Abordagem de *clustering* na otimização de portfólio com predição de série temporal baseada em algoritmos de aprendizado de máquina

Ana Paula dos Santos Gularte

Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos – SP – Brasil
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brasil
gularte@ita.br

1. Introdução

O clássico trabalho do economista Markowitz (1952, 1959) serviu como a gênese para a teoria moderna de portfólio, sendo a ferramenta mais conhecida e amplamente utilizada para alocação de capital em investimentos Qian et al. (2007). Todavia, a estrutura de *mean-variance (MV)* assume que todas as estimativas são igualmente distribuídas e os algoritmos de otimização tendem a exacerbar este problema, encontrando alocações extremas, e por sua vez, muito sensíveis a pequenas perturbações nas estimativas dos parâmetros de entrada do problema (retornos esperados e matriz de covariância) e, possivelmente, erros de modelagem Bertsimas et al. (2011); Kawas & Thiele (2011); Kim et al. (2018); Michaud (1989); Tutuncu & Koenig (2004); Xidonas et al. (2020).

A otimização de portfólio por meio da seleção e alocação adequada de ações, tem sido considerada uma das questões mais importantes em gestão de ativos e uma das principais decisões que um investidor enfrenta Elton et al. (2013). Neste contexto, a melhoria e otimização da carteira têm se tornado fundamentais na moderna pesquisa financeira e na tomada de decisão de investimentos Bodnar et al. (2017). Como sabemos, a construção do portfólio depende significativamente do desempenho futuro do mercado de ações, o que resulta em uma difícil tarefa de previsão dos preços dos ativos financeiros, pois geralmente são

não lineares, dinâmicos e caóticos. Desenvolvimentos recentes em aprendizado de máquina trouxeram oportunidades significativas para incorporar a teoria de predição na seleção de portfólio, entre as vantagens, está a capacidade de lidar com problemas não lineares e não estacionários, seguido pelo esforço em captar tendências e encontrar padrões complexos em grandes volumes de dados Henrique et al. (2019); Huck (2019); Wang et al. (2017), cujos resultados se destacam pelo potencial em gerar altos retornos de investimento e riscos de hedge Yang et al. (2019).

O modelo de otimização de portfólio de *mean-variance (MV)* de Markowitz (1952, 1959) alcança uma alocação ótima em um portfólio financeiro por meio da diversificação, medindo o risco esperado dos ativos para um retorno alvo. Estudos empíricos mostram que aumentar o número de ativos para diversificar uma carteira tem um efeito marginal positivo decrescente sobre o risco não sistemático Maringer (2005). No entanto, Tayali (2020) alerta que um grande número de ativos em um portfólio pode ter ramificações devido à maldição da dimensionalidade e aos altos custos de transação.

Uma restrição de cardinalidade supera esse obstáculo impondo um limite superior no número de ativos. Embora esse tipo de problema de otimização tenha vantagens sobre seu relaxamento, o modelo herda dificuldades computacionais, pois a restrição de cardinalidade transforma o problema em um programa inteiro misto de uma classe NP-completo. Como evidenciado no artigo publicado na revista IEEE Access® *"Optimal Portfolio Management for Engineering Problems Using Nonconvex Cardinality Constraint: A Computing Perspective"*, escrito por Khan et al. (2020). Dentre os trabalhos desta classe Ruiz-Torrubiano & Suarez (2010) afirmam que, *"Imposing limits on the number of diferente assets that can be included in the investment transforms portfólio selection into an NP-complete mixed-integer quadratic optimization problem that is difficult to solve by standard methods."*

Em virtude da dificuldade relatada, uma abordagem alternativa para a restrição de cardinalidade é reduzir o tamanho do problema antes de partir para o portfólio ideal. Redução, neste caso, é uma decisão para selecionar certos

ativos de um universo de ativos maior. Portanto, uma ferramenta de análise de dados de pré-processamento pode substituir a restrição de cardinalidade do modelo MV por um subconjunto de um universo de ativos maior. Além disso, algoritmos de agrupamento têm o potencial de minimizar ainda mais o risco medido no modelo MV e possuem a capacidade de melhorar a confiabilidade do portfólio em termos da razão entre risco previsto e realizado Tayali & Tolun (2018); Tola et al. (2008). Como um poderoso substituto para a restrição de cardinalidade no modelo MV, os métodos de agrupamento não só satisfazem a seleção de ativos e diversificação dos portfólios, escolhendo um ativo de cada *cluster* antes de executar o modelo MV, mas também aumentam a confiabilidade do portfólio, que é afetado pelos erros nos estimadores de média amostral e desvio padrão dos retornos Ren (2005); Tola et al. (2008).

Dessa forma, alguns estudos realizados integram a pré-seleção de ativos aos modelos de seleção de portfólio a fim de aliviar essa dificuldade. Além disso, métodos automatizados de *clusterização* de ativos para fins de diversificação, são inovações recentes que precisam ser mais explorados, uma vez que o conhecimento explícito do modelo MV no que diz respeito à medição de desempenho ainda é limitado Marvin (2015); Tayali (2020). Por exemplo, Ren (2005) na construção de portfólio utilizou métodos de *clustering* para agrupar ações altamente correlacionadas e, em seguida, usa esses *clusters* para executar o portfólio de *mean-variance (MV)*. Já Marvin (2015) propôs uma abordagem de *clusters* com uma medida alternativa a de similaridade de correlação, que se mostrou robusta em tempos de crise, resultando em carteiras com alto desempenho testadas em períodos pré e pós-crise. Já Paiva et al. (2019) propuseram um modelo de decisão de investimento que usa o *Support Vector Machines (SVMs)* para classificar ativos e os combina com o modelo *mean-variance (MV)* para formar uma carteira ótima. Wang et al. (2017) estudaram um método híbrido combinando uma rede neural recorrente *Long Short-Term Memory (LSTM)* com o modelo MV, que otimiza a formação do portfólio em combinação com a pré-seleção de ativos. No modelo de seleção de ações proposto por Yang et al. (2019) eles incorporam a predição das ações para capturar as tendências futuras da série do

mercado acionário, comprovando sob um nível de confiança de 95% que o novo método gera carteiras mais lucrativas, comparado com seleção de ações típicas, ou seja, sem predição de ações. Essas pesquisas mostraram que a combinação de previsão de ações e seleção de carteiras podem fornecer uma nova perspectiva para a análise financeira. Nesta pesquisa, o modelo MV é adotado para seleção de portfólio para determinar a proporção de cada ativo.

Com base nos antecedentes mencionados, esta dissertação examina a seguinte questão: Como construir uma carteira de ações incorporando técnicas de aprendizado de máquina tanto para seleção dos ativos, quanto para predição de retornos e simultaneamente, reduzir os riscos inerentes à formação da carteira?

Especificamente, dois estágios estão envolvidos neste estudo: seleção de portfólio e previsão de ações. O primeiro estágio consiste em agrupar os ativos em vários *clusters* usando os métodos: i) *Agglomerative Hierarchical Clustering*, ii) *K-means*, iii) *Partition Around Medoids (PAM)* e iv) *Uniform Manifold Approximation and Projection (UMAP)*, como um substituto para a restrição de cardinalidade e uma ferramenta de análise de dados de pré-processamento no modelo *mean-variance (MV)*. No segundo estágio, um modelo híbrido para predição de retornos das ações combinando: 1) *Gradient Boosting Machine (GBM)*, 2) *K-Nearest Neighbor (K-NN)* e 3) *Bayesian Regularized Neural Networks (BRNN)* com *ensemble* dos modelos é proposto para prever os preços das ações para o próximo período e o modelo de MV é empregado para alocação dos pesos da carteira. Utilizando a série temporal histórica de ativos listados em 3 índices: Ibovespa, Dow Jones e S&P 500 como amostra do estudo que abrange um horizonte de tempo começando em 2016 a 2020. Cada janela contínua nos dados da série temporal será composta durante os períodos de pré-crise e pós-crise para análise da performance da carteira.

2. Justificativa

A otimização de portfólio associada a previsão de ações e métodos de agrupamento mostram inúmeras vantagens em sua aplicabilidade como:

(i) Algoritmos de agrupamento são capazes de filtrar as informações relevantes em um conjunto multivariado de dados, heterogêneos entre si e mutuamente exclusivos, sendo assim, maximiza a homogeneidade dos objetos dentro dos grupos, e maximiza a heterogeneidade entre os demais grupos;

(ii) Outros estudos indicam que algoritmos de agrupamento são bastante robustos no que diz respeito a medição do ruído devido a finitude do tamanho da amostra. Isso é particularmente verdadeiro para um conjunto de variáveis hierarquicamente organizadas Tumminello et al. (2007). Ainda nesse sentido, León et al. (2017) destacam o comportamento estável de portfólios baseados em *cluster* abordando uma das questões críticas em mercados financeiros que é a volatilidade;

(iii) É testado o potencial de cada agrupamento na seleção de portfólio usando diferentes algoritmos, explorando medidas de distâncias como Chebychev, Manhattan ou Minkowski para avaliar a semelhança ou dissimilaridade dos elementos, além de critérios, parâmetros e velocidade de convergência distintos;

(iv) Alguns estudos destacam a capacidade que os algoritmos de aprendizado de máquina têm em lidar com problemas não lineares e não estacionários Chen et al. (2021). Cujos resultados mostram que a precisão desses métodos de inteligência artificial é superior aos métodos estatísticos tradicionais, contribuindo ainda mais com o aprendizado por ensemble, cujo objetivo é reduzir o viés e a variância da predição e obter melhor desempenho preditivo do que um único algoritmo. Outrossim, modelos de previsão que consideram a importância da otimização dos parâmetros do modelo, e o uso de dados recentes como a abordagem BRNN se destacam na pesquisa de Yan et al. (2016), que propõem um incremento no algoritmo para lidar com problemas de capacidade de generalização e sobreajuste na previsão;

(v) A metodologia proposta diferencia-se dos estudos existentes introduzindo um quadro de *backtesting* dinâmico de *walk forward optimization* onde os parâmetros de agrupamento para a pré-seleção de ativos são atualizados para cada janela contínua do horizonte de investimento;

(vi) Estudos recentes evidenciam que não há nenhum estudo na literatura científica relacionada sobre uma simulação histórica multiperíodo de uma estratégia de investimento com os parâmetros de agrupamento dentro da amostra, sistematicamente atualizados para investigar os efeitos sobre os dados fora da amostra gerados pelo modelo MV Tayali (2020).

3. Metodologia

São ilustradas as principais técnicas utilizadas na Figura 1 e detalhadas a seguir.

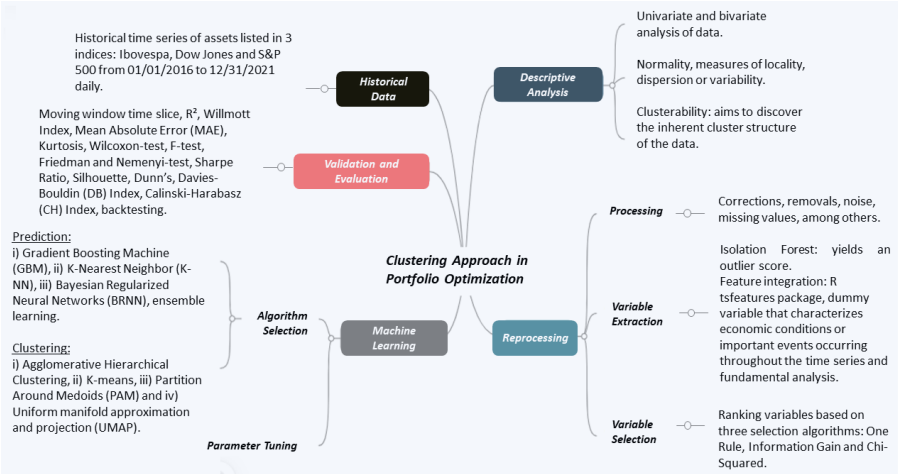

Figura 1: Pipeline da predição de ações e abordagem de *cluster* na otimização de portfólio. Fonte: Autoria própria, 2021.

Referências

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *Society for Industrial and Applied Mathematics*, *53*,

464–501. doi:10.1137/080734510.

Bodnar, T., Mazur, S., & Okhrin, Y. (2017). Bayesian estimation of the global minimum variance portfolio. *European Journal of Operational Research*, *256*, 292–307. doi:10.1016/j.ejor.2016.05.044.

Chen, W., Zhang, H., Mehlawat, M. K., & Jia, L. (2021). Mean–variance portfolio optimization using machine learning-based stock price prediction. *Applied Soft Computing*, *100*, 106–943. doi:10.1016/j.asoc.2020.106943.

Elton, E. J., Gruber, M. J., Brown, S. J., & Goetzmann, W. N. (2013). *Modern Portfolio Theory and Investment Analysis - Ninth edition.*.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226–251. doi:10.1016/j.eswa.2019.01.012.

Huck, N. (2019). Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research*, *278*, 330–342. doi:10.1016/j.ejor.2019.04.013.

Kawas, B., & Thiele, A. (2011). A log-robust optimization approach to portfolio management. *OR Spectrum*, *33*, 207–233. doi:10.1007/s00291-008-0162-3.

Khan, A. H., Cao, X., Katsikis, V. N., Stanimirović, P., Brajević, I., Li, S., Kadry, S., & Nam, Y. (2020). Optimal portfolio management for engineering problems using nonconvex cardinality constraint: A computing perspective. *IEEE Access*, *8*, 57437–57450. doi:10.1109/ACCESS.2020.2982195.

Kim, J. H., Kim, W. C., & Fabozzi, F. J. (2018). Recent advancements in robust optimization for investment management. *Annals of Operations Research*, *266*, 183–198. doi:10.1007/s10479-017-2573-5.

León, D., Aragón, A., Sandoval, J., Hernández, G., Arévalo, A., & Niño, J. (2017). Clustering algorithms for risk-adjusted portfolio construction. *Procedia Computer Science*, *108*, 1334–1343. doi:10.1016/j.procs.2017.05.185.

Maringer, D. G. (2005). *Diversification in Small Portfolios. In: Portfolio Management with Heuristic Optimization.*. doi:10.1007/0-387-25853-1_4.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*, 77–91. doi:10.1111/j.1540-6261.1952.tb01525.x.

Markowitz, H. (1959). Portfolio selection: efficient diversification of investments. *Cowles Foundation for Research in Economics at Yale University*, .

Marvin, K. (2015). Creating diversified portfolios using cluster analysis. *Independent Work Report Fall*, .

Michaud, R. O. (1989). The markowitz optimization enigma: Is "Optimized optimal?. *Financial Analysts Journal*, *45*, 31–42. doi:10.2139/ssrn.2387669.

Paiva, F., Cardoso, R., Hanaoka, G., & Duarte, W. (2019). Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications*, *115*, 635–655. doi:10.1016/j.eswa.2018.08.003.

Qian, E. E., Hua, R. H., & Sorensen, E. H. (2007). *Quantitative Equity Portfolio Management: Modern Techniques and Applications.*.

Ren, Z. (2005). Portfolio construction using clustering methods. *A Thesis submitted to the Faculty of the Worcester Polytechnic Institute. In partial fulfillment of the requirements for the Professional Masters Degree in Financial Mathematics*, .

Ruiz-Torrubiano, R., & Suarez, A. (2010). Hybrid approaches and dimensionality reduction for portfolio selection with cardinality constraints. *IEEE Computational Intelligence Magazine*, *5*, 92–107. doi:10.1109/MCI.2010.936308.

Tayali, H. A., & Tolun, S. (2018). Dimension reduction in mean-variance portfolio optimization. *Expert Systems with Applications*, *92*, 161–169. doi:10.1016/j.eswa.2017.09.009.

Tayali, S. T. (2020). A novel backtesting methodology for clustering in mean–variance portfolio optimization. *Knowledge-Based Systems*, *209*, 106454. doi:10.1016/j.knosys.2020.106454.

Tola, V., Lillo, F., Gallegati, M., & N., M. R. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, *32*, 235–258. doi:10.1016/j.jedc.2007.01.034.

Tumminello, M., Lillo, F., & Mantegna, R. N. (2007). Hierarchically nested factor model from multivariate data. *Europhysics Letters in press*, . doi:10.1209/0295-5075/78/30006.

Tutuncu, R. H., & Koenig, M. (2004). Robust asset allocation. *Annals of Operations Research*, *132*, 157–187. doi:10.1023/B:ANOR.0000045281.41041.ed.

Wang, W., Li, W., Zhang, N., & Liu, K. (2017). Portfolio formation with preselection using deep learning from long-term financial data. *Expert Systems with Applications*, *143*, 113042. doi:10.1016/j.eswa.2019.113042.

Xidonas, P., Steuer, R., & Hassapis, C. (2020). Robust portfolio optimization: a categorized bibliographic review. *Annals of Operations Research*, *292*, 533–552. doi:10.1007/s10479-020-03630-8.

Yan, D., Zhou, Q., Wang, J., & Zhang, N. (2016). Bayesian regularisation neural network based on artificial intelligence optimisation. *International Journal of Production Research*, *55*, 2266–2287. doi:10.1080/00207543.2016.1237785.

Yang, F., Chen, Z., Li, J., & Tang, L. (2019). A novel hybrid stock selection method with stock prediction. *Applied Soft Computing*, *80*, 820–831. doi:10.1016/j.asoc.2019.03.028.