

Portfolio Optimization with Rebalancing: A Hybrid Approach with Genetic Algorithms and Reinforcement Learning

Yá-Sin Barcelos Mghazli

Universidade Federal do Rio de Janeiro - UFRJ

Programa de Engenharia Elétrica

yasinbarcelos@ufrj.br

Pedro Henrique González

Universidade Federal do Rio de Janeiro - UFRJ

Programa de Engenharia de Sistemas e Computação

pegonzalez@cos.ufrj.br

RESUMO

Este trabalho apresenta um framework para otimização sequencial de portfólio com alocação discreta de ativos, incorporando transferência de conhecimento entre períodos de rebalanceamento. O problema é modelado como uma extensão do Mochila Multidimensional com Repetição, visando maximizar o Índice de Sharpe sob restrições realistas de orçamento, cardinalidade, alocação e custos de transação. São comparadas três abordagens: uma exata via programação não-linear inteira mista, um algoritmo genético discreto e uma versão híbrida com aprendizado por reforço. Experimentos com dados reais do mercado brasileiro demonstram que a abordagem híbrida apresenta maior robustez e desempenho superior em retorno ajustado ao risco, especialmente em cenários de alta volatilidade, evidenciando seu potencial para aplicações práticas em ambientes financeiros dinâmicos.

PALAVRAS CHAVE. Otimização de Portfólio, Rebalanceamento Periódico, Algoritmos Genéticos, Aprendizado por Reforço, Programação Não-Linear Inteira Mista

Tópicos: OA – Outras Aplicações em PO , IA – PO e IA

ABSTRACT

This work presents a framework for sequential portfolio optimization with discrete asset allocation, incorporating knowledge transfer between rebalancing periods. The problem is modeled as an extension of the Multidimensional Knapsack Problem with Repetition, aiming to maximize the Sharpe Ratio under realistic constraints of budget, cardinality, allocation, and transaction costs. Three approaches are compared: an exact method using mixed-integer nonlinear programming, a discrete genetic algorithm, and a hybrid version with reinforcement learning. Experiments with real data from the Brazilian market show that the hybrid approach achieves greater robustness and superior risk-adjusted returns, especially in highly volatile scenarios, highlighting its potential for practical applications in dynamic financial environments.

KEYWORDS. Portfolio Optimization, Periodic Rebalancing, Genetic Algorithms, Reinforcement Learning, Mixed-Integer Nonlinear Programming

1. Introdução

A otimização de portfólios é uma ferramenta utilizada por investidores que buscam maximizar sua riqueza ao longo do tempo, equilibrando retorno e risco em um ambiente financeiro incerto [Markowitz, 1952]. Ao distribuir recursos entre diferentes ativos, os investidores não apenas diversificam suas exposições, mas também buscam estratégias resilientes às oscilações do mercado. Essa busca por alocação eficiente torna-se ainda mais relevante em horizontes de longo prazo, nos quais decisões precisam ser ajustadas continuamente conforme novas informações surgem.

Entretanto, abordagens tradicionais frequentemente assumem um horizonte estático, em que a alocação é definida no início e mantida até o final do período. Essa suposição ignora a complexidade dos mercados reais, especialmente em economias emergentes, onde fatores como alta volatilidade, liquidez restrita, custos de transação e correlações instáveis entre ativos impõem desafios adicionais à construção de portfólios eficientes. Nessas condições, é importante que investidores adotem estratégias adaptativas, ajustando suas decisões em resposta a mudanças frequentes nas condições de mercado [Li et al., 2022; Nystrup et al., 2019].

Nesse contexto, a otimização multi-período de portfólios tem se destacado como uma abordagem mais realista. Ao permitir decisões recorrentes de alocação e rebalanceamento, essa perspectiva reconhece a necessidade de adaptação contínua frente à incerteza dos retornos futuros. Contudo, ela também introduz novas camadas de complexidade, como o acúmulo de custos operacionais, o risco de perdas prolongadas [Bienstock, 1996] e a necessidade de manter a viabilidade da carteira diante de choques adversos.

Este trabalho propõe uma abordagem metodológica para otimização e rebalanceamento dinâmico de portfólios em um cenário no qual o investidor gerencia, ao longo de vários períodos, uma carteira sujeita a custos de transação, limites operacionais e condições de mercado instáveis. O objetivo principal é maximizar a riqueza final do investidor, respeitando restrições práticas e mitigando riscos, em um processo autofinanciado, sem aportes ou retiradas adicionais de capital. O desempenho das estratégias é avaliado por meio de experimentos computacionais, com rebalanceamentos periódicos e custos operacionais realistas, em diferentes cenários de mercado.

A proposta combina técnicas heurísticas com métodos adaptativos. Inicialmente, utiliza-se um Algoritmo Genético Discreto como mecanismo robusto para explorar o espaço combinatório de composições de portfólio, respeitando restrições como limites de alocação, setorial, cardinalidade e custos de transação. Em seguida, integra-se um componente de Aprendizado por Reforço, que ajusta dinamicamente os parâmetros evolutivos com base na experiência acumulada ao longo das gerações, permitindo uma adaptação eficiente aos padrões temporais observados [Almahdi and Yang, 2017; Kuo and Chou, 2021]. Para fins comparativos, também é formulada uma abordagem exata baseada na maximização do Índice de Sharpe, modelada como um problema de programação não linear inteira mista (MINLP), resolvido com o Knitro solver [Byrd et al., 1999].

Este artigo está organizado da seguinte forma: a Seção 2 apresenta a formulação matemática do problema de otimização de portfólios multi-período, detalhando variáveis, parâmetros, função objetivo e restrições. A Seção 3 descreve a metodologia adotada, incluindo o processamento de dados, as estratégias de otimização via MINLP e Algoritmos Genéticos, além da integração com Aprendizado por Reforço. A Seção 4 apresenta os experimentos computacionais, com a descrição da amostra, configuração dos testes, análise de desempenho in-sample e out-of-sample, e discussão dos resultados. Por fim, a Seção 5 conclui o estudo com um resumo das principais descobertas e sugestões para pesquisas futuras.

2. Definição do Problema

Este trabalho aborda a otimização de portfólios sob uma perspectiva de rebalanceamento periódico com transferência de conhecimento entre períodos. Nesse contexto, o investidor deve

alocar recursos entre diferentes ativos financeiros e ajustar sua carteira periodicamente. Foi adotada uma abordagem de horizonte deslizante (*rolling horizon*), onde cada período é otimizado sequencialmente, mas incorporando os portfólios dos períodos anteriores. Fatores como custos de transação, limites operacionais e a natureza estocástica dos retornos são considerados a cada rebalanceamento [Brandt, 2005; Garlappi et al., 2009].

O problema de rebalanceamento periódico adotando uma estrutura discreta, na qual as decisões de alocação são modeladas como quantidades inteiras de unidades de cada ativo. O objetivo em cada período t é maximizar o desempenho ajustado ao risco do portfólio, enquanto consideramos o impacto das decisões passadas nas condições iniciais do período corrente. A formulação considera restrições de capital, limites de alocação individual e setorial, cardinalidade e custos operacionais, refletindo aspectos práticos da gestão de portfólios [Chang et al., 2000].

3. Metodologia

Este trabalho propõe uma abordagem híbrida para a otimização de portfólio multiperíodo sequencial com formulação discreta, combinando algoritmos genéticos (GA) com aprendizado por reforço profundo (Dueling DQN). O sistema opera em ciclos sucessivos de rebalanceamento: em cada ciclo, dados históricos são processados para estimar parâmetros de retorno, risco e correlação; em seguida, o portfólio é otimizado por um GA, cujos parâmetros são ajustados dinamicamente por um agente de RL com base em métricas populacionais como diversidade e estagnação [Huang et al., 2018; Bianchi et al., 2017]. Após um ciclo, o portfólio é rebalanceado considerando custos de transação, liquidez e restrições operacionais, e seu desempenho é avaliado por métricas como retorno anualizado, Índices de Sharpe e Sortino, drawdown máximo. As informações obtidas retroalimentam o sistema, influenciando decisões nos ciclos subsequentes ao longo do horizonte de investimento. Essa metodologia visa combinar a robustez exploratória dos algoritmos genéticos com a adaptabilidade dos métodos de aprendizado por reforço, promovendo uma otimização mais responsiva e eficiente em ambientes financeiros dinâmicos e sujeitos a incertezas.

3.1. Processamento e Análise de Dados

Os dados históricos foram obtidos por meio da biblioteca `yfinance` e submetidos a um processo estruturado de limpeza e preparação. Valores ausentes nas séries temporais foram preenchidos por interpolação linear, assegurando a continuidade dos dados. Em seguida, aplicou-se o método do intervalo interquartil (IQR) para detecção de outliers, os quais foram substituídos pela média móvel de 5 períodos, de modo a preservar a estrutura temporal das séries [Garlappi et al., 2009].

Com os dados limpos, calcularam-se os retornos diários logarítmicos para cada ativo i no tempo t . A escolha por retornos logarítmicos deve-se às suas propriedades estatísticas desejáveis, como aditividade temporal e maior aderência à normalidade [Tvede, 2012].

3.2. Janelamento Adaptativo

Em cada período de rebalanceamento t , os dados históricos são divididos em dois subconjuntos: um conjunto de treinamento, utilizado para a estimação de parâmetros e otimização do portfólio, e um conjunto de avaliação fora da amostra, que se estende do final do treinamento até o início do próximo rebalanceamento. A abordagem de janelamento utilizada neste trabalho é deslizante acumulativa: para cada t , considera-se todo o histórico disponível até o instante atual como conjunto de treinamento. Essa estratégia visa maximizar o volume de dados utilizado na estimação dos parâmetros de retorno, risco e correlação, favorecendo maior estabilidade estatística. O desempenho do portfólio é então avaliado com base no comportamento observado no intervalo subsequente, o que permite testar a capacidade de generalização da estratégia em condições de mercado não vistas durante o treinamento [Huang et al., 2018].

3.3. Framework

O framework de otimização de portfólio (Figura 1) implementa um ciclo de rebalanceamento com cinco etapas interconectadas: inicialização, processamento de dados com janelamento, otimização, rebalanceamento considerando custos e avaliação de desempenho. Dois mecanismos adaptativos caracterizam sua dinâmica: a transferência de conhecimento entre ciclos, que aprimora as heurísticas de otimização, e o janelamento temporal incremental, que incorpora continuamente novos dados de mercado, permitindo adaptação às condições variáveis do ambiente financeiro.

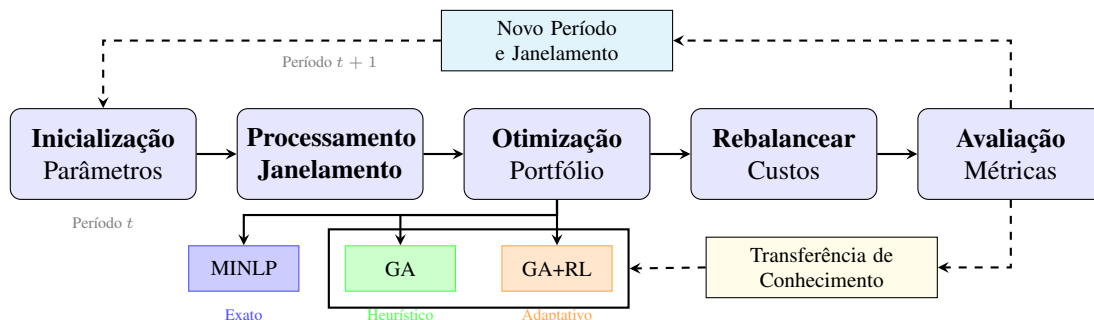


Figura 1: Framework de otimização de portfólio com rebalanceamento periódico

3.3.1. Variáveis e Parâmetros

Na Tabela 1 definem-se os conjuntos, variáveis de decisão e parâmetros, servindo como base para a formulação da função objetivo e das restrições. Os conjuntos indicam os elementos do universo do problema, como ativos financeiros, setores econômicos e períodos. As variáveis de decisão representam as escolhas feitas pelo investidor em cada período, incluindo a seleção de ativos, quantidades adquiridas ou vendidas e o controle de presença no portfólio. Já os parâmetros incluem informações de mercado, como preços, retornos, riscos, limites operacionais e custos transacionais, que influenciam diretamente o espaço viável de soluções [Fabozzi et al., 2007].

Tabela 1: Conjuntos, Variáveis de Decisão e Parâmetros

Conjuntos e Índices	
$\mathcal{I} = \{1, \dots, n\}$	Conjunto de ativos disponíveis para investimento
$\mathcal{S} = \{1, \dots, m\}$	Conjunto de setores econômicos
Variáveis de Decisão	
d_i	Quantidade de ações do ativo i no portfólio atual
$x_i \in \{0, 1\}$	Variável binária que indica se o ativo i está presente no portfólio
z_i^+, z_i^-	Quantidade de ações compradas e vendidas do ativo i
Parâmetros	
r_f	Taxa livre de risco vigente
p_i e r_i	Preço atual e retorno esperado do ativo i
σ_{ij}	Covariância entre os retornos dos ativos i e j
W	Orçamento total disponível para alocação
c^+, c^-	Custo de transação proporcional para compra e venda
K, K_s	Número máximo de ativos no portfólio e no setor s
L_i, U_i	Limites mínimo e máximo de alocação para o ativo i
$S(i) \in \mathcal{S}$	Sector ao qual o ativo i pertence
U_s	Limite percentual máximo do capital que pode ser investido no setor s
d_i^{prev}	Quantidade herdada do ativo i no portfólio anterior

3.3.2. Função Objetivo e Dinâmica Temporal

Em nossa abordagem de rebalanceamento periódico, para cada período t , o objetivo é maximizar o Índice de Sharpe do portfólio, considerando as condições iniciais herdadas do período anterior. O Índice de Sharpe é uma métrica amplamente utilizada na teoria moderna de portfólios, ao permitir avaliar a eficiência de uma estratégia ao relacionar o retorno excedente ao risco assumido Sharpe [1966]. Para cada período de rebalanceamento t , o problema de otimização é definido como:

$$\max \left[\frac{R_p - r_f}{\sigma_p} \right] \quad (1)$$

$$W_{next} = \sum_{i=1}^n p_i \cdot d_i \cdot (1 + R_p) - \sum_{i=1}^n (c^+ \cdot p_i \cdot z_i^+ + c^- \cdot p_i \cdot z_i^-) \quad (2)$$

$$d_i = d_i^{prev} + z_i^+ - z_i^-, \quad \forall i \in \mathcal{I} \quad (3)$$

Por fim, a dinâmica temporal do portfólio é descrita pelas restrições 2 e 3. A restrição 2 modela a evolução da riqueza total no tempo, considerando os retornos dos ativos e os custos de transação (de compra e venda). A Restrição 3 define a atualização da quantidade de ações mantidas no portfólio com base nas transações realizadas.

3.3.3. Programação Não-Linear Inteira Mista (MINLP)

A abordagem de Programação Não-Linear Inteira Mista (MINLP) foi empregada como método exato para resolver o problema de otimização de portfólio descrito neste trabalho. Essa escolha é motivada pela estrutura do problema, que combina uma função objetivo não-linear e não-convexa com restrições de natureza inteira e binária. A formulação do modelo inclui variáveis discretas $d_{i,t}$, que representam as quantidades inteiras de ativos alocados no portfólio, e variáveis binárias $x_{i,t}$, que indicam a presença ou ausência de cada ativo.

A modelagem foi realizada por meio da biblioteca *amplpy*, que permite a formulação matemática do problema de forma estruturada e sua integração com *solvers* especializados. A biblioteca foi utilizada para representar o modelo de otimização em cada período de rebalanceamento, com base nos dados históricos previamente processados, incluindo retornos esperados e a matriz de covariância dos ativos [Inc., 2021].

A otimização é conduzida pelo *solver* Knitro, um solver para problemas não-lineares mistos, capaz de lidar com a complexidade da estrutura do problema. O Knitro utiliza uma abordagem baseada em *Branch-and-Bound* com relaxação contínua para variáveis inteiras, combinada com métodos de pontos interiores para resolver as subproblemas contínuas. Essa estratégia permite explorar o espaço de soluções de forma eficiente, garantindo a obtenção de soluções globais ótimas mesmo em problemas não-convexos com restrições inteiras.

A formulação do problema é complementada por um conjunto de **restrições financeiras e operacionais**, que refletem aspectos práticos da alocação de ativos. A *restrição de orçamento*, expressa na Restrição 4, garante que o montante total investido não ultrapasse a riqueza disponível no início de cada período. A restrição de cardinalidade, apresentada na Restrição 5, limita o número de ativos distintos que podem compor o portfólio:

$$\sum_{i=1}^n p_i \cdot d_i \leq W \quad (4)$$

$$\sum_{i=1}^n x_i \leq K, \quad \text{onde } x_i = \mathbb{1}(d_i > 0) \quad (5)$$

As Restrições 6 a 8 impõem restrições de alocação e setoriais. A Restrição 6 assegura que a proporção alocada em cada ativo esteja entre limites inferiores L_i e superiores U_i , enquanto a Restrição 7 limita a exposição total a um setor s por meio de um limite superior U_s . A Restrição 8 restringe o número de ativos pertencentes a um mesmo setor, por meio do parâmetro K_s .

$$L_i \cdot x_i \leq \frac{p_i \cdot d_i}{W} \leq U_i \cdot x_i, \quad \forall i \in \mathcal{I} \quad (6)$$

$$\frac{\sum_{i:S(i)=s} p_i \cdot d_i}{W} \leq U_s, \quad \forall s \in \mathcal{S} \quad (7)$$

$$\sum_{i:S(i)=s} x_i \leq K_s, \quad \forall s \in \mathcal{S} \quad (8)$$

As restrições de transação e não-negatividade estão detalhadas em 9 a 11. A restrição 9 garante que as quantidades compradas e vendidas sejam não negativas e que não se possa vender mais do que se possui. As restrições 10 e 11 asseguram que todas as variáveis de quantidade de ativos e transações sejam inteiras e não negativas.

$$z_i^+ \geq 0, \quad z_i^- \geq 0, \quad z_i^- \leq d_i^{\text{prev}}, \quad \forall i \in \mathcal{I} \quad (9)$$

$$d_i \in \mathbb{Z}^+ \cup \{0\}, \quad \forall i \in \mathcal{I} \quad (10)$$

$$z_i^+, z_i^- \in \mathbb{Z}^+ \cup \{0\}, \quad \forall i \in \mathcal{I} \quad (11)$$

Apesar de oferecer garantias teóricas de otimalidade, a principal limitação da abordagem MINLP está no custo computacional, que cresce exponencialmente com o número de ativos e de períodos considerados. Em instâncias maiores, a estratégia adotada busca uma solução viável dentro de limites operacionais predefinidos, o que assegura a obtenção de resultados robustos mesmo em cenários de alta complexidade [Bonami et al., 2009].

3.3.4. Algoritmo Genético com Representação Discreta

O modelo de otimização por Algoritmos Genéticos com representação discreta foi desenvolvido com a biblioteca *DEAP* (Distributed Evolutionary Algorithms in Python), adaptado para lidar com as restrições do problema de alocação de portfólio com ativos indivisíveis e limites por setor e ativo. Cada indivíduo da população representa uma alocação de unidades inteiras de ativos no portfólio, respeitando os limites operacionais impostos. A codificação genética é dada por:

$$\mathcal{P} = (d_{1,t}, d_{2,t}, \dots, d_{n,t}), \quad \text{com } d_{i,t} \in \mathbb{Z}_{\geq 0} \quad (12)$$

Essa representação permite capturar de forma realista a indivisibilidade dos ativos financeiros, além de viabilizar a aplicação de operadores genéticos sob restrições.

A avaliação dos indivíduos é realizada por uma função de *fitness* que incorpora o objetivo de maximizar o Índice de Sharpe, com penalizações suaves aplicadas a soluções inviáveis. A penalização é proporcional ao grau de violação das restrições, promovendo a exploração de regiões viáveis do espaço de busca. O *fitness* é definido como:

$$\text{Fitness}(\mathcal{P}) = \text{Sharpe}(\mathcal{P}) - \lambda \cdot \text{Penalty}(\mathcal{P}), \quad (13)$$

onde λ é um hiperparâmetro ajustável e $\text{Penalty}(\mathcal{P})$ agrega as violações normalizadas de orçamento, cardinalidade, e limites setoriais e individuais.

A evolução é conduzida por operadores genéticos customizados. Para cada operação de crossover ou mutação, um operador é selecionado aleatoriamente dentre os disponíveis, promovendo diversidade e evitando estagnação precoce. Os operadores utilizados são:

1. **Operadores de Crossover: Um Ponto ou Dois Pontos**(seleciona um corte aleatório e troca as seções entre dois pais); **Blend Crossover** (combina os valores dos pais usando uma média ponderada); **Uniform Crossover** (para cada gene, seleciona o valor de um dos pais com base em probabilidade).
1. **Operadores de Mutação: Mutação Simples** (para cada gene, com uma probabilidade, aumenta ou diminui a quantidade em uma unidade); **Mutação Adaptativa** (ajusta a intensidade da mutação com base no valor relativo do ativo no portfólio); **Gaussian Mutation** (adiciona ruído gaussiano aos genes selecionados); **Shuffle Mutation** (reorganiza as quantidades entre ativos selecionados).

O processo evolutivo utiliza *seleção por torneio* com reposição, favorecendo indivíduos mais aptos, e mantém um *Hall of Fame* que preserva o melhor indivíduo já encontrado. Essa prática assegura que soluções de alta qualidade não se percam ao longo das gerações.

Adicionalmente, implementa-se um mecanismo de *transferência de conhecimento inter-temporal*: uma fração da população inicial de cada período de rebalanceamento é composta por mutações leves do melhor indivíduo do período anterior. Essa estratégia, inspirada em técnicas de aprendizado incremental [Kuo and Chou, 2021; Almahdi and Yang, 2017], tem como objetivo acelerar a convergência do algoritmo, preservando, ao mesmo tempo, a diversidade genética necessária para evitar a estagnação prematura da busca.

Os parâmetros evolutivos, como tamanho da população, número de gerações, taxa de crossover e de mutação, foram ajustados empiricamente com base em validação cruzada nos períodos anteriores. Essa configuração promove um equilíbrio entre qualidade da solução e custo computacional, como sugerido por Huang et al. [2018]; Kuo and Chou [2021].

3.3.5. Algoritmos Genéticos com Aprendizado por Reforço

O núcleo da nossa contribuição metodológica consiste em um esquema híbrido que integra algoritmos genéticos (GA) e aprendizado por reforço (RL) para a otimização dinâmica de portfólios. Enquanto os GAs exploram exaustivamente o espaço combinatório de alocações por meio de operadores de crossover e mutação [Kalyanakrishnan et al., 2025; Masuda and Meng, 2024], o agente de RL aprende padrões temporais e adapta as decisões de seleção e avaliação de indivíduos com base em recompensas acumuladas [Pomorski and Gorse, 2023]. Nesta arquitetura, o agente de RL atua como um controlador meta-heurístico que, em cada geração, seleciona quais operadores genéticos aplicar, ajusta parâmetros de taxa de mutação e tamanho de torneio, e decide quando injetar diversidade na população, de modo a manter um balanceamento eficaz entre exploração e diversidade [Miao and Zhang, 2023; Muhle-Karbe et al., 2016]. Essa integração capitaliza a capacidade do RL de generalizar experiências passadas para guiar a evolução populacional rumo a regiões promissoras do espaço de busca, mitigando a convergência prematura típica de GAs puros [Brandt and Santa-Clara, 2004; Peng et al., 2020]. Estudos recentes comprovam que híbridos GA+RL superam ambos os métodos isolados em problemas de otimização sequencial, incluindo cenários de rebalanceamento multiperíodo com custos de transação e restrições de risco [Li et al., 2022; Nystrup et al., 2019].

- **Estado (s_t):** Representado por um vetor que captura estatísticas da população atual (média, variância e máximo de fitness); métricas de diversidade; taxa de melhoria da melhor solução; risco de estagnação; geração atual e máximo de gerações; e o histórico de performance das ações tomadas anteriormente.
- **Ações (a_t):** O espaço de ações inclui a escolha do tipo de operador de crossover e mutação; ajuste dinâmico das probabilidades de crossover e mutação; ajuste da pressão seletiva no

torneio e taxa de elitismo; e a possibilidade de reinicialização parcial da população para manutenção da diversidade.

- **Recompensa (r_t):** A função de recompensa é dada por:

$$r_t = \alpha \cdot \Delta f_{max} + \beta \cdot D_t - \gamma \cdot S_t, \quad (14)$$

onde Δf_{max} é a melhoria na melhor fitness encontrada; D_t é a diversidade da população no tempo t ; S_t é o risco de estagnação no tempo t ; e α , β e γ são pesos configuráveis.

Para melhorar a robustez e a generalização do agente de RL, foram incorporadas estratégias de **Ruído Gaussiano** aos dados de retorno, com intensidade adaptativa ao longo do tempo, incentivando a exploração no início e a convergência nas etapas posteriores. Adicionalmente, adota-se uma abordagem de *Curriculum Learning*, onde o agente começa com versões simplificadas do problema e gradualmente enfrenta cenários mais complexos à medida que se torna mais proficiente.

Por fim, implementa-se **Transferência de Conhecimento** entre os períodos de rebalanceamento, utilizando soluções de alta qualidade de períodos anteriores para iniciar o treinamento no novo período. O histórico acumulado também é usado para ajustar a adaptação do agente às novas dinâmicas do mercado, permitindo uma evolução contínua e mais eficiente do processo de otimização [Bianchi et al., 2017]. O fluxo do processo de treinamento é descrito no Algoritmo 1.

Algorithm 1: Pseudocódigo do Dueling DQN para controle adaptativo do GA

```

1 Initialize:
2   Networks  $Q_{local}$  and  $Q_{target}$  (Dueling architecture), Replay Buffer
3   Hyperparameters:  $\alpha$  (learning rate),  $\gamma$  (discount factor),  $\tau$  (soft update factor)
4 foreach training episode do
5    $s_0 \leftarrow \text{environment.reset}()$ 
6   foreach GA generation do
7     Select action  $a_t$  via  $\epsilon$ -greedy policy
8     Apply  $a_t$  to GA environment, evolve the population
9     Observe  $r_t, s_{t+1}, done$ 
10    Store transition  $(s_t, a_t, r_t, s_{t+1}, done)$  in the buffer
11     $s_t \leftarrow s_{t+1}$ 
12    if buffer size  $\geq$  batch size then
13      Sample minibatch of experiences
14      foreach transition in minibatch do
15        if done then
16           $y_i \leftarrow r_i$ 
17        else
18           $a^* \leftarrow \arg \max_{a'} Q_{local}(s_{i+1}, a')$ 
19           $y_i \leftarrow r_i + \gamma \cdot Q_{target}(s_{i+1}, a^*)$ 
20        Minimize loss:  $\mathcal{L} = (y_i - Q_{local}(s_i, a_i))^2$ 
21      Update target network:
22       $\theta_{target} \leftarrow \tau \theta_{local} + (1 - \tau) \theta_{target}$ 
23    Update  $\epsilon$  using sigmoid decay

```

4. Experimentação Computacional

Os 42 ativos selecionados para o estudo, distribuídos entre cinco setores econômicos com base em sua representatividade no Ibovespa conforme a Tabela 2.

Tabela 2: Ativos utilizados no estudo por setor econômico e representatividade no Ibovespa

Sector	Ativos	% Ibovespa
Financeiro	BBAS3, BBDC3/4, BPAC11, SANB11, ITSA4, BBSE3, CASH3	21,85%
Materiais Básicos	VALE3, GGBR4, SUZB3, KLBN11, CSNA3, GOAU4, BRKM5, USIM5, CBAV3, RADL3	11,36%
Petróleo, Gás e Biocombustíveis	PETR3/4, PRIO3, CSAN3, RECV3, BRKM3, UGPA3, RAIZ4	13,74%
Utilidade Pública	ELET3/6, EGIE3, CMIG4, CPLE6, SBSP3, ENGI11, EQTL3, NEOE3	12,14%
Consumo Cíclico	MGLU3, LREN3, AMER3, PETZ3, CVCB3, ALPA4, YDUQ3	3,88%

Nota: Códigos seguidos de "n" indicam diferentes classes da mesma empresa

A Tabela 3 e 4 apresentam os parâmetros utilizados na implementação e execução dos experimentos.

Parâmetro	Valor	Descrição
Parâmetros do Ambiente de Investimento		
W_0	R\$10.000,00	Capital inicial
T	12-60 meses	Horizonte temporal dos experimentos
c^+, c^-	0,1%	Custos de transação de compra e venda
r_f	12,11% a.a.	Taxa livre de risco
K	10	Número máximo de ativos no portfólio
K_s	3	Limite máximo de ativos por setor
L_i	3%	Alocação mínima por ativo
U_i	25%	Alocação máxima por ativo
U_s	40%	Limite máximo de exposição por setor
Parâmetros de Análise de Dados e MINLP		
Janela	Adaptativa	Tipo de janelamento para estimação
Tamanho da janela	252-756 dias	Intervalo da janela temporal (1-3 anos)
Frequência	Diária	Frequência dos dados utilizados
Retornos	Logarítmicos	Tipo de cálculo de retorno
Método de covariância	Shrinkage	Método de estimação da matriz de covariância
IQR threshold	1,5	Limiar para detecção de outliers
Solver	Knitro	Solver utilizado para otimização não-linear
Tempo máximo	600 segundos	Limite de tempo de execução
Gap de otimalidade	1e-4	Critério de parada por gap relativo
Método	Branch-and-Bound	Algoritmo de resolução

Tabela 3: Parâmetros: Ambiente e MINLP

Parâmetro	Valor	Descrição
Parâmetros do Algoritmo Genético		
População	300	Tamanho da população
Gerações	700	Número máximo de gerações
$P_{crossover}$	85%	Probabilidade de aplicação de crossover
P_{mutao}	20%	Probabilidade de mutação
Seleção	Torneio (k=3)	Método de seleção com tamanho do torneio
Elitismo	5%	Percentual dos melhores indivíduos mantidos
% transferência	30%	Percentual de soluções transferidas entre períodos
Parâmetros do Aprendizado por Reforço		
Arquitetura	Dueling DQN	Arquitetura da rede neural
α	0,001	Taxa de aprendizado
γ	0,95	Fator de desconto
τ	0,001	Fator de atualização suave da rede-alvo
$\epsilon_{inicial}$	1,0	Valor inicial de epsilon para exploração
ϵ_{final}	0,01	Valor final de epsilon
Batch size	64	Tamanho do lote para treinamento
Replay buffer	10.000	Tamanho do buffer de experiências
α_{reward}	0,7	Peso para melhoria na função de recompensa
β_{reward}	0,2	Peso para diversidade na função de recompensa
γ_{reward}	0,1	Peso para penalidade de estagnação
σ_{base}	0,03	Desvio padrão base do ruído adaptativo

Tabela 4: Parâmetros: GA e RL

A Figura 2 mostra que o método MINLP tende a formar portfólios mais concentrados e estáveis ao longo do tempo. Já o GA apresenta maior variabilidade nas alocações, refletindo seu caráter exploratório. A abordagem GA+RL, por sua vez, equilibra diversidade e estabilidade, indicando que o componente de aprendizado por reforço contribui para decisões mais consistentes na alocação dos ativos.

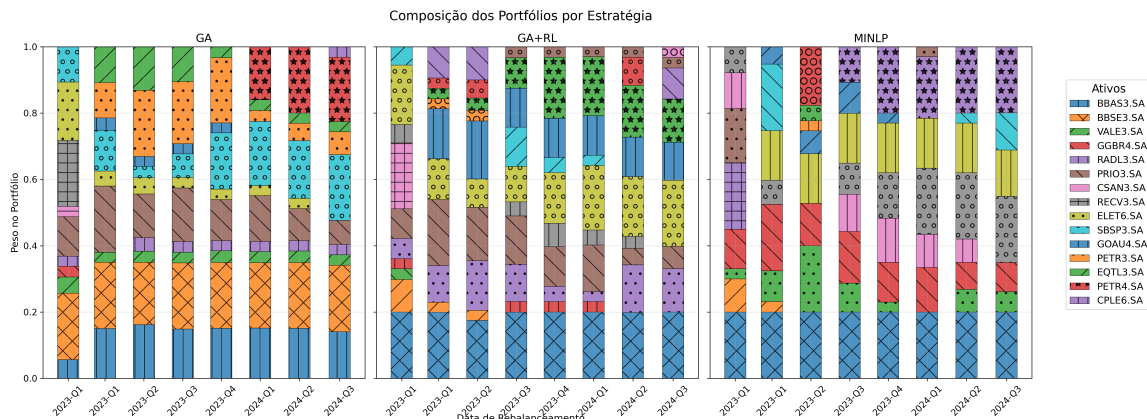


Figura 2: Composição dos Portfólios por Estratégia.

4.1. Desempenho dos Métodos de Otimização

Os resultados da fase de treinamento (in-sample) estão apresentados na Tabela 5 indicam que o método exato (MINLP) obteve o melhor desempenho em termos de Sharpe, refletindo uma ótima relação risco-retorno. O método proposto (GA+RL) apresentou desempenho competitivo, com maior consistência e menor risco que o GA, o qual, embora tenha alcançado maior retorno, exibiu maior volatilidade e drawdown.

Tabela 5: Estatísticas de performance in-sample dos métodos

Modelo	Retorno Anualizado (%)	Volatilidade (%)	Sharpe	Drawdown Máx. (%)	Sortino	Calmar	Kappa 3
MINLP	48,64 ± 06,32	1,28 ± 0,10	1,49 ± 0,17	-13,71 ± 0,96	2,31 ± 0,27	2,11 ± 1,32	1,62 ± 0,20
GA	52,92 ± 10,44	1,46 ± 0,21	1,44 ± 0,17	-15,65 ± 1,14	2,22 ± 0,25	1,96 ± 1,17	1,56 ± 0,19
GA+RL	46,85 ± 05,73	1,28 ± 0,09	1,43 ± 0,16	-14,02 ± 0,93	2,23 ± 0,26	1,94 ± 1,02	1,55 ± 0,19

Os resultados de validação (out-of-sample) apresentados na Tabela 6 indicam que o método proposto (GA+RL) obteve o melhor desempenho em termos de retorno e Sharpe, evidenciando maior capacidade de generalização e adaptação a novos dados. Como esperado, o modelo exato, otimizado para o período de treinamento, não manteve a liderança na validação, reforçando o potencial do GA+RL frente à incerteza dos cenários futuros.

Tabela 6: Estatísticas de performance out-of-sample dos métodos

Método	Retorno Anualizado (%)	Volatilidade (%)	Sharpe	Drawdown Máx. (%)	Sortino	Calmar	Kappa 3
MINLP	19,09 ± 39,15	0,90 ± 0,20	0,29 ± 2,12	-7,46 ± 6,14	0,96 ± 4,02	53,26 ± 82,22	0,60 ± 2,45
GA	19,58 ± 38,22	0,90 ± 0,21	0,30 ± 1,98	-7,30 ± 5,99	0,86 ± 3,63	51,03 ± 75,37	0,57 ± 2,29
GA+RL	23,09 ± 39,58	0,90 ± 0,19	0,50 ± 2,06	-6,99 ± 5,26	1,32 ± 3,71	66,66 ± 95,44	0,85 ± 2,41

A Figura 3 ilustra a evolução dos portfólios no período de validação, evidenciando a superioridade do método GA+RL, com valorização mais estável e robusta frente à volatilidade do mercado. Os métodos MINLP tiveram desempenho intermediário, enquanto o GA mostrou maior instabilidade e menor retorno.

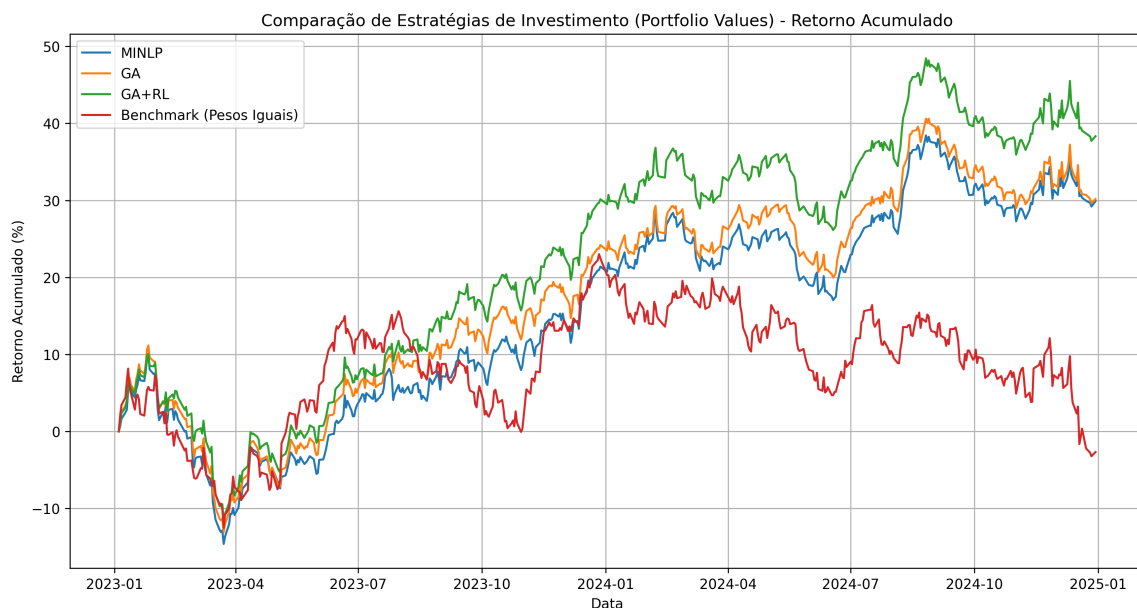


Figura 3: Comparação de Estratégias de Investimento - Retorno Acumulado

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou um framework para otimização sequencial de portfólio com variáveis discretas, integrando algoritmos genéticos e aprendizado por reforço sob restrições realistas de mercado. Os resultados computacionais indicam que, enquanto o modelo exato obteve melhor desempenho in-sample, a abordagem híbrida destacou-se na validação out-of-sample, alcançando maior robustez e retorno ajustado ao risco, especialmente em cenários de alta volatilidade. O algoritmo genético, embora competitivo em retorno, apresentou maior instabilidade e drawdown.

A análise trimestral revela a sensibilidade dos modelos às condições de mercado, com desempenho expressivo nos períodos de alta e resultados moderados durante períodos de maior instabilidade. Esta variação temporal evidencia a importância de estratégias adaptativas, que demonstraram maior consistência na preservação de capital durante diferentes ciclos de mercado.

A composição dos portfólios otimizados reflete a estrutura do mercado brasileiro, com concentração em setores de maior liquidez, especialmente nos setores Financeiro e de Petróleo, Gás e Biocombustíveis, refletindo tanto a estrutura do mercado brasileiro quanto a capacidade dos algoritmos em identificar oportunidades em ativos com maior liquidez.

O desempenho superior do modelo híbrido demonstra que a integração de algoritmos evolutivos com aprendizado por reforço proporciona ganhos significativos em adaptabilidade e robustez em ambientes de mercado dinâmicos. Esta abordagem híbrida representa uma contribuição relevante para a pesquisa em otimização de portfólios discretos multi-período.

Como trabalhos futuros, sugerimos incorporar modelos de previsão para estimação de retornos e volatilidades, o desenvolvimento e análise de mecanismos de transferência de aprendizado entre diferentes mercados e períodos, a exploração de técnicas de meta-learning para ajuste automático de hiperparâmetros, além da extensão do framework para otimização multi-objetivo e abordagens robustas que incorporem explicitamente a incerteza dos parâmetros, permitindo maior resiliência a erros de estimação em ambientes financeiros complexos.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores agradecem também ao CNPq e a FAPERJ pelo financiamento parcial desta pesquisa.

Referências

- Saeed Almahdi and Saeid Y. Yang. Adaptive portfolio management: A deep reinforcement learning approach. *Expert Systems with Applications*, 83:187–195, 2017.
- Fabio Bianchi, Sergio M. Focardi, and Tamás Tóth. Portfolio optimization with genetic algorithms and reinforcement learning: A hybrid approach. *Financial Engineering & Automated Trading*, 21:54–70, 2017. doi: 10.2139/ssrn.2963956.
- Daniel Bienstock. Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74(2):121–140, 1996.
- Pierre Bonami, Mustafa Kilinc, and Jeffrey T. Linderoth. Algorithms and software for convex mixed integer nonlinear programs. In *Mixed Integer Nonlinear Programming*, pages 1–39. Springer, 2009. doi: 10.1007/978-0-387-87987-9_1.
- Michael W. Brandt. Portfolio choice problems. In Yacine Aït-Sahalia and Lars Peter Hansen, editors, *Handbook of Financial Econometrics*, pages 269–336. Elsevier, 2005.
- Michael W. Brandt and Pedro Santa-Clara. Dynamic portfolio selection by augmenting the asset space. *NBER Working Paper*, 2004.

- Richard H. Byrd, Mary E. Hribar, and Jorge Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.
- Tzung-Jin Chang, Nick Meade, John E. Beasley, and Yasser M. Sharaiha. Heuristics for cardinality constrained portfolio optimization. *Computers & Operations Research*, 27(13):1271–1302, 2000.
- Frank J. Fabozzi, Petter N. Kolm, Dessislava A. Pachamanova, and Sergio M. Focardi. *Robust Portfolio Optimization and Management*. John Wiley & Sons, 2007.
- Lorenzo Garlappi, Raman Uppal, and Tan Wang. Portfolio selection with parameter and model uncertainty: A multi-prior approach. *The Review of Financial Studies*, 22(1):1–41, 2009. doi: 10.1093/rfs/hhn018.
- Zhiyu Huang, Guo Li, Zhi Liu, and Xiaokang Li. Multi-period portfolio optimization using genetic algorithm and particle swarm optimization. *Computers & Industrial Engineering*, 115:190–201, 2018. doi: 10.1016/j.cie.2017.10.018.
- AMPL Optimization Inc. *amplpy: Python api for ampl*, 2021. URL <https://github.com/ampl/amplpy>.
- Shivaram Kalyanakrishnan et al. Hybrids of reinforcement learning and evolutionary computation in financial applications. *ACM Transactions on Intelligent Systems*, 2025.
- Shu Yu Kuo and Yao Hsin Chou. Building Intelligent Moving Average-Based Stock Trading System Using Metaheuristic Algorithms. *IEEE Access*, 9:140383–140396, 2021. ISSN 21693536. doi: 10.1109/ACCESS.2021.3119041.
- Xiaoyue Li, A. Sinem Uysal, and John M. Mulvey. Multi-period portfolio optimization using model predictive control. *European Journal of Operational Research*, 2022.
- Harry M. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
- J. Masuda and X. Meng. Portfolio optimization using a hybrid machine learning and genetic algorithm framework. *MIT EECS Thesis*, 2024.
- Yingting Miao and Qiang Zhang. Optimal investment and consumption strategies with transaction costs under crra utility. *arXiv preprint*, 2023.
- Johannes Muhle-Karbe, Max Reppen, and H. Mete Soner. A primer on portfolio choice with small transaction costs. *arXiv preprint*, 2016.
- Peter Nystrup, Stephen Boyd, Erik Lindström, and Henrik Madsen. Multi-period portfolio selection with drawdown control. *Annals of Operations Research*, 2019.
- Cheng Peng, Young Shin Kim, and Stefan Mittnik. Portfolio optimization on multivariate regime switching garch model. *arXiv preprint*, 2020.
- Piotr Pomorski and Denise Gorse. Multi-period portfolio optimisation using a regime-switching predictive framework. *arXiv preprint*, 2023.
- William F. Sharpe. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966.
- Lars Tvede. *Quantitative Financial Risk Management*. Wiley Finance, 2012.