Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Continuous control with Stacked Deep Dynamic Recurrent Reinforcement Learning for portfolio optimization

Amine Mohamed Aboussalah, Chi-Guhn Lee*

Department of Mechanical and Industrial Engineering, University of Toronto, ON M5S 3G8, Canada

ARTICLE INFO

Article history:
Received 29 January 2019
Revised 22 July 2019
Accepted 19 August 2019
Available online 20 August 2019

Keywords:
Reinforcement learning
Policy gradient
Deep learning
Sequential model-based optimization
Financial time series
Portfolio management
Trading systems

ABSTRACT

Recurrent reinforcement learning (RRL) techniques have been used to optimize asset trading systems and have achieved outstanding results. However, the majority of the previous work has been dedicated to systems with discrete action spaces. To address the challenge of continuous action and multi-dimensional state spaces, we propose the so called Stacked Deep Dynamic Recurrent Reinforcement Learning (SDDRRL) architecture to construct a real-time optimal portfolio. The algorithm captures the up-to-date market conditions and rebalances the portfolio accordingly. Under this general vision, Sharpe ratio, which is one of the most widely accepted measures of risk-adjusted returns, has been used as a performance metric. Additionally, the performance of most machine learning algorithms highly depends on their hyperparameter settings. Therefore, we equipped SDDRRL with the ability to find the best possible architecture topology using an automated Gaussian Process ($\mathcal{GP}$) with Expected Improvement ($\mathcal{EI}$) as an acquisition function. This allows us to select the best architectures that maximizes the total return while respecting the cardinality constraints. Finally, our system was trained and tested in an online manner for 20 successive rounds with data for ten selected stocks from different sectors of the S&P 500 from January 1st, 2013 to July 31st, 2017. The experiments reveal that the proposed SDDRRL achieves superior performance compared to three benchmarks: the rolling horizon Mean-Variance Optimization (MVO) model, the rolling horizon risk parity model, and the uniform buy-and-hold (UBAH) index.

1. Introduction

The development of intelligent trading agents has attracted the attention of investors as it provides an alternative way to trade known as automated data-driven investment, which is distinct from traditional trading strategies developed based on microeconomic theories. The intelligent agents are trained by using historical data and a variety of Machine Learning (ML) techniques have been applied to execute the training process. Examples include Reinforcement Learning (RL) approaches that have been developed to solve Markov decision problems. RL algorithms can be classified mainly into two categories: actor-based (sometimes called direct reinforcement or policy gradient/policy search methods) (Baxter & Bartlett, 2001; Moody & Wu, 1997; Moody, Wu, Liao, & Saffell, 1998; Ng & Jordan, 2000; Williams, 1992) where the actions are learned directly, and critic-based (also known as value-function-based methods) where we directly estimate the value functions. The choice of a particular method depends upon the nature of the problem being addressed. One of the direct reinforcement techniques is called recurrent reinforcement learning (RRL) and it is presented as a methodology to solve stochastic control problems in finance (Moody & Wu, 1997). RRL has advantages of finding the best investment policy which maximizes certain utility functions without resorting to predicting price fluctuations and it is often incorporated with a neural network to determine the relationship (mapping) between historical data and investment decision making strategies. It produces a simple and elegant representation of the underlying stochastic control problem while avoiding Bellman's curse of dimensionality.

In the past, there have been several attempts to use a value-based reinforcement learning approach in the financial industry: a TD(λ) approach has been applied in finance (Van Roy, 1999) and Neuneier (1996) applied Q-Learning to optimize asset allocation decisions. However, such value-function methods are less-than-ideal for online trading due to their inherently delayed feedback (Moody & Saffell, 2001) and also because they imply having a discrete action space. Moreover, the Q-learning approach turns out to be more unstable compared to the RRL approach when presented with noisy data (Moody & Saffell, 2001). In fact, Q-learning algorithm is more sensitive to the value function

* Corresponding author.
E-mail addresses: amine.aboussalah@mail.utoronto.ca (A.M. Aboussalah), cglee@mie.utoronto.ca (C.-G. Lee).

https://doi.org/10.1016/j.eswa.2019.112891