論 文 内 容 の 要 旨

博士論文題目

Temporal Difference Approach in Linearly Solvable Markov Decision Processes

(線形可解マルコフ決定過程における受動的ダイナミクスのモデリングと推定)

氏名 Mauricio A. P. Burdelis

The Reinforcement learning (RL) approach to Machine Learning is a technique to learn how to make decisions in order to achieve a desired goal. The model does not include the presence of a supervisor. The agent must learn by trial and error. This is done by taking actions and observing their consequences in the form of a reward (or cost) signal. Such problems are usually formalized as a Markov decision process (MDP). The mathematical framework of MDPs relies on the Bellman equation and is very general, but finding solutions can be inefficient because of the explosion of possible future states.

The framework of linearly solvable Markov decision processes (LMDP) greatly simplifies reinforcement learning. By attending specific conditions the Bellman equation can be made linear, and it becomes possible to obtain solutions more efficiently. However, it is necessary to previously know the passive dynamics of the system (i.e. the behavior of the system in the absence of controls) which is crucial in the model, but unknown in general.

A method to calculate such passive dynamics distribution (by performing continuous embedding of known traditional MDPs) exists, but requires the previous knowledge of all transition distributions and all immediate costs. Those are usually not known beforehand in temporal difference methods. Such methods require the agent to explore the environment and learn by trial-and-error.

Here we propose a method to estimate the passive dynamics and state costs of a given system. As a consequence, such system can then be modeled as an LMDP. The method can also be combined with a temporal difference algorithm

of the LMDP framework (called Z~learning). This enables the direct application of Z learning without the need for explicit knowledge of passive dynamics nor state costs beforehand. The only required knowledge about the passive dynamics distribution of the system is which states can and which cannot be visited starting from each state. And the only remaining limitation for the direct application to real problems (with symbolic actions) is the assumption that the agent can impose any desired transition distribution it wants. Such assumption is an important premise of the LMDPs framework.

During the application of the method, new constraints regarding the passive dynamics and state costs are successively incorporated in the model from observed information of immediate costs. The resulting algorithm properly estimates the desirability and optimal cost-to-go functions, as well as the passive dynamics and state costs, when solving the resulting constrained optimization problem. The convergence speed of the new algorithm is not significantly affected when compared to pure Z learning. This represents an important step for direct application of the framework of LMDPs framework in a real temporal difference approach.

氏　名	Mauricio Alexandre Parente Burdelis

（論文審査結果の要旨）

強化学習は計算量が多く学習時時間がかかるという従来の常識に対し，Todorovによって提案されたマルコフ決定過程に対する強化学習の枠組みは，強化学習が線形のベルマン方程式に帰着するという点で驚異的なものであった．Todorovは最適制御の双対問題を考えることでその計算量を削減できることを示し，また効率的なアルゴリズムを提案している．

しかし Todorov のアルゴリズムは，入力がない時のダイナミクス（Passive Dynamics，PD）に従って学習を行うか，あるいはPDが既知であることを仮定している．実際には，適当なダイナミクスに従って学習を行う方が学習が速く進む一方で，PDは未知である場合が多い．

本研究ではこの問題を解決するため，任意のダイナミクスに従って学習を進め，同時にPDを推定する方法を提案したものである．

本研究では，Todorov の強化学習の枠組みでは瞬時報酬が PD と実際のダイナミクスの KL ダイバージェンスと状態報酬の和で表されることに着目し，PD に関する拘束条件を導いた．この条件はPDの状態遷移確率の対数について線形となるため，線形方程式の解法を利用して効率よく解くことができる．さらに，この提案法がPDが既知の場合の学習法とほぼ同等の収束性能を持つことを計算機実験により示した．

以上をまとめると，本論文は，Todorov の線形可解マルコフ決定過程の枠組みにおいて，PD が未知であっても効率よく学習する方法を提案し，かつその有効性を計算機実験によって確認している．これは強化学習の枠組みの拡張に寄与するものであり，博士（工学）の学位に値するものと認められる．