3L3-OS-26a-5

スペクトラルクラスタリングを用いた投資信託分析

Analysis of Investment Trusts Using Spectral Clustering

榊原 拓将 *1 Takumasa Sakakibara　松井 藤五郎 *2 Tohgoroh Matsui　武藤 敦子 *1 Atsuko Mutoh　犬塚 信博 *1 Nobuhiro Inuduka

*1名古屋工業大学大学院工学研究科情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

*2中部大学
Chubu University

Investing in single or similar financial instruments is not recommended from the viewpoint of risk. In order to grasp similar investment trusts, make the clustering focus on stocks companies that investment trusts are investing in. And to analyze the structure of investment trusts, we propose a clustering method that gives adequate clustering for investment trusts highly original.

1. はじめに

2014 年から少額投資非課税制度「ＮＩＳＡ（ニーサ）」が始まり投資は我々にとってより身近なものになっていくと考えられる。しかし投資には急激な価格変動や、破綻といったリスクがある。そのため単一の金融商品への投資ではなく、幅広い複数の金融商品に投資することが重要であるが、多くの金融商品一つ一つを把握し管理・運用することは容易なことではない。そこで資産の運用をプロに委託することができる投資信託という金融商品が存在する。

投資信託（ファンド）とは、販売会社を通じて複数の投資家から集めた資金を、資産運用の専門家が株や債券などの金融資産、あるいは不動産などに投資するよう指示し、運用成果をリターンとして投資家に分配する金融商品のことである。各ファンドは運用についての方針を持っており、それに従って複数の金融商品へ投資を行う。ファンドは自らが公示した投資方針に従って投資を行うため、投資家は運用方針の決定やリスク管理のためにファンドの性格を把握しておく必要がある。

そこで武田ら [1, 2] はファンドが投資している銘柄について着目しクラスタリングとネットワーク構造を用いてファンドの分析を行った。武田は投資銘柄が類似しているファンドは価格の変動も類似しているという考えから、似通ったファンドを見つけるためクラスタリング手法を用いたが、ファンドの中には他の多くのファンドとは違った銘柄に投資している独自性の高いファンドが存在しそれに対してうまくクラスタリングを行うことができなかった。

そこで本稿ではファンドの構造を分析し、独自性の高いファンドもクラスタリングできる手法について提案する。

本稿では、まず第 2 章で実験に用いた投資信託のデータとそのネットワークから見られる特徴について述べる。次に、第 3 章で実験に用いるクラスタリング手法の説明と投資信託のクラスタリングで生じる問題点を密度の観点から述べる。第 4 章で問題を解決するための提案手法について述べ、第 5 章では各手法の評価について述べる。最後に第 6 章でまとめをおこなう。

連絡先: 榊原拓将、名古屋工業大学犬塚研究室、名古屋市昭和区御器所町、052-735-5050、sakakibara@nous.nitech.ac.jp

ファンドの集合を V_f とした時
$v_i, v_j \in V_f$
ネットワーク : $G = (V, E)$
ノード集合 : $V = V_f$
エッジの重み : $w(v_i, v_j) = v_i, v_j$の間で共通する投資先銘柄数
エッジ集合 : $E = \{(v_i, v_j) | w(v_i, v_j) > 0\} \subseteq V_f \times V_f$

図 1: 図 2 のネットワーク定義

2. 投資信託の構造分析

2.1 投資信託データの取得

ファンドについての基本的な情報を得るため、Yahoo!ファイナンスの提供する MORNINGSTAR の投資信託情報の中から、投資割合上位 10 銘柄の公開されている**国内株式型投資信託**を抽出した。さらに、Python により正規表現を用いて、これらの投資信託の詳細ページの HTML ファイルから実験に用いるデータを抜き出した。このデータには、各ファンドの投資割合上位 10 銘柄の銘柄名、銘柄業種、投資割合、ファンドの純資産や分配金、取引手数料などの情報が含まれている。抽出したファンド総数は 496 であり投資先銘柄総数は 688 個であった。また、これらの 688 銘柄の投資先業種総数は、東京証券取引所の定める 33 業種であった。

2.2 投資信託の構造

ファンドの構造を分析するために各ファンドの投資割合が高い上位 10 銘柄において、「共通する銘柄数を重みとしてファンド間にエッジを張ったネットワーク」を用いる。ファンドを赤の丸で表現し重みをエッジの本数としたネットワークが図 2 である。また、そのネットワークの定義は図 1 である。

このネットワーク図から、中心に集まっている多くのファンドが似通った銘柄に投資していることが分かる。また中心から離れたところに他のファンドとは違う銘柄に投資をしているファンドが確認できる。本稿ではこれを**独自性の高いファンド**と呼び、そうではない多くのファンドと投資先が似通っているファンドについては**独自性の低いファンド**と呼ぶ。

また、銘柄ごとに投資されているファンドの数を調べるとごく少数の銘柄が 100 や 200 を超えるファンドから投資されて

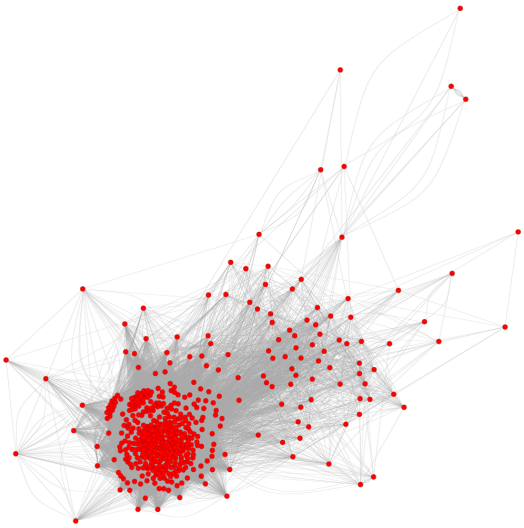

図 2: 共通投資銘柄数を重みとしたファンドのネットワーク

いる一方でほとんどの銘柄については 5 や 10 程度のファンドからしか投資をされていないことがわかった。

3. クラスタリングと問題点

複数のデータの集合をその類似度によっていくつかのグループに分けることをクラスタリングという。本研究ではファンドに対してクラスタリングを行う。そのため同じクラスタ (グループ) に含まれるファンドの投資先は類似したものになると考えられる。そこでファンドの投資先上位 10 銘柄で、投資している銘柄には 1 を投資していない銘柄には 0 を代入した長さ 688 のベクトルに対してクラスタリングを行う。

3.1 用いるクラスタリング手法と HITS アルゴリズム

本研究で用いるクラスタリング手法について述べる。

k-means 法

k-means 法は重心をクラスタの代表点とし、

$$\sum_{i=1}^{k}\sum_{x\in C_i}(dist(x,c_i))^2$$

の評価関数を最小化する。ここで k は分割数、x は対象、C はクラスタ、c はクラスタの重心であり、$dist(x,y)$ は x と y のユークリッド距離である。最適解の探索は対象の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。この手法は山登り法で、局所最適解しか求められないため、ランダムに初期値を変更して、評価関数を最小にする結果を選択する。

スペクトラルクラスタリング

スペクトラルクラスタリング [3, 4, 5] ではグラフ分割の問題としてクラスタリングを行う。クラスタリングを行う前に、データをノードとしデータ間にその類似度に応じて重みをつけたグラフを作成する。このグラフの同じクラスタで構成されたサブグラフ内は密であり、異なるクラスタで構成されたサブグラフ間は疎であるようなクラスタリングを行う。そこでサブグラフに分割する時にかかるコスト (分割するときに取り除くエッジの重みの総和) を最小とする最適解が固有値問題の解に対応することを利用したのがスペクトラルクラスタリングである。グラフの類似度行列から作られるラプラシアン行列の固有値を解くことでグラフの特徴を保持したまま低次元でのクラスタリングを行うことができ、k-means 法の欠点である初期値に大きく依存し局所最適解に陥るといったことが少ない。

階層的クラスタリング (ウォード法)

階層的クラスタリングは 1 個の対象だけを含む N 個のクラスタがある初期状態から、クラスタ間の距離に基づき、クラスタが 1 つになるまで最も距離の近い 2 つのクラスタを逐次的に併合していくクラスタリング手法であり、ウォード法 [6] は、各対象から、その対象を含むクラスタの重心までの距離の二乗の総和を最小化するようにクラスタを併合していく階層的クラスタリングの手法の一つである。距離については対象間のユークリッド距離を求めて適用する。

HITS アルゴリズム

HITS アルゴリズム [7] とは、被参照度（authority スコア）と、評価の高い Web ページへの参照度（hub スコア）から、重要性の高い Web ページを抽出する際に用いられるアルゴリズムである。被リンクは評価を受けていることを示し、発リンクはほかを評価していることを示すという考えに基づき、多くのリンクを受ける Web ページが重要な Web ページであるのと同様に、あるテーマに対して重要な Web ページに数多くリンクを張る Web ページもまたそのテーマの重要な情報源であると考える。このスコアリング手法を導入することで、単にキーワードに対しての適合度が高いだけのページではなく、多くの人間から実際に高く評価されている Web ページを高くスコアリングすることが可能となる。

3.2 クラスタリングの問題点

ファンドの属性ベクトルに対し k-means 法、ウォード法、スペクトラルクラスタリングを行い、その密度を求めるとある特徴が見て取れた。ネットワークにおける密度とはノード間にどれだけのエッジが張られているのかを測る指標である。

ノード数が n, エッジの重みの総和が m である時

$$\text{ネットワークの密度} = \frac{m}{n(n-1)/2}$$

と定義される。各手法でクラスタリングを行い同じクラスタ番号のファンドで構成された図 2 のネットワークの平均密度、最小密度、そして最小密度であったクラスタに含まれるファンド数をまとめたものが表 1 である。 表を見ると、どのクラスタリング手法にも平均より密度がはるかに小さいクラスタが存在していることがわかる。またそのクラスタに含まれるファンドの数も少なくなかった。

図 3 は k-means 法でクラスタリングし、クラスタ番号ごとにノード (ファンド) を色分けしたネットワークである。これを見ると中心に集まっている (投資銘柄が似通っている) ファンドが 4 つにクラスタリングされ、その周りの独自性の高いファンドが 1 つのクラスタで赤色に色分けされているのが分かる。この赤いファンド群が最も密度が低いクラスタの原因であり、クラスタ数を増やしてクラスタリングしていっても中心だ

表 1: クラスタ内の密度

クラスタ数	クラスタ手法	平均密度	最小密度 (ファンド数)
5	k-means 法	4.538	0.297(188)
10	k-means 法	4.4721	0.200(118)
30	k-means 法	4.117	0.153(65)
5	ウォード法	6.187	0.415(217)
10	ウォード法	5.941	0.250(117)
30	ウォード法	6.419	0.204(101)
5	スペクトラル	3.740	0.251 (165)
10	スペクトラル	4.461	0.210(139)
30	スペクトラル	5.006	0.162(84)

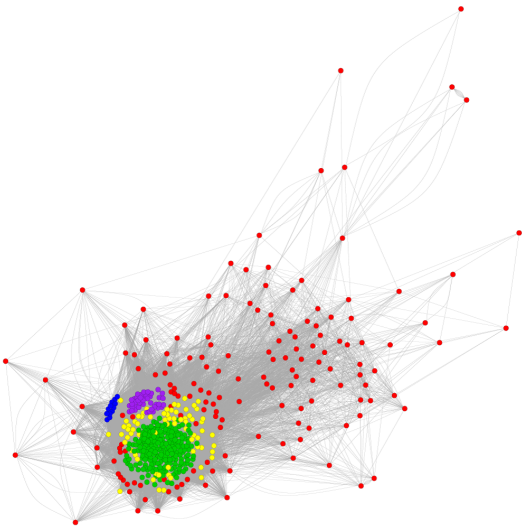

図 3: k-means 法クラスタ数 5

けが細分化され独自性の高いファンド群をうまく分けることができない。武田はこの問題を解決するために HITS アルゴリズムを用いて独自性の高いファンドを取り除いたあとでクラスタリングを提案したが、本稿では独自性の高いファンドもファンドを分析するのに重要であると考え独自性の高いファンドもクラスタリング可能な手法を提案する。

4. 提案手法

ファンドの中には他の多くのファンドとは違う銘柄に投資しているような独自性の高いファンドが存在している。これが原因でファンドに対してクラスタリングを行っても独自性の高いファンドは類似していないのにまとめられ、独自性の低いファンドだけが細かく分割されていった。

本章ではスペクトラルクラスタリングを利用して独自性の高いファンドもクラスタリングすることができる手法を提案する。

4.1 スペクトラルクラスタリングの利用

スペクトラルクラスタリングではグラフ分割の問題としてクラスタリングを行う手法であり、グラフをカットする時の最小コストの最適解が固有値問題の解に対応しているのを利用している。

しかし今回のファンドのデータでは属性ベクトルを用いているため、類似度行列に変換する必要がある。そのためスペクトラルクラスタリングは何を類似度としてデータ間に重みをつけるかによって結果が大きく影響される。

そこでファンドの類似度はその共通投資先銘柄数に比例して高いものと考え、類似度行列に変換する工程を経ず、直接類似度行列 W として図 2 の「ファンド間で共通して投資している銘柄数を重みとしたネットワーク」を与えスペクトラルクラスタリングを行う手法を提案する。

図 1 から類似度行列 W は

$$W_{ij} = w(v_i, v_j)$$

となる。ここで v はファンド、$w(v_i, v_j)$ は v_i と v_j で共通する銘柄の数である。提案手法のアルゴリズムを表 4 に示す。

入力:類似度行列 W, 分割数 k
出力:各対象が分割されたクラスタ $C_1, \cdots, C_k$

1. W から $D_{ii} = \sum_j W_{ij}$ となる対角行列 D を作る
2. W, D から $L = I - D^{-1/2} W D^{-1/2}$ となるラプラシアン行列 L を作る
3. L の固有値, 固有ベクトルを計算し固有値の値が小さい固有ベクトルを k 個列に並べた行列 U を作る
4. U に対して k-means 法を行いクラスタ C を割り当てる

図 4: 提案手法のアルゴリズム

表 2: 提案手法におけるクラスタ内の密度

クラスタ数	クラスタ手法	平均密度	最小密度 (ファンド数)
5	提案手法	2.160	0.255(52)
10	提案手法	2.869	0.444(9)
30	提案手法	3.112	0.963(11)

4.2 クラスタリング結果

提案した手法で 5 つにクラスタリングを行いクラスタ番号ごとに色分けをしたネットワークが図 5 である。k-means 法と比べると中からではなく外側から分割されているのがわかる。

表 2 は表 2 で示した密度の計算を提案手法に行った結果である。k-means 法、ウォード法、スペクトラルクラスタリングと比べると平均密度は小さいが最小の密度は大きいことが分かる。また密度が最小であるクラスタに含まれるファンドの数も少なかった。

5. 評価

本章では幅広い銘柄への投資を考慮した評価方法を示し、第 3 章で紹介した各クラスタリング手法と提案手法について評価する。

5.1 ポートフォリオにおける評価

ポートフォリオとは投資した金融商品の組み合わせである。ここではファンドの組み合わせとしてポートフォリオを構成する。各クラスタリング手法を用いてファンドを k 個のクラスタに分割し、各クラスタからランダムに一つずつ取り出した k

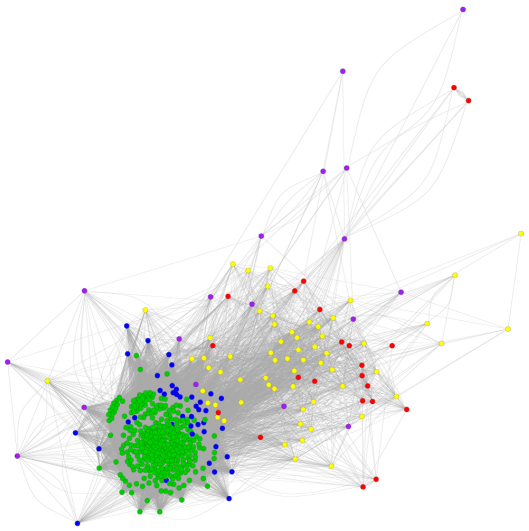

図 5: 提案手法クラスタ数 5

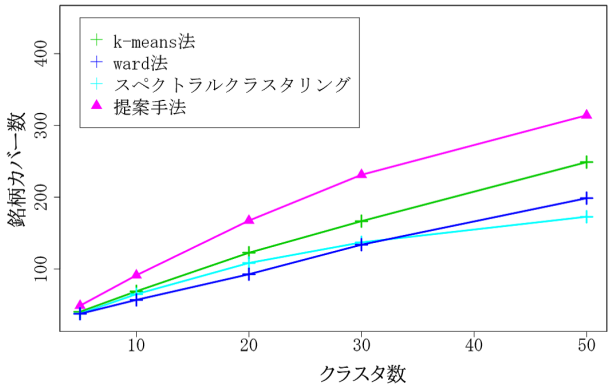

図 6: ポートフォリオの銘柄数

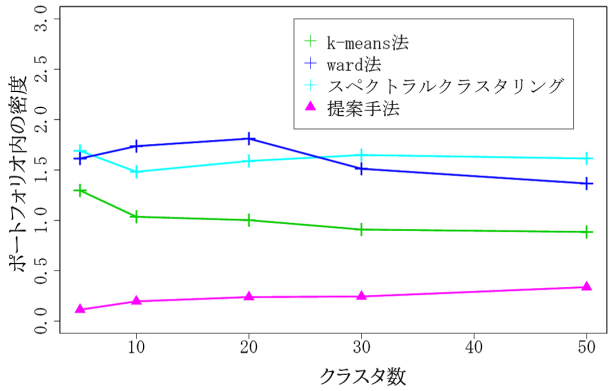

図 7: ポートフォリオの密度

個のファンドでポートフォリオを構成し、ポートフォリオが投資している銘柄数と密度について評価する。

投資銘柄数による評価

各クラスタ内のファンドは類似したものがまとめられているためポートフォリオのファンド群は幅広い銘柄に対して投資ができていると考えられる。

図 6 はポートフォリオが重複しない銘柄にどれだけ投資しているかを各クラスタリング手法でクラスタ数 (5,10,20,30,50) ごとに比較したものである。

銘柄数が多いほど、幅広い銘柄への投資ができていることを表す。

密度による評価

各クラスタ間のファンド群は非類似度が高いものと考えられる。そこでポートフォリオの密度を求める。

図 7 はポートフォリオ内のファンドの密度を各クラスタリング手法でクラスタ数 (5,10,20,30,50) ごとに比較したものである。

6. おわりに

6.1 手法の一般性についての検討

前章の実験では投資割合上位 10 銘柄の情報を用いてクラスタリングを行った。ここで、投資割合上位 10 銘柄の情報を用いずに上位 5 銘柄の情報のみを使ってクラスタリングを行うことでポートフォリオを構成した場合にも 10 銘柄の情報を利用した場合に近い結果が得られれば、投資割合上位 11 銘柄以降の銘柄情報を用いた場合についても同様の手法が有効であると考えられる。そこで、クラスタリングについては投資割合上位 5 銘柄の情報のみを使って行い、評価については投資割合上位 10 銘柄の情報を使って同様の評価を行うと銘柄数、密度ともに上位 10 銘柄を使った場合と同様の結果が得られた。

6.2 まとめ

本稿では幅広い投資を目的としてファンドのネットワーク構造、クラスタリングを試みた。ファンド間の投資銘柄に着目して重みをつけたネットワークから、他のファンドと似通った銘柄に投資している独自性の低いファンドと、他とは違う銘柄に投資している独自性の高いファンドが存在することが確認できる。そのためファンドの投資割合が高い上位 10 銘柄を属性とした属性ベクトルに対し k-means 法、ウォード法、スペクトラルクラスタリングでクラスタリングを行った結果、独自性の低いファンドが分割され独自性の高いファンドは一つのクラスタに無理やりまとめられその密度は低いものとなった。

武田は独自性の高いファンドをあらかじめ選別し取り除いておく手法を提案したが、本稿では独自性の高いファンドもファンドの特徴を分析する上で重要であると考えスペクトラルクラスタリングに属性ベクトルを与えるのではなく、初めから「ファンド間で共通する投資銘柄数を重みとするネットワーク」を類似度行列として与える手法を提案した。

提案手法の各クラスタの密度を調べると少なくとも独自性の高いファンドを一まとめにしてしまうということはなかった。各手法でクラスタリングし各クラスタから取り出したファンドで構成されたポートフォリオの投資銘柄数と密度で評価した結果、提案手法はよい結果を残した。また上位 5 銘柄だけの情報でクラスタリングした場合も上位 10 銘柄を用いた結果に近いものとなったため投資割合 11 位以降の銘柄情報を利用した場合においても一般性のある手法として適用できると考えられる。

今後の課題として、分割数 k の検討、ファンドのリターンを考慮した分析などがあげられる。

参考文献

[1] 武田祐太, 松井藤五郎, 犬塚信博, 「投資信託と銘柄のネットワーク構造に基づいた投資信託の分類」, 第 75 回情報処理学会全国大会, 5M-9 (2013)

[2] 松井藤五郎, 武田祐太, 犬塚信博, 「ネットワーク構造分析を用いた投資信託の分析」, 第 10 回人工知能学会ファイナンスにおける人工知能応用研究会 (SIG-FIN), SIG-FIN-010-09, pp. 49-54 (2013)

[3] U. von Luxburg, "A tutorial on spectral clustering", Stat. Comp. Vol. 17, Issue 4, 395-416 (2007)

[4] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis, "kernlab -An S4 Package for Kernel Methods in R"

[5] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 8, AUGUST 2000

[6] 神嶌 敏弘, データマイニング分野のクラスタリング手法 (1), 人工知能学会誌, pp.59-65, (2003)

[7] Jon Kleinberg, "Authoritative sources in a hyperlinked environment", ACM-SIAM Symposium on Discrete Algorithms,pp.668-677 (1998)