Certified Reinforcement Learning with Logic Guidance:概要

09C18707 知能システム学コース 4 年 大浦稜平

2019/10/06

1　どんなもの？

LTL から GLDBA を生成し，MDP とプロダクトを取り，この GLDBA の監視下で強化学習を用いてエージェントの MDP 上での最適方策を求める．

2　先行研究と比べてどこがすごい？批判されている理論は何？

product MDP による強化学習自体は 2014 年に提唱されたが，本研究では DRA ではなく GLDBA を用いて状態数を小さくしている（一般に DRA は状態数が大きくなり，またその受理条件の構成ゆえに reward shaping も複雑になる）．DRA を用いた Sadigh らの方法は遷移確率の近似に依存してしまい，それゆえに最適方策を生成する過程の正確さが制限されると批判している．
↓
これと比較して，提案手法は最適方策の学習と同時に，明示的に MDP のダイナミクスを獲得する．

3　技術や手法の肝はどこ？どうやって有効だと検証した？

集合 A をすべての受理領域の和集合で定義し，初めて訪れた領域を A から除去していき，すべて削除されたらまた初期化することを繰り返す更新則 Acc，及びこれに基づく報酬関数設計．連続状態空間に対しても LTL を完全に満たすような最適方策の学習を行える．

グリッドワールド・パックマン・モンテズーマの逆襲（有限状態），火星探査（連続状態）で実験．
enumerate
グリッドワールド
safty を含んだ論理制約に対して良好な結果を示した．
パックマン
およそ 20000 episode で安定した方策を獲得した．従来の方法（勝利した場合に報酬を与えるやり方）では安定した方策を獲得できなかった．
モンテズーマの逆襲
10000 episode でゴールに到達
火星探査
LCNFQ を使用．Voronoi quantizer と FVI と比較して良好な結果を得た．

4　どういう文脈・理路を辿っている？

オートマトンに GLDBA を用いる．
↓
複数の受理条件を公平に回らせるために Acc を導入．

有限状態空間：
Q learning をベースにした学習則を採用している．
∵Th2 より最適方策が存在すれば価値ベースでの学習で最適方策を獲得できる（LTL を満たすが価値ベースで最適でない方策の存在を仮定すると報酬が有限回しか貰えないことに反する）
Def14 より，最適方策が存在しない場合でも最も良い方策を Q learning で獲得できる（∵ 最も多数の受理領域を回るほうがより多くの報酬を獲得できる）．また，Th3 より，LCQL で得た方策は LTL を満たす理論上の最大確率を得る（DP ベースの価値反復で得る方策と期待割引報酬に基づく LCQL の収束により得る方策が実質同じであることが，discount を (1) 係数として解釈した場合と (2)undiscount のエピソード長が指数分布に従っていると解釈した場合から言える）

連続状態空間：

1. voronoi cells に状態空間を分割する
   セル中心との距離やオートマトンの状態が初登場か否かで状態を細かく分割していきながら Q learning
2. Fitted Value Iteration の改良
   ベルマン作用素の中の報酬項をなくして，その代わりに各状態に対する状態価値を $\mathbb{A}$ に状態が入っているかどうかで初期化する．また，ベルマン作用素の積分計算は RBF を用いて近似する．
3. MDP の事変性（定期的なもの！）を再帰的なクリプケ構造 $\mathcal{K}$ で表現し，LDBA を事変オートマトンとして拡張し，MDP と同期させる．学習は事変的な部分にのみ着目する（Q(s,a) の値を各 $k \in \mathcal{K}$ でシェアする）．これにより学習を早める

5　対象となる問題において網羅性・整合性はある？

今後，マルチエージェント系・POMDP に適用していくらしい

6　議論はある？

7　考えられる課題は？

Acc の更新則は異なる経路による受理領域への到達を区別できない（本来区別されるべき）
連続状態空間にへの対処法の妥当性について．

8　次に読むべき論文は？