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

Ana Paula dos Santos Gularte/Felipe dos Santos Alves Feitosa/Vinícius Henrique Pinto Pacheco/Vitor Venceslau Curtis

Resumo

A otimização de portfólio por meio da seleção e alocação adequada de ações é de forma mútua um tema fundamental na moderna pesquisa financeira e também uma das questões mais relevantes relacionadas às decisões de um investidor. Sob essa ótica, é visto que, de modo geral, o comportamento não-linear e caótico dos preços de ações promove cada vez mais que modelos de aprendizado de máquina possam ser empregados na predição voltada à seleção de portfólio, gerando benefícios relacionados à utilização de grandes volumes de dados e na identificação de padrões complexos que podem impactar positivamente a relação risco x retorno dos investimentos. Visto isso, através de uma série temporal de ativos listados em três índices: Ibovespa, Dow Jones e S&P 500 entre o período de 2016 a 2020, empiricamente, a pesquisa propõe utilizar-se da abordagem de *clustering* na otimização de portfólio por meio de algoritmos de aprendizado de máquina, os quais objetivam selecionar um portfólio de ações (cujos parâmetros de agrupamento serão atualizados para cada janela contínua do horizonte de investimento) e implementar um modelo híbrido (*ensemble*) de algoritmos de predição voltados à previsão dos retornos das ações para o próximo período, além de um modelo de *mean-variance* (MV) voltado à alocação dos pesos da carteira.

Palavras-chaves: Otimização de Portfólio, Aprendizado de Máquina, Clustering

Aderência à área prioritária do MCTIC

O presente projeto adere-se à área prioritária do MCTIC de Tecnologias Habilitadoras, mais especificamente ao setor de Inteligência Artificial. Pois a pesquisa irá utilizar-se de forma sistêmica da aplicação de conceitos e técnicas comuns ao setor com o objetivo de analisar séries temporais e empiricamente implementar algoritmos com foco na criação de modelos híbridos de Aprendizado de Máquina.

Método e regularidade de orientação do aluno

A orientação do projeto será feita por meio de reuniões semanais de cerca de 1 hora através da plataforma Google Meet. Nas reuniões serão propostas a avaliação das entregas previamente planejadas entre a orientadora e os orientados para a semana. As entregas estimam uma dedicação mínima de cerca de 15 horas semanais dos orientados. Cabe ressaltar que o contato direto entre os envolvidos no projeto pode ser feito sem restrições por meio do aplicativo de conversas WhatsApp.

Problema e objetivos

O problema alvo da pesquisa é: 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?

Dessa forma, a pesquisa propõe-se a empiricamente inserir uma abordagem de *clustering* na otimização de portfólio, utilizando-se de algoritmos de aprendizado de máquina tanto para Seleção do Portfólio de ações (cujos parâmetros de agrupamento serão atualizados para cada janela contínua do horizonte de investimento) de três índices de referência (Ibovespa, Dow Jones e S&P 500) no período de 2016 a 2020, quanto para propor a implementação de um modelo híbrido (*ensemble*) de algoritmos de Predição voltados à previsão dos retornos das ações para o próximo período, além da utilização de um modelo de *mean-variance* (MV) voltado à alocação dos pesos da carteira.

Justificativa e relevância científica

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 à 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).

Metodologia de pesquisa

A metodologia de pesquisa aplicada a este projeto segue as seguintes etapas:

(i) Utilização de séries temporais históricas de preços de ativos listados em 3 índices: Ibovespa, Dow Jones e S&P 500 em um horizonte de tempo entre 2016 a 2020.

(ii) Realização de uma Análise Descritiva dos dados, que consistirá em análises univariadas e bivariadas dos dados, testes de normalidade, medidas de localidade, dispersão e variabilidade, além de análises de *Clusterability* (voltada à identificação da estrutura de clusters inerentes aos dados).

(iii) O Reprocessamento dos dados será baseado em correções, remoção de ruídos e tratamento de dados faltantes, entre outros. Além disso, propõe-se a utilização de bibliotecas presentes nas linguagens de programação Python e R a fim de obter-se a identificação de Anomalias/*Outliers* (através do uso de algoritmos como *Isolation Forest*), bem como o uso de variáveis de classificação (*dummy variables*), que na série terão o objetivo de caracterizar condições econômicas e de indicadores de análise fundamentalista.

(iv) A Seleção do Algoritmo de Aprendizado de Máquina terá em caráter empírico dois principais objetivos: Seleção do Portfólio de ações e a Predição de retornos de forma a simultaneamente reduzir os riscos inerentes à formação da carteira. Dessa forma, em um

primeiro momento sob a ótica de *Clustering,* será realizada implementações que irão abranger os algoritmos de *Agglomerative Hierarchical Clustering*, *K-Means*, *Partition Around Medoids (PAM)* e *Uniform Manifold Approximation and Projection (UMAP)*. Em um segundo estágio, tem-se a implementação em um modelo híbrido (*ensemble*) de algoritmos de predição, tais como: *Gradient Boosting Machine (GBM), K-Nearest Neighbor (K-NN), Bayesian Regularized Neural Networks (BRNN)* e *Ensemble Learning*, bem como o uso de um modelo de *mean-variance* (MV) voltado à alocação dos pesos da carteira. Cabe ressaltar que em paralelo às implementações, ajustes de parâmetros dos modelos gerados serão realizados sistematicamente, com a finalidade de se obter melhores acurácias e performances.

(v) Por fim, tem-se após Validação e Avaliação dos portfólios gerados, que contará com a aplicação de cálculos de janelas temporais com diferentes amplitudes, R^2, Índice de Willmott, Mean Absolute Error (MAE), Curtose, Teste de Wilcoxon, Teste-F, Teste de Friedman e Nemenyi, Índice de Sharpe, Pontuação de Silhouette, Teste de Dunn, Índice de Davies-Bouldin (DB), Índice de Calinski-Harabasz e Backtesting.

Viabilidade de execução

Para a execução do projeto não será necessário o uso da infraestrutura física ou quaisquer recursos específicos disponibilizados pelo Instituto Tecnológico de Aeronáutica - ITA.

Cronograma e avaliação dos resultados

O Cronograma de execução do projeto seguirá as seguintes etapas:

- Projeto completo em 31/07.
- Progresso medido semanalmente com a entrega das atividades listadas na semana anterior.
- Submissão do artigo para participar do ENCITA em 21/03.
- Projeto será apresentado no ENCITA em 26/05.
- Os projetos são avaliados em arquivo disponibilizado no GitHub.
- Submissão de artigo no periódico *Expert Systems with Applications* em 31/07.

Bibliografia

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.

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.

Lopez de Prado, Marcos*. Advances in Financial Machine Learning*. 1. ed. Nova Jersey:John Wiley & Sons, Inc., 2018.

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.

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.