CALIFORNIA LUTHERAN UNIVERSITY - DEPARTMENT OF MATHEMATICS

Mutual Fund Similarity Through Graph Machine Learning

Binderiya Khurtsbaatar and Dr. Christopher Brown

California Lutheran UNIVERSITY

ABSTRACT

Introducing Fund2Vec: a graph machine learning approach to evaluate mutual fund similarity. By representing funds and assets as a weighted bipartite graph and using the node2vec algorithm, we gain a nuanced understanding of fund similarities. [1] The authors used a k-means clustering approach to tune the hyperparameters of the embedding. We proposed to replace k-means clustering with Gaussian model-based clustering for better separation of funds and assets in the embedding space. Our method outperforms traditional k-means clustering in identifying anomalous funds, aiding in risk management and investment strategy.

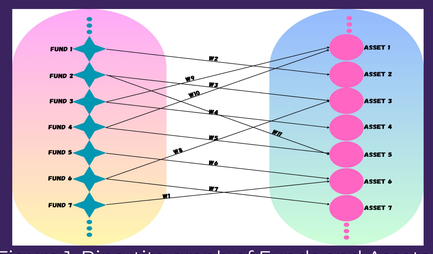

Figure 1. Bipartite graph of Funds and Assets

INTRODUCTION

Why Mutual Fund Similarity Matters

- **Explosion of Mutual Funds:** The proliferation of mutual funds makes choosing the right ones challenging.
- **Diverse Investment Goals: I**nvestors need tailored strategies to meet varied objectives.
- **Risk Management:** Understanding fund similarity helps in diversifying risk.

Challenges in Quantifying Fund Similarity

- **Subjectivity:** Traditional methods rely on human judgment or predefined categories, which may overlook similar behaviors across different categories.
- **Nonlinear Relationships:** Complex interactions between funds and assets can obscure crucial patterns.

BACKGROUND

- **Fund2Vec:** Evaluates mutual fund similarity through graph learning.
- **Node2vec Algorithm:** Embeds nodes (funds and assets) in a vector space for machine learning tasks.
- **Clustering Methods:** K-means and Gaussian model-based clustering are used to optimize node2vec hyperparameters and achieve the best separation in the embedding space.

METHOD

- **Data Collection:** SEC filings (NPORT-P, Q1 2020).
- **Hyperparameters:** Dimension (d), length of random walk (l), number of random walks per node (r), return parameter (p), in-out parameter (q).
- **Algorithms:** Node2vec for vector representation, followed by k-means and Gaussian model-based clustering.
- **Evaluation:** V-measure for clustering quality.

★ **The number of funds after cleaning:** 1052.

Reason of being removed: Incorrect data reporting, inaccurate identification numbers, and so on. We also found that most funds lay within a single connected component of the graph, and we removed all funds not in that single component.

★ **The number of assets after cleaning:** 166,531.

The **bipartite graph** has 1,052 X 166,531 (approximately 175.2 million) possible edges. The actual number of edges in the graph is 984,526. It is roughly about half a percent of the possible edges actually appeared as edges. This seems reasonable because funds do not attempt to invest in all possible assets but focus on a relative few.

The number of Funds	The number of Assets
1,052	166,531

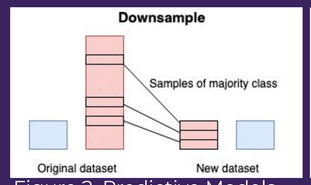

Figure 2. Predictive Models for Imbalanced Data: A School Dropout Perspective, https://www.researchgate.ne

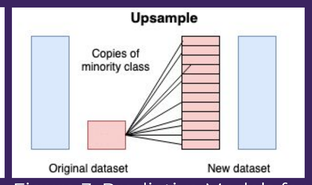

Figure 3. Predictive Models for Imbalanced Data: A School Dropout Perspective, https://www.researchgate.ne

Before clustering:

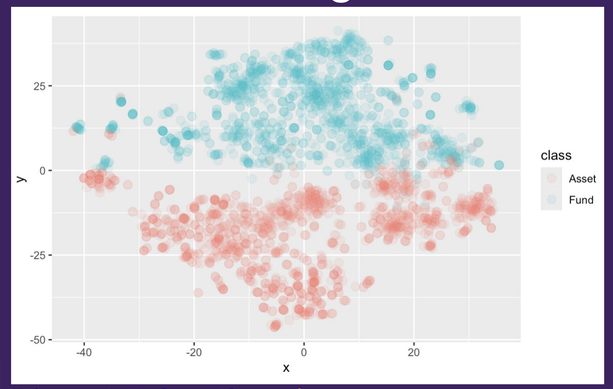

Figure 4. **Before Clustering**

After clustering:

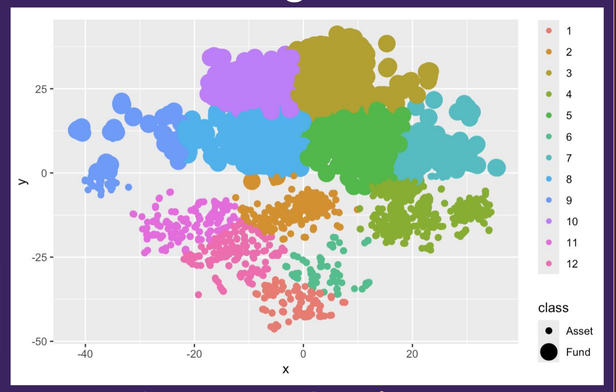

Figure 5. **After k-means clustering**

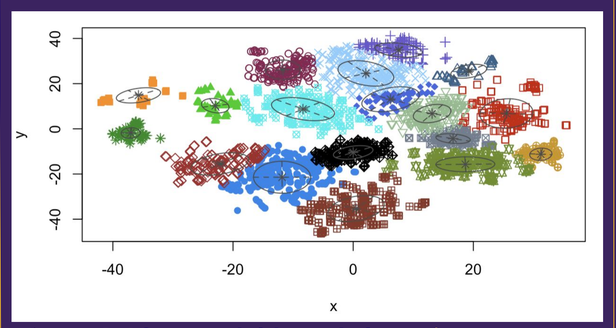

Figure 6. **After model-based clustering**

RESULTS

- **K-means Clustering:** Downsampling improved V-measure.
- **Model-based Clustering:** Identified anomalous funds, though V-measure was lower than k-means.

K-MEANS	V-measure
Downsampling	0.880
Upsampling	0.149

MODEL-BASED	V-measure
Downsampling	0.303
Upsampling	N/A

DISCUSSION

- **Anomalous Funds:** Identification aids investors in making informed decisions.
- **Future Directions:** Propose methods to assess clustering quality and further refine fund similarity evaluations.

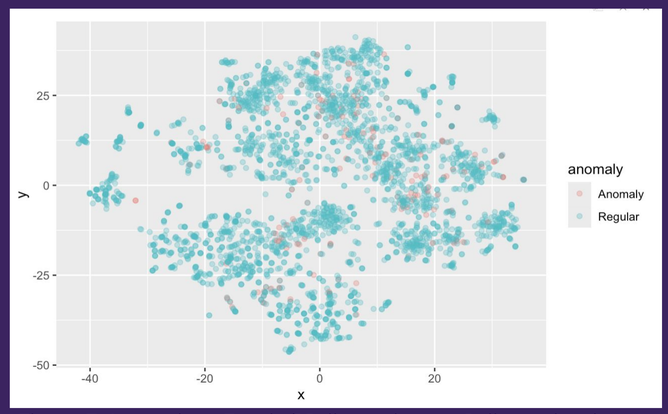

Figure 7. Anomalous Funds (16 dimensions to 2 dimensions)

REFERENCES

[1] Satone, Vipul, Dhruv Desai, and Dhagash Mehta. "Fund2vec: Mutual funds similarity using graph learning." Proceedings of the Second ACM International Conference on AI in Finance. 2021.

[2] Riesen, Kaspar, and Horst Bunke. Graph classification and clustering based on vector space embedding. Vol. 77. World Scientific, 2010.