【 97 】

| | |
|---|---|
| 氏名 | 末松伸朗 |
| 博士の専攻分野の名称 | 博士（工学） |
| 学位記番号 | 第15047号 |
| 学位授与年月日 | 平成12年1月31日 |
| 学位授与の要件 | 学位規則第4条第2項該当 |
| 学位論文名 | 部分観測環境における強化学習とアンサンブル学習法に関する研究 |
| 論文審査委員 | （主査）<br>教授 白井 良明<br>（副査）<br>講師 渡邊 亮　教授 浅田 稔　助教授 三浦 純 |

## 論文内容の要旨

本論文では、未知環境におかれたエージェントが、観測値から環境の状態を必ずしも一意に決定できない部分観測環境で、経験から適切な行動選択法（方策）を学習する強化学習の研究と、属性値の並びで記述される事例を正しいクラスへ分類することを帰納的に学習する問題において、既存の分類器学習法（基本学習器）を繰り返し適用することで分類精度を向上するアンサンブル学習法に関する研究をまとめており、以下の5章からなる。

第1章では、本研究の背景、対象問題、及び、本論文で提案する手法の概要を述べている。

第2章では、環境モデルを可変な短期記憶の確率モデルとして学習する部分観測環境での強化学習法を提案している。本手法のように環境モデルを学習する手法はモデルを用いないものより高い性能の方策を実現できるが、モデル学習の計算量が蓄えられた履歴の長さに応じて増加する問題があった。そこで、事前に設定したモデル候補集合上の事後確率分布を、一定計算量で更新できる履歴の統計量で表す手法を提案している。本手法の理論的解析を行い、実験で有効性を検証している。

第3章では、与えられた環境モデルに対して優れた方策を求める意思決定問題で、従来法の扱えない、環境が複数のマルコフ決定過程モデルで表される場合の解法を提案している。本手法は、従来の意思決定手法「方策反復」を拡張し、複数のモデルに同時に対応できるような確率的な方策空間での山登り法を行う。環境モデルの学習の初期にモデルを一つに決めると、それは真のモデルと大きく異なる危険性が高いが、複数のモデルを取り出して本手法を適用すればこの危険性は軽減できることを実験的に示している。

第4章では、ブースティングと呼ばれるアンサンブル学習法に発想を得た確率モデル学習法を提案している。ある環境で過去の履歴から将来の観測値を分類により予測する場合、履歴に依存する観測値の出現確率分布(確率モデル)を学習しなければならない。一方、ブースティングの理論的基盤は決定的な真理関数の学習にあり、それに基づく従来法は確率モデルを学習できなかった。提案手法は、ブースティングの、学習事例に与えた重みの更新と基本学習器の適用を繰り返す手続きを、求める確率モデルを基本学習器の作る確率モデルで逐次的に因数分解する手続きへ応用し、確率モデルを学習する。本手法は確率モデルを学習するように拡張されていながら、確率モデルの学習が必要で

ない分類を従来法と同等の精度で行えることを実験で示している。

第5章では、本論文のまとめと今後の課題を述べている。

## 論文審査の結果の要旨

一般に、未知環境の観測によって将来の観測値を予測したり、その環境で適切な行動を選択することは困難であるが、この能力を学習によって獲得することは重要なことである。従来の多くの研究は、観測によって環境の状態を完全にわかる場合を扱っているが、ここでは、観測値から環境の状態を必ずしも一意に決定できない部分観測環境を扱っている。

本論文は、部分観測環境において経験から適切な行動選択法（方策）を学習する強化学習と、環境に関する観測値から環境を正しいクラスへ分類することを、既存の分類器学習法を利用して能力を高めるアンサンブル学習を提案したもので、その主な成果は次のとおりである。

(1) 部分観測環境で適切な行動を学習する強化学習法のためには、環境のモデルを学習する必要があるが、従来の手法は過去の観測と行動の履歴の長さに応じて計算量が増加する問題があった。ここでは、事前に履歴のモデル候補集合を設定し、学習では一定計算量でその出現確率を更新する手法を提案している。この履歴モデルを用いて現在の状態を推定し、強化学習を行なうことができる。本手法の理論的解析を行い、実験で有効性を確かめている。

(2) 与えられた環境モデルに適した方策を求める意思決定問題で、従来の手法は環境が一つのマルコフ決定過程モデルで表される場合しか扱えなかった。環境が複数のマルコフ決定過程モデルで表される場合、学習の初期にモデルを一つに決めると、それは真のモデルと大きく異なる危険性が高い。この問題を解決するために、複数の環境モデル候補に対して、方策空間の山登り法によって最適な方策を得るアルゴリズムを提案している。本アルゴリズムの有効性を実験的に示している。

(3) 環境の観測の履歴からその環境を分類して将来を予測する場合、一般に将来の環境の観測値は、履歴に依存する確率分布（確率モデル）に従う。一方、ブースティングと呼ばれるアンサンブル学習法は決定的な真理関数の学習を行ない、通常は確率モデルを学習できなかった。提案手法は、ブースティングの学習事例の重みの更新法と基本学習器の適用の繰り返し手続きを応用して、確率モデルを基本学習器の作る確率モデルで逐次的に因数分解することにより、確率モデルを学習する。本手法は確率モデルを学習するように拡張されているが、確率モデルの学習が必要でない分類（本手法の特殊例）でも、従来法と同等の精度で分類が行えることを実験で示している。

以上のように本論文は、部分観測環境において適切な行動選択法の学習と、環境を正しいクラスへ分類する方策の学習を提案するとともに、観測と行動を伴う一般的なシステムの方策を学習する研究の発展に寄与することが大きい。よって本論文は博士論文として価値のあるものと認める。