

リターン時系列に基づく投資信託の動的クラスタリング

Dynamic clustering of mutual funds based on the return series

松井 藤五郎 *1
Tohgoroh Matsui米田 一樹 *2
Kazuki Yoneda森山 甲一 *2
Koichi Moriyama武藤 敦子 *2
Atsuko Mutoh犬塚 信博 *2
Nobuhiro Inuzuka*1中部大学
Chubu University*2名古屋工業大学
Nagoya Institute of Technology

This paper proposes a method to analyze mutual funds using dynamic clustering based on the return series. For the time series data divided into some terms, the proposed method (1) converts the original high-dimensional data to two dimensional data using t-SNE for each term, (2) applies dynamic clustering using x -means for each term, and (3) detects the cluster transitions using FBL-MONIC. This paper shows the experimental results for 29 Japanese mutual funds that track TOPIX, including four ETFs. The results indicate that there are three clusters at terms 1, 2, 3 and 4, and four clusters at term 5. We consider that the clusters are valid because one of the clusters consists of ETFs for all terms. FBL-MONIC could detect the transitions from a cluster at term 4 to a new cluster at term 5.

1. はじめに

投資信託（ファンド）は投資家から資金を預かり、ファンド・マネージャーと呼ばれる専門家が資産運用を行う金融商品である。ファンドの中には、東証株価指数（TOPIX）や日経平均株価のような株式指標（インデックス）と連動するように運用されているものがあり、これらはインデックス・ファンドと呼ばれる。

ファンドが運用の目標としている指標をベンチマークといい、同じ指標をベンチマークとしているインデックス・ファンドは同じような値動きをするはずである。しかし、実際には、同じ指標をベンチマークとしているインデックス・ファンドであっても、値動きが異なっていることがある。これは、ファンドごとに資産内容が異なるためだと考えられる。

これまで、ファンドの投資先類似度に基づくクラスタリング [1, 2] を行ってきたが、この方法は投資先がわかっていないと用いることができない。運用会社は自社が運用するファンドの投資先ならわかるが、一般的にはファンドの投資先を正確に知ることはできない。年に数回発行される運用報告書には投資先が記載されているが、ファンドごとに基準日が異なり、類似度を正確に求めることができない。

そこで本研究では、外部からでも観測可能な月次リターンの系列に基づいて、ファンドをクラスタリングする。月次リターンの系列を複数の期に分割し、期ごとにクラスタリングを行うことによって動的クラスタリングを行う。

我々は、動的クラスタリングにおいて、クラスター遷移を抽出する手法を提案している [3]。変化前のクラスターに属するデータと変化後のクラスターに属するデータの一致率を求め、変化前のクラスターを基準とした前向き分析によってクラスター遷移を検出する MONIC (MONItoring Cluster transitions) [4] という手法が提案されているが、異なる時刻に観測されたデータは同一でないと判断するために、特徴量が時間によって変化するデータに対して適用することができない。そこで我々は、ラベルの同一性に基づいてクラスター遷移を検出するよう MONIC を拡張した。これを本論文では FL-MONIC (Forward

Label-based MONIC) と呼ぶ。また、FL-MONIC に対して、変化後のクラスターを基準とした後ろ向き分析を追加した。これを本論文では FBL-MONIC (Forward-Backward Label-based MONIC) と呼ぶ。FBL-MONIC と FL-MONIC は、異なる時刻のデータであってもラベルが同じであれば一致すると判断して分析するため、特徴量が時間によって変化するデータに対して同じラベルを割り当てることによってクラスター遷移を検出することができる。

本論文では、インデックス・ファンドに対して月次リターンの系列に基づいて動的クラスタリングを行い、FBL-MONIC を用いてクラスター遷移を検出し、その結果を示す。

2. 分析手法

対象となるデータは、複数の期に分割された時系列データの集合 X_1, \dots, X_n である。ここで、 n は期の数を表す。期の長さを揃える必要はないが、同じにしておく結果を解釈しやすい。また、期が重なっていてもいいが、重複する期間が長いほどクラスタリング結果が類似しやすく、動的クラスタリングにおけるクラスターの変化が小さくなる。

まず、期ごとに、時系列データ X_t ($t = 1, \dots, n$) に対し、時系列に含まれる変数を時刻ごとに独立であると考え、t-SNE (t-distributed Stochastic Neighbor Embedding) [5] を用いて 2次元空間上のデータ集合 Y_t に変換する。

$$Y_t \leftarrow \text{t-SNE}(X_t)$$

t-SNE は、基準となる点 x から見た点 x' の近さを、 x を中心とした正規分布を考えて、 x が観測されたときに x' も観測される確率（条件付き確率）として表し、変換後の点の近さが自由度 1 の t 分布（正規分布よりも裾が重い分布）となるように変換することによって、近いデータをより近くに、遠いデータをより遠くに配置する。t-SNE を用いることによって、次元を圧縮することができ、高次元空間では最近傍であっても距離が大きくなってしまいう「次元の呪い」を解くことができる。また、2次元空間上の点に変換することによって、クラスタリング結果を可視化しやすくなる。ただし、t-SNE はデータ間の距離のみを考慮して次元圧縮を行うため、変換後の座標空間における軸の向きや大きさには意味がない。

Require: 期ごとの時系列データの集合 (X_1, \dots, X_n)

- 1: $n \leftarrow$ 期の数
- 2: **for** $t = 1$ **to** n **do**
- 3: 2次元空間上の点の集合 $Y_t \leftarrow$ t-SNE(X_t)
- 4: クラスターの集合 $\Gamma_t \leftarrow$ x-means(Y_t)
- 5: **end for**
- 6: **for** $t = 2$ **to** n **do**
- 7: クラスター遷移 $T_t \leftarrow$ FBL-MONIC(Γ_{t-1}, Γ_t)
- 8: **end for**
- 9: **return** $(\Gamma_1, \dots, \Gamma_n), (T_2, \dots, T_n)$

図1 分析手法のアルゴリズム

期ごとに、2次元空間上のデータ Y_t に対して x-means [6] を用いてクラスタリングを行い、クラスターの集合 Γ_t を得る。

$$\Gamma_t \leftarrow \text{x-means}(Y_t)$$

x-means は、非階層クラスタリングである k-means をクラスター数を自動的に決めるように拡張したものであり、クラスター数を2から1つ増やしながらか k-means を行い、BIC (Bayesian Information Criterion) を用いて最適なクラスター数を決める。x-means を用いることによって、k-means のクラスター数 k を与えなくてもよくなる。最適なクラスター数は期ごとに決められるため、本手法の動的クラスタリングにおいてはクラスター数は一定ではない。また、x-means や k-means では、クラスター割り当てをランダムに初期化するため、クラスター番号の変化には意味がない。

t-SNE と x-means の性質により、動的クラスタリングにおいて2次元空間上での位置やクラスター番号を用いて前後のクラスターを対応付けることはできない。

そこで、クラスターに属するデータのラベルに基づいたクラスター遷移検出法である FBL-MONIC [3] を用いる。クラスターの集合 $\Gamma_1, \dots, \Gamma_n$ に含まれる連続する二つのクラスター集合 Γ_{t-1} と Γ_t に対して、FBL-MONIC を用いてクラスター遷移 T_t を検出する。

$$T_t \leftarrow \text{FBL-MONIC}(\Gamma_{t-1}, \Gamma_t)$$

このアルゴリズムを図1に示す。このアルゴリズムは、複数の期に分割された時系列データの集合 X_1, \dots, X_n を入力すると、クラスターの集合 $\Gamma_1, \dots, \Gamma_n$ とクラスター遷移 T_2, \dots, T_n を返す。

3. インデックス・ファンドの分析

3.1 分析データ

Yahoo!ファイナンス内のモーニングスターのファンド検索において、「TOPIX」をキーワードとして検索し、検索されたファンドの中から東証株価指数 (TOPIX) 全体をベンチマークとしているインデックス・ファンドを取り出した*1。取り出されたファンドは29だった。その一覧を表1に示す。表中の*印はETF (上場投資信託) を表す。

これらのファンドについて、2013年11月から2018年10月までの5年分の月次リターンの時系列データを取得した。ファ

*1 例えば、野村アセットマネジメントが運用する「(NEXT FUNDS) 素材・化学上場投信」は、「TOPIX」をキーワードとした検索結果に含まれるが、対象インデックスが TOPIX ではなく TOPIX-17 素材・化学であるため除外した。

表1 分析対象の TOPIX 連動型ファンド。*印はETFを表す。

ファンド名	ラベル
トビックス・インデックス・オープン	Nomura-TIO
TOPIX 連動型上場投資信託*	Nomura-ETF
野村 インデックス F・TOPIX	Nomura-Fi
インデックスファンド TSP	Nikko-TSP
上場インデックスファンド TOPIX*	Nikko-ETF
インデックスファンド TOPIX(日本株式)	Nikko-DC
三菱 UFJ TOPIX・ファンド	MUFJ-TF
MAXIS トビックス上場投信*	MAXIS-ETF
トビックスオープン	MUKAM-TO
三菱 UFJ トビックスインデックスオープン	MUFJ-TIO
eMAXIS TOPIX インデックス	eMAXIS
三菱 UFJ トビックスオープン	MUFJ-TO
トビックス・インデックスファンド	Daiwa
ダイワ 上場投信-トビックス*	Daiwa-ETF
明治安田 TOPIX オープン	MYAM
JA TOPIX オープン	JA
ニッセイ TOPIX オープン	Nissei
MHAM トビックスファンド	MHAM-TF
DIAM 国内株式バッシブ・ファンド	DIAM
MHAM TOPIX オープン	MHAM-TO
ニュートビックスインデックス	One
東京海上 日本株 TOPIX ファンド	TMAM
りそな・TOPIX オープン	Risona
しんきん トビックスオープン	Shinkin
日本株式インデックス e	SMT-Jle
日本株式インデックスファンド	SMT-JIF
SMT TOPIX インデックス・オープン	SMT-TIO
日本株式インデックス・オープン	SMT-JIO
国内株式指数ファンド (TOPIX)	SMA

ンドの月次リターンから TOPIX の月次リターンの値を引いて TOPIX 超過リターンを求め、1年分ずつに分割してそれぞれ1つの期とした。すなわち、1つのデータは長さ12の時系列として表されている。

例として、5期 (2017年11月から2018年10月まで) の TOPIX 超過リターンの推移を図2に示す。折れ線の色は、後述するクラスタリング結果のクラスター番号を表している。1年分の折れ線グラフでは線がほとんど重なってしまうため、最初の4ヶ月分を拡大したものを図3に示す。このように、多くのファンドがかなりよく似た値動きをしている。緑色のファンドは全てETFであり、互いに非常によく似た値動きをしているが、他のファンドと比較して TOPIX 超過リターンが高い傾向にあることがわかる。

3.2 分析結果

t-SNE によって変換された2次元空間と x-means によるクラスタリング結果を図4-8に示す。データの記号と色はクラスター番号を表している。また、クラスター番号を表2に示す。

表2は、FBL-MONIC によって検出されたクラスター遷移のうち、前向き分析において生存していると判定されたクラスターを同じ色で表している*2。前向き分析だけの FL-MONIC では、 $C_{4,2}$ (4期の2番クラスター) に属するメンバーの4/7が $C_{5,4}$ にも属していることから、「 $C_{4,2}$ は $C_{5,4}$ として生存している」と判定されたが、 $C_{4,2}$ と $C_{5,2}$ の関係は検出されなかった。後ろ向き分析も行う FBL-MONIC では、 $C_{5,2}$ に属するメンバーの2/3が $C_{4,2}$ にも属していることから、「 $C_{5,2}$ は $C_{4,2}$ の影響が強い」と判定された。

*2 一致率の閾値を $\tau = 0.5$ とした。

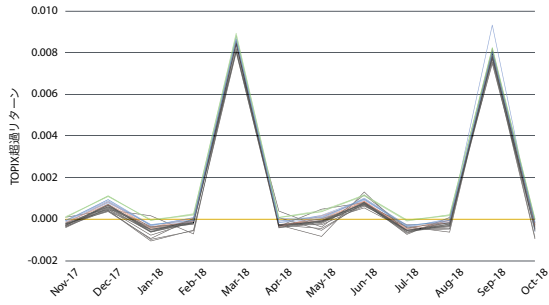


図2 29 ファンドの5期（2017年11月から2018年10月まで）のTOPIX 超過リターン。色はクラスター番号を表す。

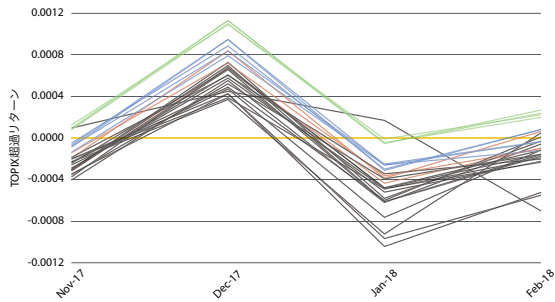


図3 29 ファンドの5期の最初の4ヶ月（2017年11月から2018年2月まで）のTOPIX 超過リターンの拡大図。色はクラスター番号を表す。

表2 動的クラスタリングの結果（クラスター番号）と信託報酬。セルの色は対応するクラスターを表す。同じクラスターに属するファンドが並ぶようにソートされている。

ラベル	1期	2期	3期	4期	5期	信託報酬
Nomura-TIO	1	3	1	1	1	0.67%
Nikko-TSP	1	3	1	1	1	0.56%
Nikko-DC	1	3	1	1	1	0.67%
MUFJ-TF	1	3	1	1	1	0.68%
MUKAM-TO	1	3	1	1	1	0.67%
MUFJ-TO	1	3	1	1	1	0.65%
Daiwa	1	3	1	1	1	0.67%
MYAM	1	3	1	1	1	0.65%
JA	1	3	1	1	1	0.59%
MHAM-TO	1	3	1	1	1	0.59%
One	1	3	1	1	1	0.65%
TMAM	1	3	1	1	1	0.65%
Risona	1	3	1	1	1	0.65%
Shinkin	1	3	1	1	1	0.86%
SMT-JIF	1	3	1	1	1	0.65%
MHAM-TF	1	3	2	1	1	0.65%
DIAM	1	3	1	1	2	0.49%
SMT-JIO	3	2	1	2	1	0.65%
MUFJ-TIO	3	2	1	2	2	0.70%
Nissei	3	2	2	2	2	0.54%
Nomura-Fi	3	2	2	2	4	0.43%
eMAXIS	3	2	2	2	4	0.43%
SMT-JIe	3	2	2	2	4	0.40%
SMT-TIO	3	2	2	2	4	0.40%
SMA	3	2	2	3	4	0.43%
Nomura-ETF	2	1	3	3	3	0.12%
Nikko-ETF	2	1	3	3	3	0.10%
MAXIS-ETF	2	1	3	3	3	0.08%
Daiwa-ETF	2	1	3	3	3	0.12%

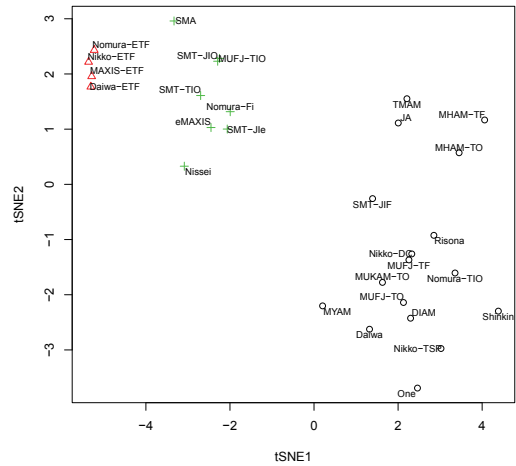


図4 1期（2013年11月から2014年10月まで）の超過リターン時系列に基づくTOPIX ファンドのクラスタリング結果。

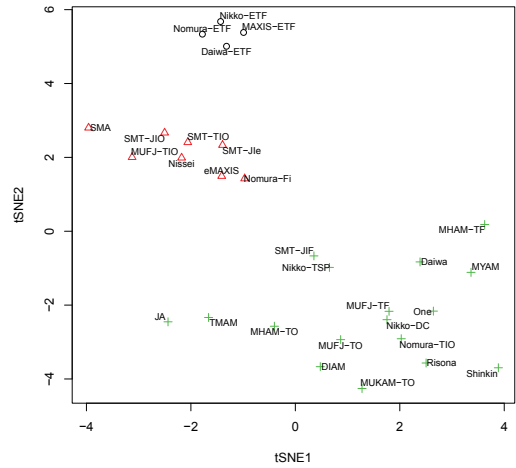


図5 2期（2014年11月から2015年10月まで）の超過リターン時系列に基づくTOPIX ファンドのクラスタリング結果。

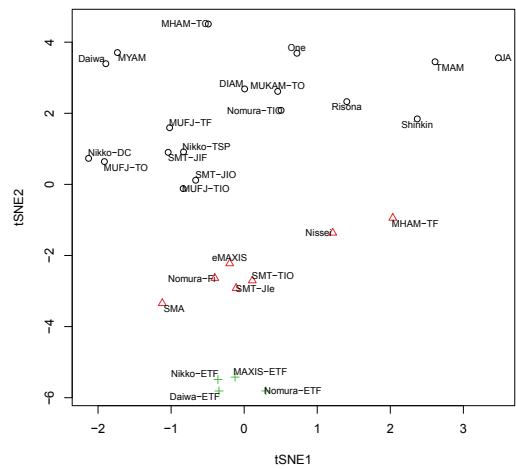


図6 3期（2015年11月から2016年10月まで）の超過リターン時系列に基づくTOPIX ファンドのクラスタリング結果。

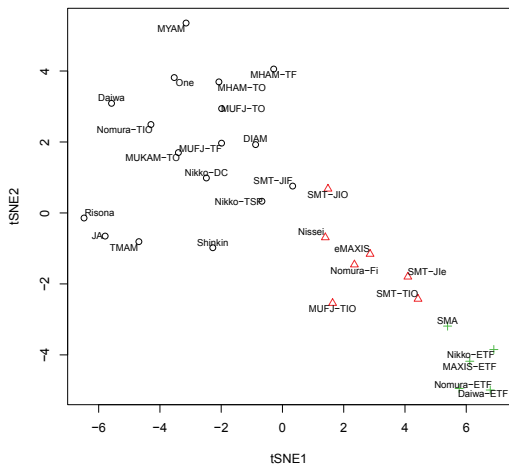


図7 4期(2016年11月から2017年10月まで)の超過リターン時系列に基づく TOPIX ファンドのクラスタリング結果。

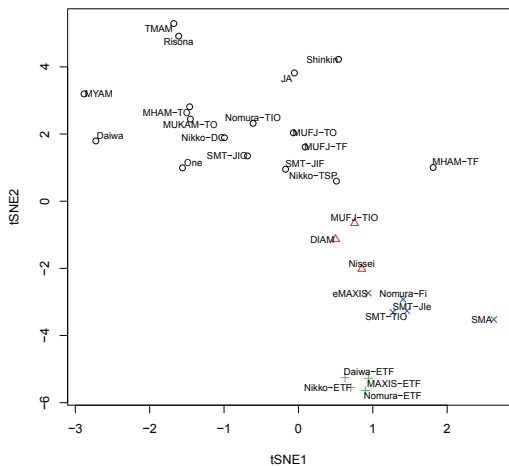


図8 5期(2017年11月から2018年10月まで)の超過リターン時系列に基づく TOPIX ファンドのクラスタリング結果。

4. 考察

今回の分析対象には、株式市場で売買されている ETF が 4 つ含まれている。ETF は株式市場で売買されており、証券会社や銀行などに支払う販売コストがかからないことから、信託報酬を小さくすることができる。

t-SNE を用いてリターン時系列を 2 次元空間に圧縮した図を見ると、全ての期において ETF は極めて近い位置に配置されている。このことから、t-SNE による次元圧縮は適切に行われていると考えられる。また、クラスタリングの結果を見ると、全ての期において ETF が同じクラスターに含まれており、かつ、そのクラスターはほぼ ETF だけで構成されていた。このことから、t-SNE によって圧縮された 2 次元空間上での x-means によるクラスタリングも適切に行われていると考えられる。

t-SNE と x-means の性質により、2 次元空間上での位置とクラスター番号だけからでは動的クラスタリングにおけるクラスターの遷移や対応関係を求められない。提案手法では、クラス

ターに属するメンバーのラベルに基づいてクラスター遷移を検出することによって、クラスターの遷移と対応関係を求めることができる。

前向き分析だけの FL-MONIC は、5 期の 2 番クラスター $C_{5,2}$ を 4 期のクラスターとは関係のない新しいクラスターと判定したが、後ろ向き分析も行う FBL-MONIC は、 $C_{5,2}$ を $C_{4,2}$ の影響が強いクラスターと判定した。実際に、 $C_{5,2}$ を構成する DIAM, MUFJ-TIO, Nissei のうち、MUFJ-TIO と Nissei は $C_{4,2}$ にも含まれるため、 $C_{5,2}$ は $C_{4,2}$ から分割された小さなクラスターに新たに DIAM が加わったものとして捉えることができる。このことから、FBL-MONIC による後ろ向き分析によって前向き分析だけでは捉えることができないクラスター遷移を検出できるようになったことがわかる。

また、全ての期において ETF クラスターに最も近いクラスターに所属していたファンドは、全て信託報酬が 0.43% 以下だった。ファンドの基準価格からは信託報酬等の間接コストが控除されており、運用が同じなら間接コストが大きい方がリターンが小さくなる。本論文では TOPIX との連動を目指すファンドだけを対象にしたためリターン時系列の違いは小さいが、信託報酬等の間接コストがリターンに影響した結果が反映されたクラスターを獲得することができたと考えられる。

5. おわりに

本論文では、複数の期に分割されたファンドのリターン時系列を対象として、(1) 期ごとに t-SNE を用いて 2 次元空間に圧縮し、(2) 期ごとに x-means を用いてクラスタリングを行い、(3) FBL-MONIC を用いてクラスター遷移を検出することによって、動的クラスタリングを行ってそのクラスター遷移を検出する方法を提案した。

提案手法を用いて TOPIX 連動型インデックス・ファンドを分析したところ、1 期から 4 期は 3 つ、5 期は 4 つのクラスターが得られた。全ての期において、クラスターの 1 つに全ての ETF が含まれており、かつ、このクラスターには ETF 以外のファンドがほとんど含まれていなかったことから、提案手法によって妥当な動的クラスタリングを行うことができたと考えられる。また、5 期に新しく増えたクラスターは、前向き分析では 4 期のクラスターからの遷移として検出されなかったが、後ろ向きでは 4 期のクラスターの 1 つの影響が強いと判定され、クラスター遷移として検出することができた。

参考文献

- [1] T. Sakakibara, T. Matsui, A. Mutoh, N. Inuzuka: Clustering mutual funds based on investment similarity, *Procedia Computer Science*, 60:881–890, 2015
- [2] 米田一樹, 松井藤五郎, 武藤敦子, 犬塚信博, 森山甲一: 投資割合を考慮した投資先類似度に基づく投資信託のクラスタリング, 第 31 回人工知能学会全国大会, 2D1-4, 2017
- [3] 米田一樹, 松井藤五郎, 森山甲一, 武藤敦子, 犬塚信博: 動的クラスタリングにおけるクラスタの変化分析, 情報処理学会数理モデル化と問題解決研究会, 2019-MPS-122(6):1–6, 2019
- [4] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, R. Schult: MONIC—Modeling and monitoring cluster transitions, *KDD 2006*, 706–711, 2006
- [5] L. van der Maaten, G. Hinton: Visualizing data using t-SNE, *J Mach Learn Res*, 9:2579–2605, 2008
- [6] 石岡恒憲: クラスター数を自動決定する k-means アルゴリズムの拡張について, *応用統計学*, 29(3):141–149, 2000