Chapter 10
Model-Based Multi-Objective Reinforcement Learning by a Reward Occurrence Probability Vector

Tomohiro Yamaguchi
Nara College, National Institute of Technology (KOSEN), Japan

Shota Nagahama
Nara College, National Institute of Technology (KOSEN), Japan

Yoshihiro Ichikawa
Nara College, National Institute of Technology (KOSEN), Japan

Yoshimichi Honma
Nara College, National Institute of Technology (KOSEN), Japan

Keiki Takadama
The University of Electro-Communications, Japan

ABSTRACT

This chapter describes solving multi-objective reinforcement learning (MORL) problems where there are multiple conflicting objectives with unknown weights. Previous model-free MORL methods take large number of calculations to collect a Pareto optimal set for each V/Q-value vector. In contrast, model-based MORL can reduce such a calculation cost than model-free MORLs. However, previous model-based MORL method is for only deterministic environments. To solve them, this chapter proposes a novel model-based MORL method by a reward occurrence probability (ROP) vector with unknown weights. The experimental results are reported under

DOI: 10.4018/978-1-7998-1382-8.ch010

the stochastic learning environments with up to 10 states, 3 actions, and 3 reward rules. The experimental results show that the proposed method collects all Pareto optimal policies, and it took about 214 seconds (10 states, 3 actions, 3 rewards) for total learning time. In future research directions, the ways to speed up methods and how to use non-optimal policies are discussed.

INTRODUCTION

Reinforcement learning (RL) is a popular algorithm for automatically solving sequential decision problems such as robot behavior learning and most of them are focused on single-objective settings to decide a single solution. A single objective RL can solve a simple learning task under a simple situation. However, in real world robotics, a robot often faces that the optimal condition on its own objective changes such as an automated driving car in a public road where many human driving cars move. So the real world learner has to treat multi-objective which may conflict by subsumption architecture (Tajmajer 2017)or the weights of these objectives may depend on the situations around the learner. Therefore, it is important to study multi-objective optimization problems in both research fields for robotics and reinforcement learning.

In multi-objective reinforcement learning (MORL), the reward function emits a reward vector instead of a scalar reward. A scalarization function with a vector of n weights (weight vector) is a commonly used to decide a single solution. The simple scalarization function is linear scalarization such as weighted sum. The main problem of previous MORL methods is a huge learning cost required to collect all Pareto optimal policies. Hence, it is hard to learn the high dimensional Pareto optimal policies. To solve this, this chapter proposes the novel model-based MORL method by reward occurrence probability (ROP) with unknown weights. There are two main features. The first feature is that the average reward of a policy is defined by inner product of the ROP vector and the weight vector. The second feature is that it learns ROP in each policy instead of Q-values. Pareto optimal deterministic policies directly form the vertices of a convex hull in the ROP vector space. Therefore, Pareto optimal policies are calculated independently with weights and just once. The experimental results show that the authors' proposed method collected all Pareto optimal policies under three dimensional stochastic environments, and it takes a small computation time though previous MORL methods learn at most two or three dimensions deterministic environments.