F Engle and Kenneth F Kroner. Multivariate simultaneous generalized arch. *Econometric theory*, Vol. 11, No. 1, pp. 122–150, 1995.

[ELL93] BS Everitt, S Landau, and M Leese. Cluster analysis. 1993. *Edward Arnold and Halsted Press,*, 1993.

[Eng82] Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pp. 987–1007, 1982.

[Eng02] Robert Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, Vol. 20, No. 3, pp. 339–350, 2002.

[Fam70] Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, Vol. 25, No. 2, pp. 383–417, 1970.

[fam15] A five-factor asset pricing model. *Journal of Financial Economics*, Vol. 116, No. 1, pp. 1 – 22, 2015.

[FB66] Eugene F Fama and Marshall E Blume. Filter rules and stock-market trading. *The Journal of Business*, Vol. 39, No. 1, pp. 226–241, 1966.

[FF88] Eugene F Fama and Kenneth R French. Dividend yields and expected stock returns. *Journal of financial economics*, Vol. 22, No. 1, pp. 3–25, 1988.

[FF92] Eugene F Fama and Kenneth R French. The cross-section of expected stock returns. *the Journal of Finance*, Vol. 47, No. 2, pp. 427–465, 1992.

[FF93] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, Vol. 33, No. 1, pp. 3–56, 1993.

[FF12] Eugene F Fama and Kenneth R French. Size, value, and momentum in international stock returns. *Journal of financial economics*, Vol. 105, No. 3, pp. 457–472, 2012.

[FF17] Eugene F Fama and Kenneth R French. International tests of a five-factor asset pricing model. *Journal of financial Economics*, Vol. 123, No. 3, pp. 441–463, 2017.

[FHJ51] Evelyn Fix and Joseph L Hodges Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, California Univ Berkeley, 1951.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, Vol. 1. Springer series in statistics New York, NY, USA:, 2001.

[For65] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, Vol. 21, pp. 768–769, 1965.

[FP14] Andrea Frazzini and Lasse Heje Pedersen. Betting against beta. *Journal of Financial Economics*, Vol. 111, No. 1, pp. 1–25, 2014.

[Fri01] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

[GGR06] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, Vol. 19, No. 3, pp. 797–827, 2006.

[GMW07] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*, Vol. 20. Siam, 2007.

[Gri02] John M Griffin. Are the fama and french factors global or country specific? *The Review of Financial Studies*, Vol. 15, No. 3, pp. 783–803, 2002.

[Ham94] James Douglas Hamilton. *Time series analysis*, Vol. 2. Princeton university press Princeton, 1994.

[HHCH09] Sheng-Hsun Hsu, JJ Po-An Hsieh, Ting-Chih Chih, and Kuei-Chu Hsu. A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, Vol. 36, No. 4, pp. 7947–7951, 2009.

[HL05] Peter R Hansen and Asger Lunde. A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics*, Vol. 20, No. 7, pp. 873–889, 2005.

[HLZ16] Campbell R Harvey, Yan Liu, and Heqing Zhu. ⋯ and the cross-section of expected returns. *The Review of Financial Studies*, Vol. 29, No. 1, pp. 5–68, 2016.

[Hua12] Chien-Feng Huang. A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, Vol. 12, No. 2, pp. 807–818, 2012.

[HW98] Wen-Jyi Hwang and Kuo-Wei Wen. Fast knn classification algorithm based on partial distance search. *Electronics letters*, Vol. 34, No. 21, pp. 2062–2063, 1998.

[Ita75] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 23, No. 1, pp. 67–72, 1975.

[JT93] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, Vol. 48, No. 1, pp. 65–91, 1993.

[JT01] Narasimhan Jegadeesh and Sheridan Titman. Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of finance*, Vol. 56, No. 2, pp. 699–720, 2001.

[Keo] Eammon Keogh. The ucr time series data mining archive. `http://www.`

cs.ucr.edu/\~{}eamonn/TSDMA/index.html.

[KL05] Eamonn Keogh and Jessica Lin. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and information systems*, Vol. 8, No. 2, pp. 154–177, 2005.

[KLW12] Hakan Kaya, Wai Lee, and Yi Wan. Risk budgeting with asset class and risk class approaches. *The Journal of Investing*, Vol. 21, No. 1, pp. 109–115, 2012.

[KMPV18] Ralph SJ Koijen, Tobias J Moskowitz, Lasse Heje Pedersen, and Evert B Vrugt. Carry. *Journal of Financial Economics*, Vol. 127, No. 2, pp. 197–225, 2018.

[KP00] Eamonn J Keogh and Michael J Pazzani. Scaling up dynamic time warping for datamining applications. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 285–289. ACM, 2000.

[KR87] Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

[KS19] Nakagawa Kei and Sashida Shingo. The ksnn package. 2019.

[Leh90] Bruce N Lehmann. Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, Vol. 105, No. 1, pp. 1–28, 1990.

[LFOS84] Breiman Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. *Wadsworth International Group*, 1984.

[Lin75] John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. In *Stochastic Optimization Models in Finance*, pp. 131–155. Elsevier, 1975.

[Man63] Benoit Mandelbrot. The variation of certain speculative prices. *The journal of business*, Vol. 36, No. 4, pp. 394–419, 1963.

[Men10] Lukas Menkhoff. The use of technical analysis by fund managers: International evidence. *Journal of Banking & Finance*, Vol. 34, No. 11, pp. 2573–2586, 2010.

[MM61] Merton H Miller and Franco Modigliani. Dividend policy, growth, and the

valuation of shares. *the Journal of Business*, Vol. 34, No. 4, pp. 411–433, 1961.

[Mos66] Jan Mossin. Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, pp. 768–783, 1966.

[MP16] R David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *The Journal of Finance*, Vol. 71, No. 1, pp. 5–32, 2016.

[MST+94] Donald Michie, David J Spiegelhalter, CC Taylor, et al. Machine learning. *Neural and Statistical Classification*, Vol. 13, , 1994.

[Mur99] John J Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[NIAI19] Kei Nakagawa, Tomoki Ito, Masaya Abe, and Kiyoshi Izumi. Deep recurrent factor model: Interpretable non-linear and time-varying multi-factor model. In *AAAI-19 Workshop on Network Interpretability for Deep Learning*. arXiv preprint arXiv:1901.11493, 2019.

[NR07] Vit Niennattrakul and Chotirat Ann Ratanamahatana. On clustering multimedia time series data using k-means and dynamic time warping. In *Multimedia and Ubiquitous Engineering, 2007. MUE'07. International Conference on*, pp. 733–738. IEEE, 2007.

[NUA18] Kei Nakagawa, Takumi Uchida, and Tomohisa Aoshima. Deep factor model. In *ECML PKDD 2018 Workshops*, pp. 37–50. Springer, 2018.

[PI07] Cheol-Ho Park and Scott H Irwin. What do we know about the profitability of technical analysis? *Journal of Economic Surveys*, Vol. 21, No. 4, pp. 786–826, 2007.

[PQS04] Jeng-Shyang Pan, Yu-Long Qiao, and Sheng-He Sun. A fast k nearest neighbors classification algorithm. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. 87, No. 4, pp. 961–963, 2004.

[RR04] Greg Ridgeway and Maintainer Greg Ridgeway. The gbm package. *R Foundation for Statistical Computing, Vienna, Austria*, Vol. 5, No. 3, 2004.

[Sha64] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, Vol. 19, No. 3, pp. 425–442, 1964.

[Sto74] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pp. 111–147, 1974.

[TAR10] Terry M Therneau, Beth Atkinson, and Maintainer Brian Ripley. The rpart package. 2010.

[TDO10] Lamartine Almeida Teixeira and Adriano Lorena Inacio De Oliveira. A method for automatic stock trading combining technical analysis and nearest neighbor classification. *Expert systems with applications*, Vol. 37, No. 10, pp. 6885–6890, 2010.

[TK14] Prodromos E Tsinaslanidis and Dimitris Kugiumtzis. A prediction scheme using perceptually important points and dynamic time warping. *Expert Systems with Applications*, Vol. 41, No. 15, pp. 6848–6860, 2014.

[VF10] Bruce Vanstone and Gavin Finnie. Enhancing stockmarket trading performance with anns. *Expert Systems with Applications*, Vol. 37, No. 9, pp. 6602–6610, 2010.

[VFH12] Bruce Vanstone, Gavin Finnie, and Tobias Hahn. Creating trading systems with fundamental variables and neural networks: The aby case study. *Mathematics and computers in simulation*, Vol. 86, pp. 78–91, 2012.

[WHW12] Baohua Wang, Hejiao Huang, and Xiaolong Wang. A novel text mining approach to financial time series forecasting. *Neurocomputing*, Vol. 83, pp. 136–145, 2012.

[YHKL09] Haiqin Yang, Kaizhu Huang, Irwin King, and Michael R Lyu. Localized support vector regression for time series prediction. *Neurocomputing*, Vol. 72, No. 10-12, pp. 2659–2669, 2009.

[YKJ05] Paul D Yoo, Maria H Kim, and Tony Jan. Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *Computational Intelligence for Modelling, Control and Au-*

tomation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on, Vol. 2, pp. 835–841. IEEE, 2005.

[YSYT03] Yuu Yamada, Einoshin Suzuki, Hideto Yokoi, and Katsuhiko Takabayashi. Decision-tree induction from time-series data based on a standard-example split test. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 840–847, 2003.

[伊藤 10] 伊藤博志. 論考テクニカル分析のパフォーマンス考察–その有効性と問題点の視座から–. 大阪経大論集, Vol. 60, No. 5, pp. 79–104, 2010.

[沖本 10] 沖本竜義. 経済・ファイナンスデータの計量時系列分析. 朝倉書店, 2010.

[刈屋 03] 刈屋武昭. 金融時系列分析入門. 経済時系列の統計』, 岩波書店, 2003.

[久保 07] 久保田敬一, 竹原均. Fama-french ファクターモデルの有効性の再検証. 現代ファイナンス, No. 22, pp. 3–23, 2007.

[袴田 02] 袴田守一. マルコフ切り換え確率的トレンドモデルを用いた topix のトレーディング戦略. 2002.

[今村 18a] 今村光良, 中川慧, 吉田健一. ベンチマークデータを用いた時系列勾配ブースティング木の実験評価. 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集. 社団法人 人工知能学会, 2018.

[今村 18b] 今村光良, 中川慧, 吉田健一. 価格変動パターンによる証券/為替/仮想通貨市場の分析. 電気学会論文誌 C (電子・情報・システム部門誌), Vol. 138, No. 8, pp. 992–998, 2018.

[佐藤 17] 佐藤賀一. テクニカル分析に基づくペアトレードの有効性と日本の株式市場の効率性. 行動経済学, Vol. 10, pp. 22–49, 2017.

[山田 03] 山田悠, 鈴木英之進, 横井英人, 高林克日己ほか. 動的時間伸縮法に基づく時系列データからの決定木学習. 情報処理学会研究報告知能と複雑系 (ICS), Vol. 2003, No. 30 (2002-ICS-132), pp. 141–146, 2003.

[神蔦 03] 神蔦敏弘ほか. データマイニング分野のクラスタリング手法 (1): クラスタリングを使ってみよう! 人工知能学会誌, Vol. 18, No. 1, pp. 59–65, 2003.

[総務] 総務省ＡＩネットワーク社会推進会議. 国際的な議論のためのＡＩ開発ガイドライン案. http://www.soumu.go.jp/main_content/000499625.pdf.

[筒井 09] 筒井義郎, 平山健二郎. 日本の株価: 投資家行動と国際連関. 東洋経済新報社, 2009.

[白浜 11] 白浜公章. 経済データに対する値と形状に基づく時系列類似尺度の比較. 国民経済雑誌, Vol. 204, No. 5, pp. 71–79, 2011.

[和泉 11] 和泉潔, 後藤卓, 松井藤五郎ほか. 経済テキスト情報を用いた長期的な市場動向推定. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315, 2011.

[藏本 13] 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志. 新聞記事のテキストマイニングによる長期市場動向の分析. 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296, 2013.

関連業績リスト

Referred papers

第 3 章 Kei Nakagawa,Mitsuyoshi Imamura and Kenichi Yoshida,"Stock Price Prediction using k^*-Nearest Neighbors and Indexing Dynamic Time Warping", International Workshop: Artificial Intelligence of and for Business (AI-Biz2017),2017

Kei Nakagawa,Mitsuyoshi Imamura and Kenichi Yoshida, "Stock Price Prediction with Fluctuation Patterns using Indexing Dynamic Time Warping and k^*-Nearest Neighbors",New Frontiers Artificial Intelligence: JSAI-isAI Workshops, JURISIN, SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI, Tsukuba, Tokyo, November 13-15, 2017, Revised Selected Papers, pp97-111, 2018

第 4 章 中川 慧, 今村 光良, 吉田健一,"価格変動パターンを用いた市場予測 k-Medoids Clustering with Indexing Dynamic Time Warping の株式市場への適用", 電気学会論文誌C（電子・情報・システム部門誌）, Vol. 138, No. 8, pp.986-991, 2018

Kei Nakagawa,Mitsuyoshi Imamura and Kenichi Yoshida,"Stock Price Prediction using k-Medoids Clustering with Indexing Dynamic Time Warping",Electronics and Communications in Japan, Vol.102, No.2, pp.3-8, 2019

第 5 章 Kei Nakagawa,Mitsuyoshi Imamura and Kenichi Yoshida,"Time-Series Gradient Boosting Tree for Stock Price Prediction",Submitted to International Journal of Data Mining, Modelling and Management

Unreferred papers

第 3 章　中川 慧, 今村 光良, 吉田健一,”株価変動パターンの類似性を用いた株価予測” 人工知能学会全国大会論文集 第 31 回全国大会, pp. 2D11-2D11, 2017

Kei Nakagawa,Shingo Sashida ,”ksNN Package”, 2019, CRAN

第 4 章 中川 慧, 今村 光良, 吉田健一,”価格変動パターンを用いた市場予測　IDTW Based k-medoids clustering の株式市場への適用”, 第 16 回情報科学技術フォーラム (FIT2017), 2017

第 5 章 中川 慧, 今村 光良, 吉田健一,”時系列勾配ブースティング木による分類学習 金融時系列予測への応用”, 人工知能学会全国大会論文集 第 32 回全国大会, pp. 2J203-2J203, 2018