A New Hybrid Model for Associative Reinforcement Learning

H. Montazeri
Soft Computing Laboratory
Computer Engineering and IT Department
Amirkabir University of Technology
Tehran, Iran
montazeri@autac.ir

M. R. Meybodi
Soft Computing Laboratory
Computer Engineering and IT Department
Amirkabir University of Technology
Tehran, Iran
mmeybodi@autac.ir

Abstract—**In this paper, a new model, addressing the Associative Reinforcement Learning (ARL) problem, based on learning automata and self organizing map is proposed. The model consists of two layers. The First layer comprised of a SOM which is utilized to quantize the state (context) space and the second layer contains of a team of learning automata which is used to select an optimal action in each state of the environment. First layer is mapped to the second layer via an associative function. In other words, each learning automaton is in correspondence with only one neuron of the self organizing map. In order to show the performance of the proposed method, it has been applied successfully to classification applications on Iris, Ecoli, and Yeast data sets, as examples of ARL task. The results of experiments show that the proposed method is reached the accuracy near to or even higher than the highest reported accuracy. The results obtained for Ecoli and Yeast data sets indicate that the method is able to classify in relatively high dimensional context space and high number of classes.**

I. INTRODUCTION

Associative reinforcement learning (ARL) task defined originally by Barto and Anandan [1] is one that requires the learning element to establish a connection between input and output. ARL tasks involve the following interaction between the environment and the learning system. At time step n, the environment provides the learning system with some context vector $X(n)$ selected from a set of vectors $X \subseteq \Re^n$, where $\Re$ is set of real numbers. Based on this input, the learning system selects an action $\alpha(n)$ among its action set. The environment evaluates the action $\alpha(n)$ in the context of the input $X(n)$ and sends to the learning system a real-valued evaluation signal $\beta(n+1) \in [0,1]$ at time step *n+1*, with $\beta(n+1) = 0$ denoting the maximum evaluation.

Many works have been done on ARL, from which only a few are mentioned here. In original definition of ARL tasks, Barto and Anandan [1] only considered learning tasks for which the evaluation is a binary-valued success/failure signal (i.e. $\beta(n) \in \{0,1\}$). Their work in such tasks led to the development of the associative reward-penalty algorithm (A_{R-P}).

Some solutions of ARL tasks utilize a set of parameters. In this case, the learning task is to find the optimal values of the parameters. The complementary reinforcement backpropagation algorithm (CRBP) [2], an approach based on neural network, consists of a feed-forward network mapping an encoding of the context vector to an encoding of the action. The action is determined probabilistically from the activation of the output units. A supervised training procedure is used to adapt the network as follows. If the evaluation of the selected action is success ($\beta = 1$), then the network is trained to use the action whenever this context vector is provided. If an action fails to generate reward, CRBP will endeavor to generate an action that is different from current choice. Another work take advantage of set of parameters is presented In [3]. In this algorithm, each action (or alternative) has an attribute vector associated with it, and for each the action it selects, it gets either a success or failure. The method incorporates a learning method like Widrow-Hoff rule and a probabilistic selection strategy to solve the ARL problem. Another method of this category is an automaton-based approach which will be described later.

Some methods consider ARL tasks from an analytical point of view. Streth shows that associative prediction problem, a version of ARL, can be reduced to cost-sensitive classification and then to standard classification [4]. In [5], confidence bounds are used to deal with situations which exhibit an exploitation-exploration trade-off. In [6], efficient algorithms to solve a restricted class of RL problem are proposed. These algorithms can learn efficiently action policies that can be expressed as propositional formula in k-DNF (disjunctive normal form). Abe et al. consider the problem of reinforcement learning with immediate rewards in a worst-case theoretical framework. This work provides bounds on the per-trial regret that go to zero as the number of trials approaches infinity [7]. In [8, 9], immediate reward reinforcement learning problem is reduced to a reward-weighted nonlinear regression problem, which greatly