Neural Contextual Bandits with UCB-based Exploration

Dongruo Zhou  
Department of Computer Science  
University of California, Los Angeles  
Los Angeles, CA 90095  
drzhou@cs.ucla.edu

Lihong Li  
Google Research  
lihong@google.com

Quanquan Gu  
Department of Computer Science  
University of California, Los Angeles  
Los Angeles, CA 90095  
qgu@cs.ucla.edu

Abstract

We study the stochastic contextual bandit problem, where the reward is generated from an unknown function with additive noise. No assumption is made about the reward function other than boundedness. We propose a new algorithm, NeuralUCB, which leverages the representation power of deep neural networks and uses a neural network-based random feature mapping to construct an upper confidence bound (UCB) of reward for efficient exploration. We prove that, under standard assumptions, NeuralUCB achieves $\tilde{O}(\sqrt{T})$ regret, where $T$ is the number of rounds. To the best of our knowledge, it is the first neural network-based contextual bandit algorithm with a near-optimal regret guarantee. We also show the algorithm is empirically competitive against representative baselines in a number of benchmarks.

1 Introduction

The stochastic contextual bandit problem has been extensively studied in machine learning [13, 8, 14]: at round $t \in \{1, 2, \ldots, T\}$, an agent is presented with a set of $K$ actions, each of which is associated with a $d$-dimensional feature vector. After choosing an action, the agent will receive a stochastic reward generated from some unknown distribution conditioned on the action’s feature vector. The goal of the agent is to maximize the expected cumulative rewards over $T$ rounds. Contextual bandit algorithms have been applied to many real-world applications, such as personalized recommendation, advertising and Web search.

The most studied model in the literature is linear contextual bandits [4, 2, 9, 13], which assume that the expected reward at each round is linear in the feature vector. While successful in both theory and practice [15, 8, 11], the linear-reward assumption it makes often fails to hold in practice, which motivates the study of nonlinear or nonparametric contextual bandits [10, 19, 6, 20]. However, they still require fairly restrictive assumptions on the reward function. For instance, Filippi et al. [10] makes a generalized linear model assumption on the reward, Bubeck et al. [6] require it to have a Lipschitz continuous property in a proper metric space, and Valko et al. [20] assume the reward function belongs to some Reproducing Kernel Hilbert Space (RKHS).
In order to overcome the above shortcomings, deep neural networks (DNNs) \[11\] have been introduced to learn the underlying reward function in contextual bandit problem, thanks to their strong representation power. We call these approaches collectively as neural contextual bandit algorithms. Given the fact that DNNs enable the agent to make use of nonlinear models with less domain knowledge, existing work \[17, 21\] study neural-linear bandits. That is, they use all but the last layers of a DNN as a feature map, which transforms contexts from the raw input space to a low-dimensional space, usually with a better representation and less frequent updates. Then they learn a linear exploration policy on top of the last hidden layer of the DNN with a more frequent update. These attempts have achieved great empirical success, but no regret guarantees are provided.

In this paper, we consider provably efficient neural contextual bandit algorithms. The new algorithm, NeuralUCB, uses a neural network to learn the unknown reward function, and follows a UCB strategy for exploration. At the core of the algorithm is the novel use of DNN-based random feature mappings to construct the UCB. Its regret analysis is built on recent advances on optimization and generalization of deep neural networks \[12, 8, 7\]. Crucially, the analysis makes no modeling assumptions about the reward function, other than that it be bounded. While the main focus of our paper is theoretical, we also show in a few benchmark problems the effectiveness of NeuralUCB, and demonstrate its benefits against several representative baselines.

Our main contributions are as follows:

- We propose a neural contextual bandit algorithm that can be regarded as an extension of existing (generalized) linear bandit algorithms \[1, 10, 15, 16\] to the case of arbitrary bounded reward functions.
- We prove that, under standard assumptions, our algorithm is able to achieve $\tilde{O}(d\sqrt{T})$ regret, where $\tilde{d}$ is the effective dimension of a neural tangent kernel matrix and $T$ is the number of rounds. The bound recovers the existing $O(d\sqrt{T})$ regret for linear contextual bandit as a special case \[11\], where $d$ is the dimension of context.
- We demonstrate empirically the effectiveness of the algorithm in both synthetic and benchmark problems.

References


