東京電機大学大学院

学位論文内容の要旨

報告番号	先端科学技術甲第168号	氏 名	神 谷 匠
論文題目	Softsatisficing：確率論的満足化方策		

21世紀に入ってからの急速な変化として、深層学習技術の進展に支えられて、機械学習が社会の広汎な範囲で実用に供されるようになってきた。機械学習の中でも強化学習分野は、ある特定の環境において自律的な試行錯誤を通じて適切な行動を獲得するもので、深層学習との組み合わせである深層強化学習技術の進展に伴い、ロボットの身体制御やプランニングなども射程に入っており、発展の暁には産業的な価値も非常に高い。実際、現在までに強化学習技術により、数々のビデオゲームのみならず囲碁や将棋を含むボードゲームなどにおいても人間のチャンピオンを超えるパフォーマンスを示している。

今後強化学習がゲームのような仮想空間における限定された問題でなく現実の問題に応用できるほど成熟するためには、数々の障害がある。その中でも最も根本的な障害としては、問題が現実に近づき複雑化するほど（つまり可能な状態数と行動種類が増えるほど）深刻化する、試行錯誤の必要回数の爆発の問題がある。そこで本論文は、動物や人間、そして動物の群れや社会的組織などの意志決定傾向として定性的に観察されている「満足化satisficing」の新たなアルゴリズムを提案し、深層強化学習への応用の可能性を切り拓くことを目的とした。満足化とは、適切な行動を試行錯誤を通じて探索する際に、ある基準を上回る行動が見つかり次第探索を打ち切るという行動選択の方策（ポリシー）であり、会社組織（黒字になるまで事業内容を変更）、人間（合格点を取るまで勉強）、動物（生存に必要なカロリーを得るまで採餌）に実際に観察される行動傾向のモデルである。論文は６章構成である。

第１章は序論であり、近年の急速な深層強化学習技術の発展を概観したのち、現実的タスクが提示する問題を論じ、それへの解答としての限定合理性の理論と満足化の概念を導入している。また、1950年代に提案された限定合理性の理論が近年どのようにリバイバルしつつあるかを概観している。満足化を用いた国内外の先行研究を整理し、背景を明確にした上で、研究の目的を述べている。

第２章は第１章で簡単に触れた限定合理性の理論と満足化の概念を、従来の強化学習分野の最適化の概念と対比し、より明確にしている。満足化方策の実装の問題点の一つである探索の仕方については、これまで具体的な議論が少なく、従来の研究では単純な一様乱数を用いて行動をランダムに選択してきた。しかし例えばアスリートであれば、自己ベストの更新を目標とするならば、自己ベストを出した際のフォームを微調整するだろう。他方で世界記録を目指して大きく自己記録を上回る必要があれば、フォームの大幅な変更や肉体改造も視野に入れるだろう。このような探索・試行錯誤方式の動的な調整の必要性という先行研究になかった点が議論されている。

第３章は本論文において具体的な強化学習タスクとして扱うＫ本腕ベルヌーイバンディット問題（以下バンディット問題）の問題設定とパフォーマンス指標を述べ、またバンディット問題のベンチマークアルゴリズムとして代表的な Upper Confidence Bounds (UCB)アルゴリズムとThompson Sampling (TS)アルゴリズムを導入している。

第４章は、先行研究で導入された、環境の不確実性（リスク）を考慮した満足化アルゴリズムである Risk-sensitive Satisficing (RS) を分析し、UCBやTSと比較可能な形式(RSβ) を導いている。バンディット問題におけるRSの基本的なパフォーマンスを示し、またUCBやTSとの挙動の違いを具体的な数値実験で実証している。また、RSの振る舞いと

して「非満足時損失均衡」という概念を提案し、RSの探索の仕方が、ある種の均衡を目指す形で行われていることを示している。

第５章は、これまでの準備と分析に基づき、RSと同等の挙動をもたらす確率論的な方策である Softsatisficing アルゴリズムを導いている。そのためにまず、強化学習の行動選択に用いられる Softmax 関数を導入し、非満足時損失均衡を仮定することでRSの挙動を確率分布に変形し、Softsatisficing 関数を導出している。また、Softmax において行動選択のランダム性を制御するハイパーパラメータである逆温度に対応する値が Softsatisficing にも存在し、それが自律的に制御されることを示している。さらには、数値実験でRSとSoftsatisficingが同等の探索傾向とパフォーマンスを示すことを確認し、SoftmaxとSoftsatisficingの振る舞いの違いをも詳細に分析している。最後に、Softsatisficing（とRS）の探索の傾向について明らかになったことをまとめている。強化学習における探索には、指向性を持った探索（環境の不確実性を縮減することを目的とする）と、ランダムな探索（一様乱数を用いた探索）という二種類の探索がある。Softsatisficingは、指向的探索とランダム探索の間のギャップを連続的に繋ぎ、満足化基準の設定値に応じて、自律的に指向性とランダム性を制御できることを示している。また、二種類のエントロピーを用いて、Softsatisficingの探索の仕方を明らかにしている。

第６章は、本論文の主張をまとめ、今後の課題を述べている。Softsatisficing の提案により、満足化による意思決定が深層強化学習においても容易となる。様々な工学的応用が可能となるのみならず、満足化という行動方策を具体的にはどのように組織や人間や動物が行っているかに関する、定量的・科学的な検証も可能となった。人や動物の意思決定が Softsatisficing により満足化として記述できるか、あるいは個人の基準の推定が可能かどうかの検討と、その実証が今後の課題である。