Emotional Reinforcement Learning for Portfolio Selection

Ali Abbaspour, Caro Lucas

Control and Intelligent Processing Center of Excellence,
Electrical and Computer Eng. Dept., University of Tehran, Tehran, Iran
and
School of Intelligent Systems, Institute for Studies on Theoretical Physics and Mathematics,
Tehran, Iran

Abstract:

Reinforcement learning algorithm has been successfully used in prediction and decision making [5,11]. The main contribution of this paper is to provide decision making using reinforcement learning approach to allocate resources optimally in stochastic conditions in a well known example; in the portfolio selection. The modern theories of portfolio selection consider some presumptions. But if they don't hold, these methods are no longer efficient. So these days, some papers have been written by using the artificial intelligent methods. In this paper, appropriate emotional reinforcement signal is composed for portfolio selection. For this purpose, the reward signal is taken as the output of a linguistic fuzzy inference system with the return of portfolio and the risk of portfolio as inputs, then we implement the Q-learning neural network and we train this network with the proposed reward signal.

Keywords: Q-learning, Neural networks, Emotional learning, Portfolio selection

I. Introduction

Reinforcement learning from machine learning point of view is a collection of algorithms that can be used to optimize a decision making task. Initially it was presented as "trial and error" method to improve the interaction with dynamical systems [6]. Later it has been established that it can be regarded as a heuristic kind of Dynamic Programming (DP). The objective is to find a policy, a function that maps the states of the systems to control the actions, that optimizes a performance criterion. When the future performance is known for each state, the given the present state it is possible to select the "preferred" next state. For this a model of the system is required. Because such a model of the system is not always available, model free RL techniques were developed. Q-learning is model free RL. The idea is that the sum of future reinforcement can be approximated as a function of the state and the action. This function is called the Q-function and it has a value for each state and action combination [8,10].

In this paper we introduce a Q-learning neural network for portfolio selection by considering the multi-objective such increasing the return and decreasing the risk of stock market. The critic has a big role in this method and criticizes the overall performance of the decision maker and then, it makes the emotional reinforcement signal.

The paper consists of six sections. In section two, we briefly discuss the preliminary of portfolio selection. The main aspects of Q-learning methods are described in the third section. The neural implementation of Q-learning methods is explained in section four. The results of applying the proposed emotional multi-Objective Q-learning method to allocate investments in the portfolio between individual common stock are reported and analyzed in section five. Finally, the last section includes some concluding remarks.

II. Portfolio selection

In this section we introduce a formal model of the portfolio selection problem in a stochastic stock market. Portfolios are an effective way of increasing returns while decreasing risk when investing in the stock market. For this reason there has been considerable attention to portfolio selection strategies in the financial. The first theory of modern portfolio backs to the paper by "Harry Markowitz" in the title of "Portfolio Selection" [2]. In his paper and the other modern portfolio theories [1,3,4], they assumed that the probability distribution function of the shares is normal and time invariant. Beyond these presumptions these methods are the efficient way for investment. But if they do not hold, these methods are no longer efficient. but these days, some papers has been written by using the artificial

intelligent methods such neural networks or expert systems for allocating investments between types of securities, such as bonds, stocks, venture capitals, and real estates [12,13,14].

A portfolio in market of N stocks in a single investment period is represented as a vector $W = (w_1,\ldots,w_N)$ where $w_i \geq 0$ and $\sum_{i=1}^{N} w_i = 1$. A fraction w_i of wealth is invested in stock i at the start of period. The total change in wealth over the period depends on the change in price of the stocks held in the portfolio. Given a vector of "price relatives", $X = (x_1,\ldots,x_N)$ where x_i is ratio of closing price to opening price over the period for stock i, then the wealth of an agent with portfolio W increases (or decreases) by a factor of $W \cdot X = \sum_{i=1}^{N} w_i x_i$. This is the simple gross return from portfolio W. The return on investment, R_s from portfolio selection is $R_s = W \cdot X$. After calculating the return, the wealth of investor is easily calculable with these formulas:

$$\begin{aligned} I_1 &= 1 \\ I_t &= I_{t-1}(1 + R_s^{t-1}) \end{aligned} \qquad (1)$$

In this formula, I_t is the wealth of investor in period t and R_s^{t-1} is the return of portfolio in the period $t-1$. The optimal portfolio strategy will depend on the risk preference of the investor. Typically investors are risk averse. The model for risk can be considered as variance, semi-variance, mean absolute deviation and etc. A good investment strategy makes a tradeoff between the expected final-period wealth and the risk of portfolio.
Respectively, the variance and semi variance of the output of ANFIS can be calculated:

$$P_1 = \mathrm{var}(\mathrm{Re}\,t_p) = \frac{1}{N_s}\sum_{i=1}^{N_s}(\mathrm{Re}\,t_i - \frac{1}{N_s}\sum_{k=1}^{N_s}\mathrm{Re}\,t_k)^2 \qquad (2)$$

and

$$P_2 = \frac{1}{N_s}\sum_{i\in Q}(\mathrm{Re}\,t_i - \frac{1}{N_s}\sum_{k=1}^{N_s}\mathrm{Re}\,t_k)^2$$
$$Q = \left\{ i \mid \mathrm{Re}\,t_i > \sum_{k=1}^{N_s}\frac{\mathrm{Re}\,t_k}{N_s} \right\} \qquad (3)$$

where N_s is the number of data.

III. Q-learning

The first version of Q-learning is based on the temporal difference of order 0, $TD(0)$, while only considering the following step. The agent observes the present state, x_t and executes an action, a_t, according to the evaluation of return that it makes at this stage. It updates its evaluation of the value of the action while taking un account, a) the immediate reinforcement, r_t, b) the estimated value of the new state, $V_t(x_{t+1})$, that is defined by:

$$V_t(x_{t+1}) = \max_{b\in A_{t+1}} Q(x_{t+1}, b) \qquad (4)$$

The update corresponds to the equation:

$$Q(x_t, a_t) \leftarrow Q(x_t, a_t) + \beta\{r_t + \gamma V_t(x_{t+1}) - Q(x_t, a_t)\} \qquad (5)$$

β is a learning rate such that $\beta \to 0$ as $t \to \infty$. This equation can be written:

$$Q(x_t, a_t) \leftarrow (1-\beta)Q(x_t, a_t) + \beta\{r_t + \gamma V_t(x_{t+1})\} \qquad (6)$$

Q-learning, in addition to its simplicity, presents several characteristics. The evaluations of Q, the Q-values, are independent of the policy followed by the agent. This one can follow any policy, while continuing to construct correct evaluations of the value of actions. Q-values are exploitable a long time before the formal convergence that can be sometimes very slow. Lastly, there are proofs of convergence toward the optimal policy.
After convergence of Q-Learning, the optimal policy is performed while choosing the action that, to every state, maximizes the Q-function:

$$a = \arg\max_{b\in A_x} Q^*(x, b) \qquad (7)$$

This policy is called *greedy*. However, in the beginning of the training, the values of $Q(x,a)$ are not meaningful: applying the greedy policy too quickly often drive to local minima. To get an useful evaluation of Q, it is necessary to sweep and to evaluate the set of possible actions, for all states: it is what one calls the phase of *exploration* in opposition to the *exploitation* one, at the end of training. The Exploration/Exploitation dilemma can be expressed by: at each state, the agent must choose between

- an action for which the expected reward is supposed to be good quality
- an action whose quality can be, to this precise instant of the choice, less good but whose application could drive it in promising but not explored zones.

IV. Q-learning Neural Networks Implementations

A neural implementation seems to offer many advantages: quality of the generalization and limited memory requirement for storing the knowledge. The memorization function uses the weight set of the neural network. The memory size required by the

system to store the knowledge is defined, a priori, by the number of connections of the network. It is independent of the number of explored situation-action pairs. The ideal neural implementation will provide, in a given situation, the best action to undertake and its associated Q value. For discrete state and action spaces of reduced size, Q-values are stocked in a look-up table, $q_{ij} = Q(x_i, a_j)$ for $x_i \in X$ and $a_i \in A$. If the size of X or of A increases considerably, or if X is continuous, it becomes impossible to visit all the states and to test all actions in one reasonable time. For this reason, capacities of interpolation of Artificial Neural Networks (ANN) have been exploited [17, 25] in the case of a continuous state space and a discrete action space. Let J be the number of actions in A. A neural implementation is J ANN with one output: one output by action, in accordance with the architecture of Figure 1.

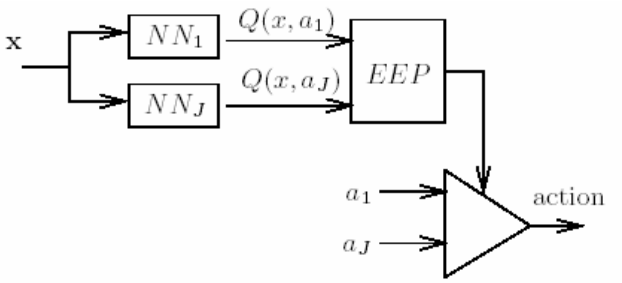

Figure 1. Neural Systems with J actions

This architecture, proposed initially by Lin [7], is more effective. The process is the next one for Q(0):

1. Presentation of the state vector **x**. Every ANN calculates an evaluation of $Q(x, a_j)$, $j = 1$ to J.

2. A module of exploration/exploitation, EEP, chooses the action to apply, a_{j^*}.

3. The new state is **y** and the immediate reinforcement is r.

4. Presenting **y** as a new input of the ANNs, its value is calculated by:

$$V(y) = \max_{1 \le j \le J} \mathrm{ANN}_j(y) \qquad (8)$$

where ANN_j is the output of the ANN number j

5. The new evaluation of $Q(x, a_j)$ becomes $r + V(y)$: the deference between the elder and the new value is considered as the error committed by ANN_{j^*}. This error is used to modify the weights.

V. Experimental result

In this part of article, for the practical experiment, the data of some stocks related to the price of some blue chips in the S&P index 500 are chosen for this benchmarking portfolio selection. The critic has a big role in reinforcement learning and criticizes the performance of Q-learning approach. With the daily returns of stocks, it evaluates the return and the risk of portfolio and the other parameters that it needs. Then, it makes the reward signal. The critic has an interaction with learning element and knowledge base. So, it produces the appropriate action to the environment (Stock market) that consists of the weights of portfolio. For this paper, we consider two methods for producing the reward signal. First, we only regard the return of portfolio to evaluate the reward signal for Q-learning as follows:

$$r_t = \begin{cases} 1 & \text{if the wealth of investor increases} \\ -1 & \text{if the wealth of investor decreases} \\ 0 & \text{if the wealth of investor doesn't change} \end{cases}$$

Second, as it is said before, the risk has a big role in portfolio selection to obtain the optimal portfolio strategy so we must consider the criterion of the risk beside the return of portfolio as the important objectives to make the reward. For this purpose, the reward signal is taken as the output of a linguistic fuzzy inference system with the return of portfolio and the risk of portfolio as inputs. One time we consider the variance of portfolio as a criterion for the risk. Next we replace the semi-variance of portfolio with the variance as the other criterion for the risk. Three Gaussian membership functions are used for the inputs. Figure 2 shows the surface generated by the fuzzy rules of emotional critic. Then it makes the reward signal for Q-learning method.

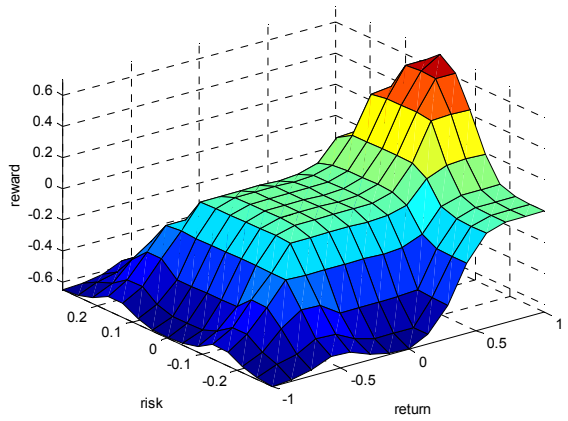

Figure 2: The output surface of a linguistic fuzzy inference system for producing reward signal

The results of applying the proposed emotional multi-Objective Q-learning method to allocate

investments in the portfolio between individual common stock are shown in figure 3. This result is obtained when we only use the increase of return of portfolio as a reward for Q-learning.

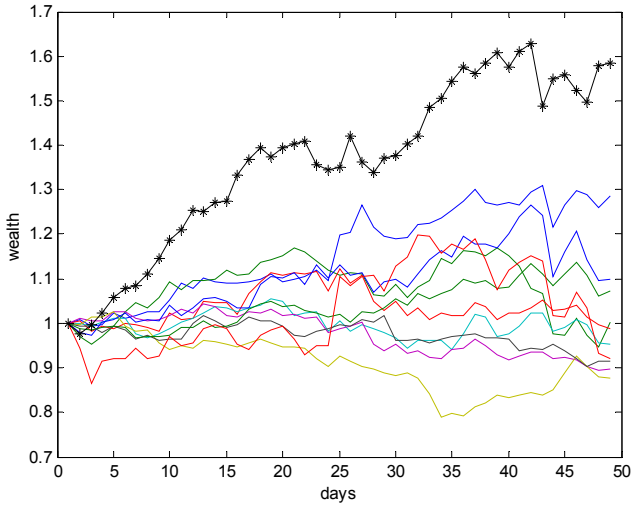

Figure 4. The results of applying the proposed Q-learning when we only use the return of portfolio as the reward

The comparison between three situations is depicted in figure 4. First, when we only use the return of portfolio. Next, when we apply the criterion of risk beside the return of portfolio is depicted in figure 4. In this figure, we separately consider the semi-variance and variance as the criteria for the risk of portfolio.

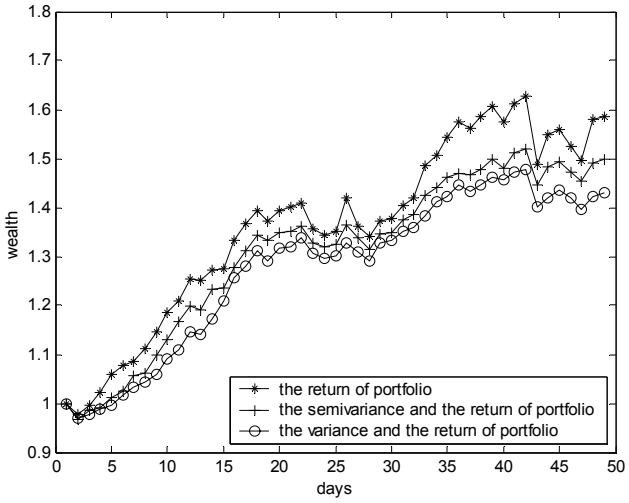

Figure 4. The results of applying the proposed emotional multi-Objective Q-learning with consideration of different criterion for the risk

The inability to utilize the device results in slower wealth accumulation though the portfolio volatility has improved by the extra risk aversion forced by that fact. Secondly, we argue that price fluctuation is not always bad. In fact, the investor would avoid unexpected occasional high upturns. It is the price decrease that should worry the investor. Therefore, we propose semi-variance as a superior criterion for portfolio risk.

VI. Conclusion

Emotional Q-learning algorithm has been used as a method to improve the portfolio selection toward several goals and requirements. We implement a neural Q-learning and its training is based on the method that describe in this article. The definition of reinforcement signal is an important aid in this learning algorithm, which provides high degrees of freedom. The experimental results confirm this argument as the consideration of semi-variance has resulted improvement in the performance of the portfolio. In fact, the investor would avoid unexpected fluctuations and considering the semi-variance as a criterion for the risk causes less fluctuations.

References:

[1] A.R. Havgen, *Modern Investment theory*, (prentice Hall. Inc, 1997)

[2] H. Markowitz, portfolio selection, *Journal of finance,* 1952

[3] W.F. Sharpe, "A simplified model of portfolio", *Management science,* 1963.

[4] W.F. Sharpe, "capital Asset prices : A theory of Market Equilibrium under Conditions of Risk". *Journal of Finance,* 1964.

[5] Touzet P., "Neural reinforcement learning for behavior synthesis", *Proc. of CESA'96, IMACS Multi-conference*, Lille, July 1996.

[6] Sutton R.S., Barto A.G., *Introduction to reinforcement learning*, MIT Press/Bradford Books, Cambridge, MA, 1998.

[7] Lin L-J., Self-improvement based on reinforcement learning, planning and teaching", *Proc. of* 8th *Workshop on Machine Learning ML'91*, 1991.

[8] Watkins C., Dayan P., "Q-learning", *Machine Learning*, 8, p. 279-292, 1992.

[9] Sutton R.S., \Learning to predict by the method of temporal difference", *Machine Learning*, 3, p. 9-44, 1989.

[10] Glorennec P.Y., Jou_e L., "Fuzzy Q-Learning",*Proc. of FUZZ-IEEE'97*, Barcelona, July 1997.

[11] Barto A., Sutton R., Watkins C., "Learning and sequential decision making", in *Learning and Computational Neuroscience*, MIT Press,Cambridge, 1990.

[12] J. Lee, H. SOO KIM, "*Intelligent Stock Portfolio Management System*", Expert Systems,1989

[13] Trippi R., Turban E., "*Investment Management Decision Support and Expert system*" (Boyd & Fraser publishing company, 1990)

[14] white, HJ, "*Economic prediction Using Neural Networks : the case of IBM Daily stock patterns*" proceedings of the IEEE International conferences on Neural Networks, JULY, 1988.