Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 881 – 890

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Clustering Mutual Funds Based on Investment Similarity

Takumasa Sakakibara[a], Tohgoroh Matsui[b], Atsuko Mutoh[a], Nobuhiro Inuzuka[a,*]

[a]*Nagoya Institute of Technology*
[b]*Chubu University*

Abstract

It is risky to invest to single or similar mutual funds because the variance of the return becomes large. Mutual funds are categorized based on the investment strategy by a company that rated funds based on performance, but the fund categories are different from its actual operations. While some previous studies have proposed methods to cluster mutual funds based on the historical performances, we cannot apply these methods to new mutual funds. In this paper, we clusters mutual funds based on the investment similarity instead of the historical performances. The contributions of this paper are: 1. To propose two new methods for classifying mutual funds based on the investment similarity, 2. To evaluate the proposed methods based on actual 551 Japanese mutual funds.

Peer-review under responsibility of KES International

Keywords: clustering; mutual fund; *k*-means; spectral

1. Introduction

A mutual fund is a financial instrument that distributes revenue gained by entrusting the management of funds collected from individuals to a specific expert, according to their contribution amounts. Each mutual fund has an operation policy that determines how it invests in accordance with multiple financial products. To invest in their own mutual fund in accordance with its public investment policy, investors need to know the type of fund in order to determine the risk management of the investment policy. We can refer to mutual fund categories published by the financial instruments industry to discover the nature of the funds but it is hard to confirm that the category has been sufficiently classified in consideration of the type of mutual funds. There are problems of false declaration and cases where the category does not change following a change in the mutual fund investment policy (for example, when a mutual fund categorized as an active mutual fund has management aspects such as an index mutual fund[1,2]). Although there is a method to determine the funds type using return information, this method cannot be used to new mutual funds that have no performance history. Therefore, we propose a method to perform clustering focused on the companies where the mutual fund is invested. Since funds that invest in the same company tend to have similar

* Takumasa Sakakibara. Tel.: +81-52-735-5469 ; fax: +81-52-735-5469.
E-mail address: cke17557@stn.nitech.ac.jp

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.251

operation policies, it is considered that they can provide a more appropriate fund classification by clustering rather than using fund categories.

There are several related works. Baghai-Wadji et al.[3] have used Self-Organizing Maps (SOM) to find homogeneous groups of hedge funds based on similar characteristics. They identified nine hedge fund classes based on a ten-year sample of 2,442 dead and active hedge funds. Moreno et al.[4] have proposed a method to improve the classification of Spanish mutual funds using SOM, k-nearest neighbors and the k-means algorithm. They reported that over 40% of mutual funds were misclassified in the official Spanish mutual funds classification. Alimi et al.[5] have proposed a method to cluster mutual funds using Ward method and k-means with fuzzy theory based on six characteristics including rate of return, variance, semivariance, turnover rate, Treynor index and Sharpe index for multi-objective portfolio optimization. These methods classifies mutual funds based on the historical investment performance similarities. However, they cannot classify the brand-new funds because we cannot yet measure the performance just after the fund was established. We need a new method to classify the mutual funds based on the current information instead of the past one. The contributions of this paper are: 1. To propose two new methods for classifying mutual funds based on the investment similarity, 2. To evaluate the proposed methods based on actual 551 Japanese mutual funds. Using these two method we can find some funds categorized to actively-managed funds but actually closet index ones where the fund manager tracks a benchmark stock index. In this paper, we use the two different approaches of k-means method and spectral clustering. In section 2 we describe the mutual fund data used in the experiment and the features of mutual funds that can be seen from the network . Next, in section 3, we describe the proposed method and in section 4 we outline our experimentation and evaluation. The summary is in section 5.

2. Network structure of mutual funds

2.1. Acquisition of investment trust data

To obtain basic information about the mutual fund, we extracted the "domestic stock investment trusts" top 10 investment stocks from the fund information that is published in Morningstar provided by Yahoo! Finance. Furthermore, using a regular expression in Python, we extracted data from the HTML file detail page of these funds in the experiment. These data include the brand name of the investment rated in the top 10 stocks of each fund, the stock industry, the investment rate, the net assets of the fund and the dividends,, and the transaction fee. The total number of extracted mutual funds is 551 and the number of investment stocks is 773, divided between 33 industries as stipulated by the Tokyo Stock Exchange.

2.2. Mutual fund network

To analyze the structure of the mutual funds, we use a network stretched edge between funds with weights for the number of common stocks in the top 10 stocks in each mutual fund. Figure 1 is a network that represents the mutual funds as circles and the weight as the number of edges. In addition, the definition of the network is shown in Figure 2. Let $I(v)$ be the set of investment stocks of mutual fund v, and $I_{10}(v)$ be the set of top 10 stocks of v. The weight of edge (v_i, v_j) is defined as follows:

$$w(v_i, v_j) = |I_{10}(v_i) \cap I_{10}(v_i)|.$$

The network diagram in Figure 1 shows that many mutual funds that invest to similar stocks are gathered in the center and we call them "common funds." On the other hand, there are some mutual funds that invest to different stocks that are away from the center. We call them "unique funds."

Examining the number of funds that are invested in each stock, it can be seen that very few stock contain in excess of 100 or 200 mutual funds, while most stocks contain investments from between 5 and 10 mutual funds.

2.3. Nikkei 225 and TOPIX

Nikkei 225 and TOPIX are both typical stock indexes in the Japanese stock market. The Nikkei 225 is calculated as a modified average from the highest 225 stocks trading in the activity and liquidity in the stocks listed on the First

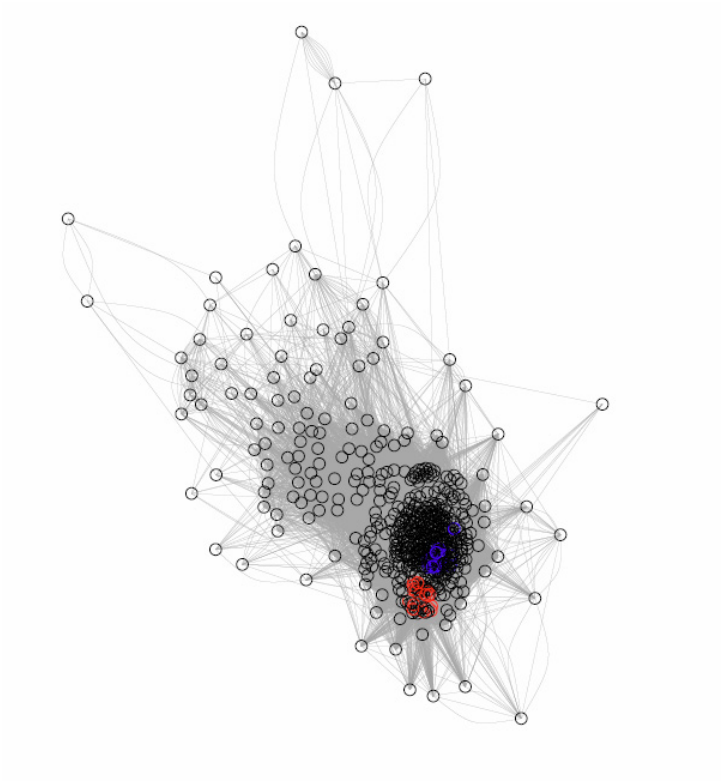

Fig. 1. Mutual fund network

Defined as V_f a set of mutual fund
Network: $G = (V, E)$
Node: $V = V_f$
Edge weight: $w(v_i, v_j)$ = the number of common shares in $v_i \in V_f$ and $v_j \in V_f$
Edge: $E = \{(v_i, v_j) \mid w(v_i, v_j) > 0\} \subseteq V_f \times V_f$

Fig. 2. Definition of network

Section of the Tokyo Stock Exchange using a calculation method based on the stock price average type system of the Dow Jones Industrial Average stock price.

TOPIX is a free-float adjusted market capitalization-weighted index that is calculated based on all the domestic common stocks listed on the TSE (Tokyo Stock Exchange) First Section. TOPIX shows the measure of current market capitalization assuming that market capitalization as of the base date (January 4 ,1968) is 100 points[6].

Although both indexes are calculated from the First Section of the Tokyo Stock Exchange, TOPIX is characterized by being less affected by stock price movements of specific industries and companies compared to the Nikkei 225. Therefore, it is desirable to classify the different categories of funds. Figure 1 represents a mutual fund that invests in the Nikkei 225 in red and in TOPIX in blue. Both red and blue circles in Figure 1 are clearly visible in the common funds group.

3. The proposed method

Here, we describe the method proposed in this study. In section 2.3, although the fund shows that there are typical stock indicators for the Nikkei 225 and TOPIX indexes, the Morningstar categorization is unable to divide them into

Input: A dataset $X = \{x_1, ..., x_N\}$, The number of clusters k
Output: Clusters $C_1, \cdots, C_k$
$dist(x_i, x_j)$: The Euclidean distance between $x_i \in X$ and $x_j \in X$

1. Assign $\forall x \in X$ randomly to $C_1, \cdots, C_k$.
2. Calculate the centroid $c_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \forall C_i$
3. Assign $\forall x \in X$ to the closest C_i: $i = \arg\min_{i=1,...,k} dist(x, c_i)$
4. **if** all of the cluster assignment does not change $\forall x \in X$ **then** end
   else go to Step 2

Fig. 3. k-means algorithm

two groups. In this study, a number that is divided into separate clusters of Nikkei 225 fund group and TOPIX fund group by increasing the number of fund divisions by clustering is defined as the optimal number of divisions. Next, we describe the k-means method and spectral clustering as the clustering method to be used.

k-means method

In k-means method, clusters are represented by the mean or centroid that minimizes an evaluation function:

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

where k is the number of clusters, x is the subject, C is the cluster, c is the centroid of the cluster, $dist(x, y)$ is the Euclidean distance between x and y. A search for the optimal solution is performed by iteratively recalculating the assignment and the representative point to the target cluster alternately. The k-means algorithm is shown in Figure 3. This approach is based on a hill-climbing method and selects the result of minimizing an evaluation function by changing the initial value randomly because we only want a local optimal solution. In this paper, we give a set of attributes a vector of length 773 (the number of investment destination brands) as a target set X that is given to the k-means method. Elements of the attribute vector represents the stock that has been invested as 1, otherwise 0, in the top 10 stocks of the investment proportion of funds. The distance between two mutual funds v_i and v_j is defined as the investment dissimilarity between v_i and v_j, which is defined as follows:

$$dist(v_i, v_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{jl})^2},$$

where if stock $s_l \in I_{10}(v_i)$ then x_{il} is 1, otherwise 0, and m is the number of stocks.

Spectral clustering

Clustering is performed as a matter of graph partitioning in spectral clustering [7,8,9]. This approach gives a similarity matrix for clustering a given graph unlike k-means. This approach performs clustering as a subgraph that is dense in the same cluster and is otherwise sparse. Spectral clustering takes advantage of the fact that the optimal solution cost to split into a sub-graph (the sum of the weights of the edges to remove when split) to the minimum corresponds to the solution of the eigenvalue problem. By solving eigenvalues of the Laplacian matrix derived from the similarity matrix of the graph, it is possible to perform clustering in a low dimension while maintaining the characteristics of the

Input: Similarity matrix W, The number of clusters k
Output: Clusters $C_1, \cdots, C_k$

1. Make a diagonal matrix D that satisfies $D_{ii} = \sum_j W_{ij}$ from W.
2. Make a diagonal matrix L that satisfies $L = I - D^{-1/2}WD^{-1/2}$ from W and D.
3. Calculate the eigenvalues and eigenvectors of L and create matrix C, which arranges the eigenvectors in order of k number of columns from those values of the eigenvalues that are small.
4. Apply k-means method to U.

Fig. 4. Spectral clustering algorithm

graph, meaning it is less likely to fall into the local optimal solution. In this paper, we define the similarity between mutual fund v_i and v_j as the number of common shares in v_i and v_j. And the similarity matrix W is defined as follows:

$$W = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix},$$

where $W_{ij} = w(v_i, v_j)$ described in the previous section and n is the number of mutual funds.

The spectral clustering algorithm is shown in Figure 4.

4. Experimental result

To evaluate two type of clustering methods, we have experimented with 551 Japanese mutual funds described in Section 2. We obtained the data of 551 Japanese mutual funds and the top 10 shares in their investment from Yahoo! Japan Finance, that are provided by Morningstar Inc. We then applied two proposed methods to the mutual funds in order to get fund clusters. Finally we calculated the average return and the variance of the returns for each cluster, that are based on the next one-month returns. We compared them with the categories provided by Morningstar Inc.

k-means

The optimal number of divisions in the k-means method is 4. Figure 5 is a network that is color-coded for each cluster to which the fund belongs, and Figure 6 is a graph showing the return mean and variance return of funds in each cluster. Here, a cluster that contains Nikkei 225 is color-coded blue, while TOPIX clusters are yellow.

Figure 6 shows that the k-means method is split into three common funds group and the unique funds group is combined into one. Furthermore, we can see that both the Nikkei 225 and the TOPIX clusters have low variance of returns, but the means of returns are different. Therefore it is necessary to divide them into different clusters. And Figure 6 indicates that clustering based on only the top 10 investment stocks classifies mutual funds well, since their clusters have low variance of returns.

Spectral clustering

The optimal number of divisions in spectral clustering is 7. Figure 7 is a network that has is color-coded for each cluster to which the fund belongs, and Figure 8 is a graph showing the return mean and variance return of funds in each cluster. Here, is a cluster that contains Nikkei 225 is color-coded in red, while TOPIX clusters are green.

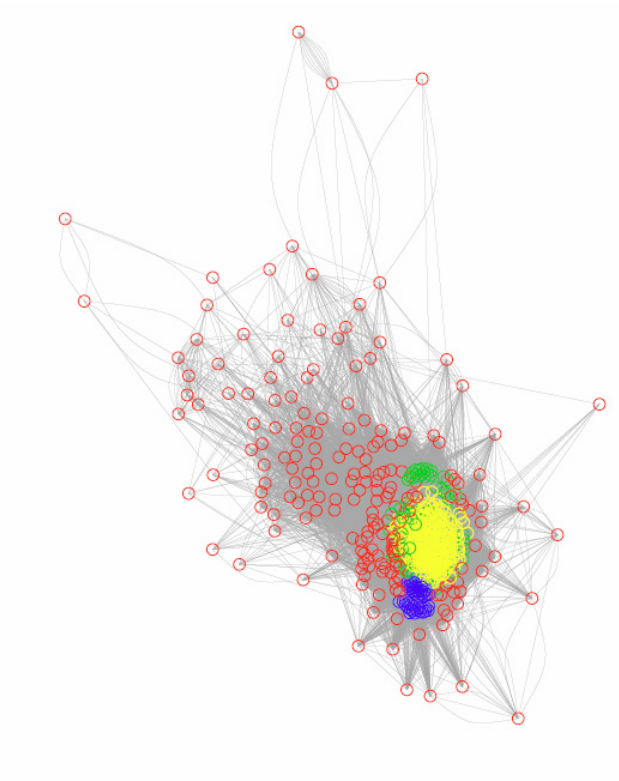

Fig. 5. *k*-means network

Spectral clustering shows that division has been performed in its unique funds group, which was combined into one in the *k*-means method. As with the *k*-means result, we can see that both the Nikkei 225 and the TOPIX clusters have low variance of returns, but the means of returns are different. Therefore, it is necessary to divide them into different clusters as in the *k*-means method. The spectral clustering that the group of funds summarized by the green cluster in the *k*-means method form several low clusters that are dispersive rather than split. And the orange cluster especially has low variance of returns. We got several the opinions about this cluster by stockbrokers. One of them is that we could find the third closet index funds group and the funds in their three index clusters are passive funds even in the case of calling themselves active mutual fund.

Morningstar category

Morningstar Inc. classifies Japanese mutual funds into nine categories based on the investment strategy, which is a combination of "investments in small, medium, and large-sized stock" and "value invested in undervalued stocks, growth invested in high-growth stock and blend combining value and growth." Figure 9 is a mutual funds network diagram in which the small value funds are in red, the medium value funds in green, the large value funds in blue, the small blend funds in yellow, the medium blend funds in purple, the large blend funds in orange, the small growth funds in black, the medium growth funds in light blue and the large growth funds in brown. Figure 10 is a graph showing the return mean and variance in return of funds in each category.

The network diagram shows that it is difficult to classify mutual funds properly by the investment stragety.

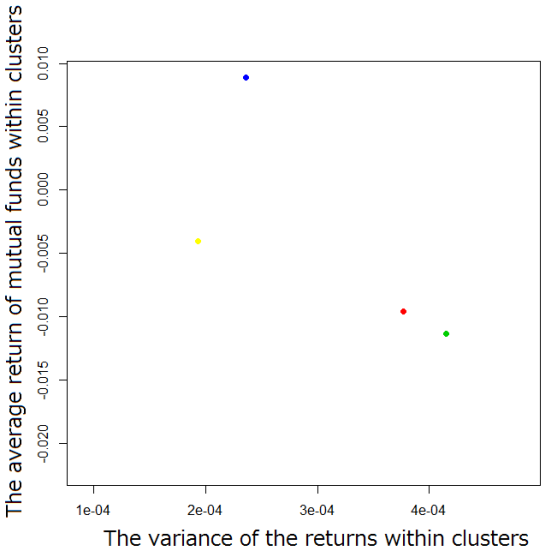

Fig. 6. *k*-means graph

5. Conclusion and future work

In this paper, we focus on the investment similarity of mutual funds and proposed two clustering method based on the investment similarity. The proposed methods can find mutual funds that use a different operation from the categorized strategy. One of the good points of the proposed methods is they can categorize new mutual funds because they do not need the historical performances. We can acquire the optimal number of clusters for each clustering technique when the Nikkei225 and TOPIX are firstly splitted into two clusters. We confirmed that the proposed methods can classify Japanese mutual funds more properly than the categories provided by Morningstar Inc. Thus, the proposed methods would help to find mutual funds that use different operation from the classified category strategy. Comparing the two proposed methods, i.e., the *k*-means method and spectral clustering, it was confirmed that the *k*-means method can divide finely in the center part of the network whereas spectral clustering can divide finely in the outer part.

A future challenge is a classification that considers the degree of similarity between stocks. Although there are several stocks that are similar (e.g., Toyota and Honda), they are determined to be invested in totally different stocks by the proposed method it may be possible to split mutual funds better if we can ameliorate this problem. Although in this study, Nikkei 225 and TOPIX provide the best division numbers when divided into two, it is necessary to verify whether this method is correct. Clustering methods used here to classify the given data always cluster somewhere, but there are instances that are not similar to any others in the mutual funds. Therefore, there is a chance of being wrongly grouped into a cluster. We may consider using extracts communities to extract only high similarity data groups rather than using completely divided data to solve this problem.

Acknowledgments

We thank SPARX Asset Management Co., Ltd. for their useful technical comments to our experimental results.

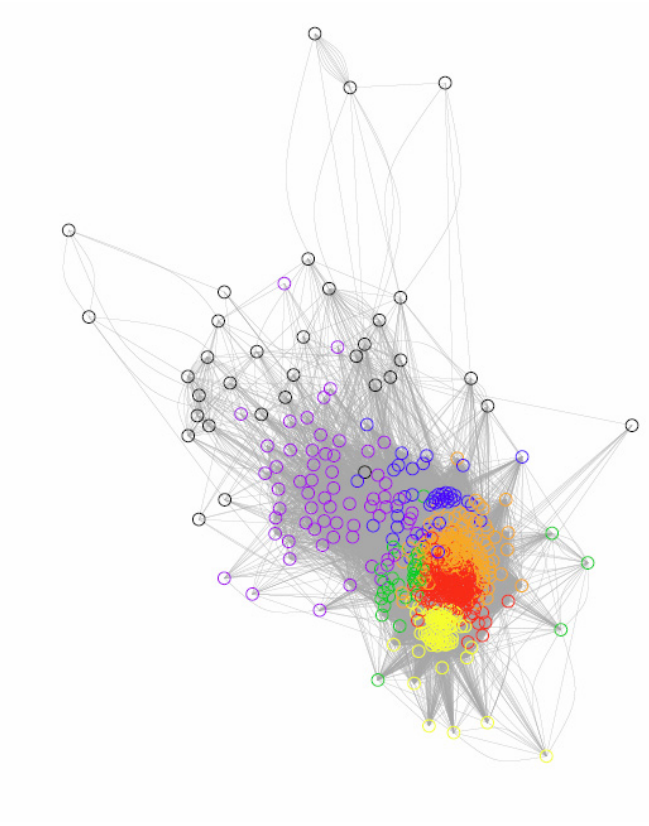

Fig. 7. Spectral clustering network

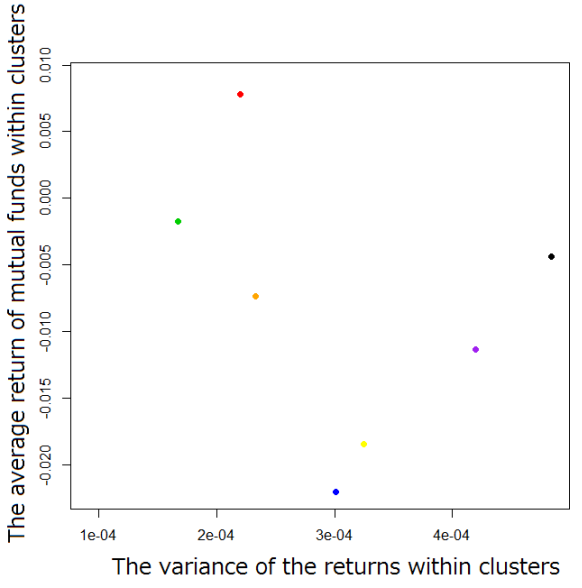

Fig. 8. Spectral cluster graph

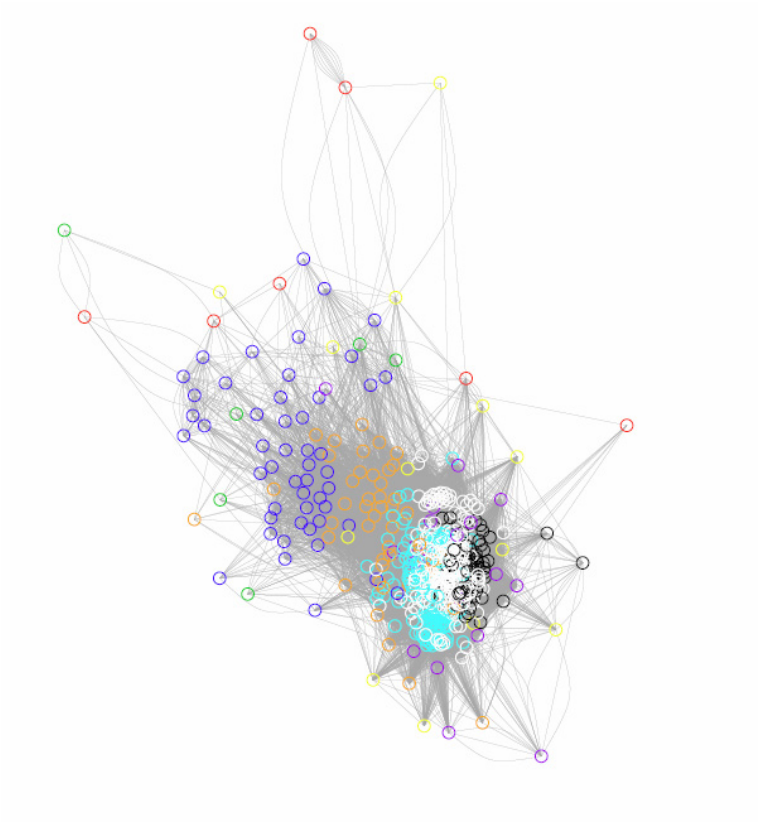

Fig. 9. Morningstar category network

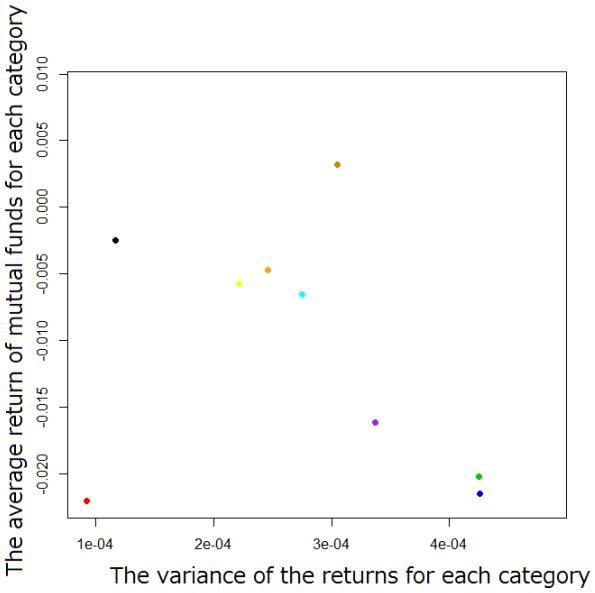

Fig. 10. Morningstar category graph

References

1. Cremers, KJ Martijn and Petajisto, Antti: "How active is your fund manager? A new measure that predicts performance," Review of Financial Studies, 22(9):3329–3365 (2009)
2. Suto, Megumi and Menkhoff, Lukas and Beckmann, Daniela: "Behavioural biases of institutional investors under pressure from customers: Japan and Germany vs the US," WP WIF-05-006, Waseda University Institute of Finance (2005)
3. Ramin Baghai-Wadji, Rami El-Berry, Stefan Klocker, and Markus Schwaiger: "The Consistency of Self-Declared Hedge Fund Styles ? A Return-Based Analysis with Self-Organizing Maps," Financial Stability Report 9:64–76, Oesterreichische Nationalbank (2005)
4. David Moreno, Paulina Marco, and Ignacio Olmeda: "Self-organizing maps could improve the classification of Spanish mutual funds," European Journal of Operational Research 174(2):1039–1054 (2006) uta
5. A. Alimi, M. Zandieh, and M. Amiri: "Multi-objective portfolio optimization of mutual funds under downside risk measure using fuzzy theory," International Journal of Industrial Engineering Computations, 3(5):859–872 (2012)
6. Japan Exchange Group: "What is TOPIX?," http://www.jpx.co.jp/english/markets/indices/topix/ (2015)
7. U. von Luxburg, "A tutorial on spectral clustering," Statistics and Computing, 17(4):395–416 (2007)
8. Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis, "kernlab—An S4 Package for Kernel Methods in R," Journal of Statistical Software, 11(9):1–20 (2004)
9. Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):888–905 (2000)