

# Q-learning e gerenciamento de portfólios no mercado de criptomoedas

Pedro Albuquerque and Ludmilla Mattos\*


*FACE - Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas  
Públicas - Universidade de Brasília*

E-mail:



## Resumo

## Introdução

(só um rascunho)

Gerenciar um portfólio financeiro consiste na redistribuição constante de um montante em diferentes instrumentos financeiros. Em geral, a alocação de ativos é formalizada como  o Problema Markoviano de decisão (MDP) e pode ser otimizada com a aplicação das técnicas de Reinforcement Learning.

Em sua forma mais simples, O MDP é descrito por um conjunto finito de estados e ações e um conjunto de probabilidades de transição. Cada estado possui uma *Value-function*  $V$ , que indica o o quão vantajoso é para o agente estar naquele estado e seguir uma *policy*  $\pi$ , que é uma regra para decidir qual ação tomar. O objetivo é encontrar a *policy* que maximiza

  $V$  para todos os estados. O valor ótimo de  $V$  é calculado utilizando a equação de Bellman  segundo duas abordagens principais:

- A programação dinâmica uma solução padrão, que assume que as probabilidades de transição e os valores esperados são conhecidos *model-based*.
- O Q-learning é um método de aprendizado por reforço que não exige um modelo do sistema conhecido (*model-free*). Ele busca a solução ótima através de amostras de estados e retornos enquanto interage com o sistema.


## Referencial teórico

O *Q-learning* (QL) é uma técnica de aprendizado por reforço, primeiramente introduzida em 1998 por Watkins<sup>1</sup> em sua tese de PHD. Em 1992 a convergência do método foi provada por Watkins e Dayan<sup>2</sup>.


A história do aprendizado por reforço e sua aplicação na área de finanças é recente. Em 1996, Neuneier<sup>3</sup> foi um dos pioneiros em lançar mão do *Q-learning* para alocação ótima de ativos. Utilizando um modelo simplificado do mercado, sob suposições de que o agente não possui aversão ao risco, demonstra a aplicabilidade do aprendizado por reforço à um problema com espaço de estados de alta dimensionalidade. Mostrou que a estratégia de alocação utilizando *Q-learning* é no mínimo equivalente à programação dinâmica. Os métodos foram testados com a tarefa de investir capital líquido no mercado de ações alemão e redes neurais são utilizadas para aproximação da função de valores.

Em 1998 Moody et al.<sup>4</sup> utilizou o *Q-learning* para comparar recompensas futuras e imediatas. (este artigo é importante, mas ainda não consegui baixar o pdf na íntegra) No mesmo ano, Moody et al.<sup>5</sup> inovou ao formular uma técnica de aprendizado por reforço baseado em redes neurais, utilizando *back propagation* para atualizar diretamente a função *Q*.

Já em 2000, Xio Gao<sup>6</sup> propõe um método de gerenciamento de portfólio utilizando QL e Sharpe Ratio. Alguns estudos, ao modo de<sup>3</sup>, utilizam QL para maximizar o lucro, porém assumem que o investidor não possui aversão ao risco. Aqui os autores propõem uma forma

de mensurar o risco utilizando Sharpe Ratio, uma medida de lucro relativo ajustado ao risco, dada pela razão do retorno médio pelo desvio padrão do retorno. 

Os modelos em finanças podem ser reduzidos à formas simplificadas que possibilitam estudos menos (complexos), entretanto, a negociação de ativos é uma tarefa complexa quando todas as nuances da realidade do mercado de ações são consideradas. Assim, novas abordagens de aprendizado por reforço foram sendo desenvolvidas para possibilitar a aplicação no universo das transações financeiras. Enquanto os primeiros estudos baseados em *Q-Learning* estudam apenas um agente .

Lee e Jangmin<sup>7</sup> inovaram trazendo uma aplicação de *Deep Q-Learning* para múltiplos agentes cooperativos, que se comunicam compartilhando *training episodes* e *policies* aprendidas. Os parâmetros de transação são otimizados com QL, utilizando redes neurais para aproximar a função Q. O modelo visa maximizar os ganhos dos investimentos considerando não apenas a tendência global mas também o movimento diário dos preços dos ativos. Isso é feito pelos múltiplos agentes, cada um com seu objetivo: 

- Buy signal agent: faz predição, estimando informação de curto e longo prazo para produzir sinal de compra.
- Buy order agent: não faz predição, apenas determina um preço de compra.
- Sell signal agent: produz sinal de venda.
- Sell order agent: determina preço de venda.

Os estados são definidos como matrizes que representam os indicadores de longo e curto prazo. Quanto as ações, os agentes podem tomar dois tipos de ação: o Buy agent pode assumir BUY ou NOT-BUY, enquanto o sell agent pode HOLD ou SELL. A recompensa é definida como a taxa de lucro, considerando os custos da transação.

O método de QL é comparado como um método de aprendizado supervisionado de RN sendo ambos aplicados no mercado de ações Coreano. O experimento realizado com QL se

mostrou superior em retorno e gerenciamento de risco, mesmo em um momento de queda brusca do mercado, enquanto o método RN perde severamente, o QL apresenta perdas pequenas. [resumir esse blah infinito...]

:

Em 2012, Necchi<sup>8</sup> Desenvolve um algoritmo baseado em RL, Policy Gradient Methods

<sup>9</sup> Esse me parece ser o estado da arte. Artigo muito complexo, utilizando convolutional neural networks (e várias coisas que eu não entendi) no mercado de criptomoedas (sim, CRIPTOMOEDAS!). Abstract:Financial portfolio management is the process of constant redistribution of a fund into different financial products. This paper presents a financial-model-free Reinforcement Learning framework to provide a deep machine learning solution to the portfolio management problem. The framework consists of the Ensemble of Identical Independent Evaluators (EIIE) topology, a Portfolio-Vector Memory (PVM), an Online Stochastic Batch Learning (OSBL) scheme, and a fully exploiting and explicit reward function. This framework is realized in three instants in this work with a Convolutional Neural Network (CNN), a basic Recurrent Neural Network (RNN), and a Long Short-Term Memory (LSTM). They are, along with a number of recently reviewed or published portfolio-selection strategies, examined in three back-test experiments with a trading period of 30 minutes in a cryptocurrency market. Cryptocurrencies are electronic and decentralized alternatives to government-issued money, with Bitcoin as the best-known example of a cryptocurrency. All three instances of the framework monopolize the top three positions in all experiments, outdistancing other compared trading algorithms. Although with a high commission rate of 0.25

## Referências

- (1) Watkins, C. J. C. H. Learning from delayed rewards. Ph.D. thesis, King's College, Cambridge, 1989.

- (2) Watkins, C. J.; Dayan, P. *Machine learning* **1992**, 8, 279–292.
- (3) Neuneier, R. Optimal asset allocation using adaptive dynamic programming. *Advances in Neural Information Processing Systems*. 1996; pp 952–958.
- (4) Moody, J.; Wu, L.; Liao, Y.; Saffell, M. *Journal of Forecasting* **1998**, 17, 441–470.
- (5) Moody, J. E.; Saffell, M.; Liao, Y.; Wu, L. Reinforcement Learning for Trading Systems and Portfolios. *KDD*. 1998; pp 279–283.
- (6) Gao, X.; Chan, L. An algorithm for trading and portfolio management using Q-learning and sharpe ratio maximization. *Proceedings of the international conference on neural information processing*. 2000; pp 832–837.
- (7) Lee, J. W.; Jangmin, O. A multi-agent Q-learning framework for optimizing stock trading systems. *International Conference on Database and Expert Systems Applications*. 2002; pp 153–162.
- (8) Necchi, P. G.
- (9) Jiang, Z.; Xu, D.; Liang, J. *arXiv preprint arXiv:1706.10059* **2017**,