—論文題目—

IL-RBMの強化学習タスクへの応用

英語のタイトル

指導教授

萩原 将文 教授

学習指導副主任

遠山 元道 准教授

慶應義塾大学 理工学部 情報工学科

平成　27年度

学籍番号 61212346

筒井佑一郎

目 次

あらまし

ffffffffffffffffffffffffffffffffffffffffffff こ

第 1章

はじめに

近年、機械学習についての研究が盛んである。機械学習の学習手法は大きく三つに分類され、それぞれ教師あり学習、教師なし学習、強化学習である。とりわけ、強化学習は複雑で正確な教師データの与えられない実環境において、ロボット制御や最適化をおこなう有望な手法として注目されている。また、近年のディープラーニング (深層学習: Deep Learning) の成功をうけ、ディープニューラルネットワークを用いた強化学習に対する研究も盛んに行われている。そのような研究の中でも最も有名な研究の一つとして、ディープラーニングと強化学習を組み合わせビデオゲームを解いたものがある [Mnih 13]。この研究では、ディープニューラルネットワークを用いて行動価値観数を近似する手法を用い、複数種のビデオゲームにおいて人間を上回る性能を学習させることに成功した。

一方ニューラルネットワークの学習手法に対する研究も盛んに行われている。一般にニューラルネットワークの学習には多くのパラメータを設定する必要がある。また、ニューラルネットワークは、新規のデータを学習させた場合、過去に学習したデータを忘却する、破壊的干渉と呼ばれる現象がおこる [津守 10]。

これらの問題点を解決するため、多くの手法が提案されてきた。ニューラルネットワークの学習におけるパラメータは、学習率、素子数、層数など多く存在する。例えば、学習率の自動決定として、Adagrad[Duchi 11], や Adam[Kingma, 15] といったものが提案されている。

また、素子数についての研究として Bengio らによる研究 [] などがある。この研究では RBM の中間素子数がデータ数＋ 1 であれば理論上データセットを完全に学習できることが示された。しかしながら実際に使用される際の RBM の主要な役割はデータ

セット内から特徴を抽出することであり、この役割においてデータセットを完全に学習してしまうような素子数を設定することは適切ではない。

ニューラルネットワークは追加学習が出来ないという点について、大澤らは素子を新たに追加するRBMを提案することで、既学習情報を破壊せず追加学習を可能にした[大澤 14]。RBMに対し与えられた学習データを未学習か既学習かを判定し、未学習であれば新たに素子を追加し学習を行うことで、追加学習を行った。

また、大澤は彼らの提案したRBMを強化学習、行動選択タスクに対し応用した。提案したRBMに正の報酬が与えられた際の環境と行動を学習させた。その後、与えられたデータを未学習が既学習かを判定するプロセスを応用することで、新たに選択する行動によってもたらされる環境が、学習済みの環境か否かを判別し、最適行動選択を行った。

しかし、大澤の提案したRBMではデータセットの未学習、既学習は判定できるものの、与えられたデータが正の報酬に関連づいたデータであるのか、負の報酬に関連ついたデータであるのかを判定することが出来ない。

本研究では大澤の提案したRBMを改良し、与えられたデータが正の報酬に関連したデータであるか、府の報酬に関連したデータであるかを判別可能なRBMを提案する。異なるデータセットに対しそれぞれネットワークを割り当てることで、異なる種類のデータに対して未学習、既学習を判定することが可能である。また、正の報酬に関連したデータセットを学習するネットワークと負の報酬に関連したデータセットを学習するネットワークを持つRBMを使用したエージェントを用い、簡単な強化学習タスクへと応用する。正の報酬が与えられた際の環境と行動、負の報酬が与えられた際の環境と行動をエージェントに学習させ、未学習、既学習を判定することで、長期的な観点から正の報酬を獲得し、負の報酬を避けようとする最適行動選択が可能であることを示す。

以下、2章で既存手法の説明をし、3章で提案手法について詳細を述べ、4章で評価実験について示し、5章をまとめとする.

第 2章

強化学習の概要

機械学習の一分野である強化学習について述べる。

2.1　強化学習

2.1.1　強化学習について

2.1.2　代表的な手法

2.1.2.1　モンテカルロ法

2.1.2.2　Q 学習

2.2　ニューラルネットワーク

ニューラルネットワークは、神経科学的な特徴を反映させた計算モデルである。人間の脳のニューロンとシナプスを模しており、ある値を持つノード（人口ニューロン）とその結合荷重によって表現される。

近年、多層に重ねたニューラルネットワークを用いたディープラーニングが画像認識や音声認識などのパターン認識や、データマイニングにおいて非常に目覚ましい成果を上げており、近年注目を集めている。

2.3 ニューラルネットワークの問題

ニューラルネットワークは、多次元量の線形分離不可能なデータを比較的少ない計算量で扱えるといった長所がある一方、以下に列挙するような問題をはらんでいる。

2.3.1 破壊的干渉

あるデータセットを学習済みのニューラルネットワークに対し、新たなデータセットを学習させる際、既に学習したデータを忘却してしまう、という問題がある。これを破壊的干渉という。

異なる性質のデータセットをその違いを考慮し、一つのニューラルネットワークで扱う場合、既学習情報を破壊せず新たなデータセットを学習する必要がある。

このような問題を解決するために、学習データを保持し、新たなデータセットと合わせて学習するという手法が提案された。しかし、このような手法は学習データを保持し続けなければならず、学習時間が長くなり、メモリを大量に必要とする、という欠点がある。

2.3.2 メタパラメータの設定

ニューラルネットワークの学習には、様々なメタパラメータの設定が必要不可欠である。例えば、層の数、それぞれの層におけるノード数、学習率、荷重減衰、モーメンタムなどである。このようなメタパラメータの組み合わせは莫大な数に上り、これらのパラメータの異なるそれぞれのモデルに対し、それぞれ学習を行い予測精度を比較することは、手間と時間が非常にかかる。そのため、これらのメタパラメータの自動決定、最適化は非常に需要のある研究である。

データセットに対し、データセットの特徴を十分学習する素子数の計算する研究がある。

また、学習率の自動決定に関しては、多くの手法が提案されており、その例として、Adagrat、Adam などが挙げられる。

2.3.3 局所解への収束問題

ニューラルネットワーク、特に多層ニューラルネットワークにおいて、局所解への収束問題は非常に重要な問題である。ニューラルネットワークの学習において基本的に用いられる手法に誤差逆伝播法があるか、この誤差逆伝播法は設定された誤差関数を現象させるようにノード間の結合荷重などのパラメータを調整する。したがって、誤差関数の局所解へ収束していしまい、最適解へ辿りつけないという問題がある。

局所解への収束問題はとりわけ多層ニューラルネットワークにおいて顕著であり、局所解への収束問題を解決するために事前学習 (pre-training) という手法が提案されている。

2.4 Restricted Boltzmann Machine

Restricted Boltzmann Machine(RBM) とは、ニューラルネットワークの一種である。

統計的な変動を用いたホップフィールドネットワークの一種である Boltzmann Machine の、可視層間、隠れ層間の結合を制限したものである。

その学習における結合の重みの変化の過程が、脳における神経科学的な学習則であるヘブ則とも類似しているなど、ニューラルネットワークの現在非常に有力な手法の一つである。

RBM は可視層と隠れ層の二層からなる。可視層に入力データを入れ学習させることで、隠れ層にその特徴をよく表すようなパラメータが出力される、という特徴がある。

通常、可視層、隠れ層の各ノードには 0 か 1 の値が入る。可視層に入力データを入れることで、隠れ層の各ノードの取る値の条件付き確率が計算できる。ゆえに、RBM は確率モデルであり生成モデルである。

2.5 Deep Belief Network

Deep Belief Network(DBN) とは、Deep Neural Network の一種である。

Deep Neural Network の一般的な学習方法である、誤差逆伝播法の問題の一つに、局所解への収束問題がある。誤差逆伝播法はある誤差関数を最小化するように各パラ

メータを変化させる手法であるが、その性質上誤差関数の局所的な解に収束し、誤差関数の最適解へ収束しない現象が起こり得る。この現象は Deep Neural Network の層数が増えるに従ってより起こりやすくなる。この局所解への収束問題を解決する手法として事前学習 (pre-training) が考案された。

事前学習の手法にはいくつかあるが、そのうちの一つが前述した RBM を用いるもので、RBM を用いた事前学習を行う Deep Neural Network を DBN と呼ぶ。

DBN では RBM を多段に重ねた構造を取る。そして、一番下の層から入力データを RBM に学習させ、RBM の隠れ層に特徴が現れるよう各パラメータを学習させる。その後、RBM が生成モデルである利点を活かし、一番下の層に入力データを入れた後に計算される、隠れ層の条件付き確率を次層の RBM にの可視層に入力し、同じように特徴を学習させる。

このようにしてパラメータを学習させた Deep Neural Network は、それぞれの層において入力データの特徴をうまく抽出するようパラメータが決定されているため、初期パラメータをランダムに決めていた多層パーセプトロンに比べ局所解へ収束しづらい。

こうして事前学習されたネットワークを最後に教師あり学習で学習させる (fine-tuning) 手法を DBN と呼ぶ。

第 3章

システムの概要

この章ではシステムの概要について説明する。

3.1　システム全体の構成

本システムは、強化学習の基本的な考え方に則り、”環境”と”エージェント”が”行動”と”報酬”により相互に影響を及ぼし合うように構成されている。

3.1.1　システム全体の流れ

システムの流れは以下のとおりである。

1. 二つのエージェントを定義する。
2. 先行のエージェントに盤面の状態を与える。
3. 先行のエージェントは盤面の状態から行動を選択する。
4. 環境は、エージェントの行動を受け取り、自身の状態を更新する。
5. 環境の状態に応じて、エージェントに報酬を与える。
6. 環境の状態が終了条件を満たせばゲームを終了する。
7. 上記手順を先攻、後攻を交代し繰り返す。

環境部分は、常にある状態をを持ち、その状態はエージェントの行動によって変化させられる。また、環境部はその状態とエージェントの行動によってエージェントに与える報酬を決定する。

3.2 Incremental Learning-RBM(IL-RBM)

この節では、本システムの根幹となる IL-RBM について説明する。IL-RBM は追加学習を行うと既学習情報を失うという RBM の欠点を改良したものである。

3.2.1 追加学習

IL-RBM では追加学習を行う。追加学習とは、既にあるデータセットに対し学習済みのニューラルネットワークが、新たなデータセットに対し、既学習情報を失わない形で学習を行うことである。

通常の RBM の隠れ層のノード数は一定である。しかし、IL-RBM は新たなデータセットを学習する際に隠れ層のノード数を追加する操作を行う。IL-RBM の学習は次のような流れをとる。

1. 初期データセットに対し IL-RBM を学習させる。
2. 追加データセットに対し、適切な追加ノード数を決定し、Il-RBM の隠れ層にノードを追加する。
3. 上記手順で追加したノードのみを用いて、追加データセットを学習させる。
4. 上記 2、3 の手順を追加データセットの分だけ繰り返し行う。

3.3 IL-RBM の強化学習への応用

本システムでは、IL-RBM を用いて強化学習タスクを行う。IL-RBM を用いたエージェントを設定し、このエージェントに、ある環境下において得られる報酬を最大化するような行動を学習させる。

3.3.1 強化学習

強化学習とは、あるエージェントがある環境内にて、得られる報酬を最大化するような行動を学習するような機械学習のことである。

強化学習には重要な 4 つの概念がある。

- 環境 … エージェントの行動に応じて、報酬をエージェントに与える。また、エージェントの行動に応じて、エージェントの観測する環境も更新される。
- 報酬 … エージェントの行動に応じて環境からエージェントに与えられる。この得られる報酬を最大化するようエージェントは行動を学習する。
- 行動 … エージェントは観測した環境に応じて、行動を選択する。
- エージェント … 環境を観測し、報酬を受け取り、行動を選択する。

明確な教師データが与えられる教師あり学習や、全く教師データが与えられない教師なし学習と異なり、報酬という限定されたフィードバックのみが与えられる点に特徴がある。不確実な環境を取り扱えるという点で、応用上非常に有望な機械学習手法の一つである。

3.3.2 未学習データセットの判別

エージェントに与えられたあるデータを既学習であるか、未学習であるかを判別する手法について述べる。エージェントに与えられたデータの既学習判定に RBM のエネルギーを用いる。RBM は学習済みのデータに対して、エネルギーが低くなる、という特徴を持つため、データを入力した際の RBM のエネルギーを見ることで入力されたデータが既学習であるか、未学習であるかを判別することができる。

3.3.3 エージェントの概要

エージェントの概要について具体的に説明する。

本システムで用いたエージェントは、IL-RBM と出力層からなる DBN(以下、IL-DBN) を用いている。この IL-DBN は入力データを環境、出力を行動として学習する。エージェントは、常に今自分が置かれている環境を観測することが可能であり、環境から報酬を与えられた場合には、それを感知することが可能である。

エージェントは環境と行動の組からなるデータセットを自動的に構築し、そのデータセットを逐次的に学習することで、最適行動を学習する。また、IL-RBM のエネルギーに注目することで、未学習データセットを検出できるという特徴を用い、後述するサブゴールを獲得し、長期的な戦略を獲得することが可能である。

3.3.4 データセットの獲得

本システムがデータセットを環境中から自動で学習するメカニズムについて説明する。以下、三目並べタスクの最適行動を学習するエージェントを例に説明する。

学習エージェントが対戦相手と三目並べを行う環境下で、三目並べに勝利した場合と敗北した場合にそれぞれ正の報酬と負の報酬を与えられるとする。この時、エージェントが観測する環境は盤面の状態であり、出力する行動は次に選択する手の盤面上の位置となる。

3.3.4.1 勝利データセットの獲得

まず、エージェントは勝利データセットを収集する。ある行動を出力した際、盤面が勝利条件を満たし、正の報酬が与えられたとする。その勝利条件を満たした際の行動と、その行動をとった際の環境を組みにしてデータセットとしてエージェントは保存する。

データセットがある一定数に達した場合、または試行回数が一定数に達した場合、エージェントは保存したデータセットを用いて学習を行う。

3.3.4.2 サブゴールの獲得

前述した、既学習、未学習判定を用いてサブゴールを獲得し、長期的な戦略を獲得するメカニズムについて説明する。

勝利データセットの獲得において、盤面が勝利条件を満たした場合、その時にとった行動と、行動を選択した際の環境を組みにしてデータセットとして保存した。この場合、勝利条件を満たした盤面の状態をゴールとしてデータセットを採集している。

サブゴールとは、ゴールにつながり得る環境の状態を指す。サブゴールを適切に設定し、サブゴールに至る行動と、その行動を選択した際の環境を更にデータセットに加える事で、長期的な戦略を獲得することができる。

サブゴールの設定方法について説明する。ある行動を選択し、環境が更新されたとする。その際の環境をエージェントの RBM が既学習であるか未学習であるかを前述したエネルギーによる判定法で判定する。そして、得られた環境が既学習であった場合、その環境はゴールへと至る可能性が高いため、その環境をサブゴールとして設定する。そして、サブゴールに至る直前の環境と、その状態にて選択した行動をデータセットとして新たに保存し、追加学習を行う。

このようにして、サブゴールについても学習した IL-RBM は、また新たに”サブゴールのサブゴール”を扱うことが可能になる。観測された環境が、既に学習済みであるサブゴールである環境と一致、近似すると判定された場合、さらにその環境に至る行動とその直前の環境をデータセットとして加えることで、サブゴールのサブゴールを学習でき、長期的な戦略を獲得することが可能である。

3.3.5 負のネットワーク、負のサブゴール

前述したサブゴールの設定法には負の報酬と行動の抑制を扱えない、という欠点があった。ある RBM がある環境を表す入力データを未学習か既学習か判定できたとしても、その環境が正の報酬に紐付いているか、負の報酬に紐付いているかを判定できないからである。

そこで、負の報酬と負の報酬を得る環境へ至る行動を抑制するために負の報酬をあつかうネットワークをエージェントに追加することにした。

エージェントは IL-RBM を二つ使用する。それぞれ正の報酬を扱うネットワークと負の報酬を扱うネットワークである。また、保持するデータセットも二種類用意する。正の報酬を得ることのできるゴールに至る行動と、その直前の環境の組、サブゴール

に至る行動とその直前の環境の組をデータセットとして扱う正のデータセットと、負の報酬について同様に扱う負のデータセットである。

負のデータセットを扱う負のネットワークでは、負の報酬を得ることになるゴール、サブゴールに至る行動が出力され、その直前の環境が入力データとなる。したがって、観測された環境をこのネットワークに入力した際の出力された行動は抑制されるべき行動である。

本エージェントでは負のネットワークからの出力を正のネットワークからの出力から引くことで、負の報酬を得る環境へ至る行動を抑制している。しかし、正のネットワークと負のネットワークを同列に扱うべきか、荷重をかけどちらかを優先すべきかなどは研究途中である。

第 4章

評価実験

本研究では、提案手法の強化学習への妥当性の検証のため、三目並べタスクを行った。実験を行うために制作したシステムの詳細について述べた後、実験と結果についての報告と考察を行う。

4.1　三目並べタスクの強化学習システムの詳細

本システムは強化学習の基本的な考え方に則り、環境とエージェントの相互作用を取り扱う。

4.1.1　環境部

環境部はエージェントと環境の相互作用を制御する main オブジェクトと、三目並べタスクの詳細の動作、情報を制御する ThreePiece オブジェクトからなる。

4.1.2　main オブジェクト

main オブジェクトは、プログラム全体の制御を行う。

まず、行うタスクを定義する。今回行うタスクは三目並べであるため ThreePiece オブジェクトをタスクとして定義する。その後、タスクの要請する数のエージェントを定義する。今回の実験では、先行のエージェントをランダムエージェント、後攻のエージェントを IL-DBN エージェントとした。

その他、エージェントと環境の相互作用を制御し、三目並べの終了時に ThreePiece オブジェクトとエージェントのリセットを行ったり、三目並べを何ターン行うかの制御、それぞれのエージェント数の勝利数の保持を行う。

4.1.3 ThreePiece オブジェクト

ThreePiece オブジェクトは三目並べタスクを実際に執り行う。ThreePiece オブジェクトはターン制でそれぞれのエージェントに盤面の情報を渡し、行動情報を受け取る。盤面の情報の次元数や行動情報の形式は ThreePiece オブジェクトが決定し、エージェントがアーキテクチャをその形式に合わせる。

盤面の情報はそれぞれのマスに対し、空白ならば (0,0), 白石が置かれていれば (0,1)、黒石が置かれていれば (1,0) の 2bit で表現する。したがって 9 マス全体の表現は 18 次元の (0,0,0,1,1,0,0,0,0,1,0,0,0,0,0,0,1,0) のような表現になる。

一方エージェントの行動は 0〜8 の整数値で表される。それぞれの数値が石を置く盤面上の位置を示している。

ThreePiece オブジェクトはエージェントから渡された石の置き位置を示す値がルール上正当なものかを判定する。エージェントが石を置こうとしている場所に既に石が置かれている場合はエージェントに再度行動を選択するように命令する。指定された石の置き位置が正当であれば、オブジェクトの保持する盤面の状態を更新する。

その後、盤面の状態が三目並べの終了条件を満たしているかを判定する。盤面の状態を監視し、縦、横、斜めいずれかの方向に石が 3 つ並んでいれば、勝利エージェントへ報酬 1 を、敗北エージェントへ報酬-1 を与える。

4.1.4 エージェント部

エージェントは基本的なエージェントの振る舞いを規定する Agent クラスがあり、それを継承することでそれぞれの Agent のクラスが作られている。

エージェントは環境から盤面状態を受け取った後、行動選択を行う。ランダムエージェントであれば 0〜 8 で乱数を返し、学習エージェントは学習した行動を出力する。

環境が行動を受け取った後、エージェントは環境から報酬を受け取る。学習エージェントは報酬に応じて学習を行う。その学習の具体的なプロセスについては後述する。

4.1.4.1 学習エージェント

学習エージェントは次のような流れで学習を行う。

1. 環境から状態が与えられる。
2. 環境が記憶済みの状態の場合、一つ手前の状態とその状態でとった行動をデータセットに追加する。
3. 行動選択後、環境から報酬が与えられる。
4. 報酬に応じて状態と行動の組をデータセットに追加する。
5. 上記の 1 から 4 を一定回数繰り返す。
6. 採集したデータセット用いて追加する RBM のノード数を決定する。
7. ノード数の確定した RBM を採集したデータセットで学習させる。
8. データセットを空にリセットし 1 から再度データセットを採集する。

4.2 予備実験と各種パラメータの決定

学習率、エポック数等の各種パラメータを決定するために予備実験を行った。

4.2.1 RBM の学習率とエポック数

RBM の学習時の学習率は文献を [] を参考に決定した。学習エージェントが実際に採集した、入力を盤面の状態とし、出力を次に石を置く盤面の位置とするデータセットを利用する。RBM の素子数はほにゃららの自動決定法により決定した。エポック数は以下のグラフのようになった。以上の結果より

4.2.1.1 出力層の学習率とエポック数

出力層の学習率とエポック数を決定するため予備実験を行った。文献 [] を参考に学習率を決定した。学習エージェントが実際に採集した、入力を盤面の状態とし、出力を次に石を置く盤面の位置とするデータセットを利用する RBM の素子数はほにゃらの自動決定法により決定した。エポック数は以下のグラフのようになった。以上の結果より

4.3 提案手法の評価実験

以下の条件下で実験を行った。- 1 ターン 1000 回の三目並べの試行を行う。- ターンの終了ごとに追加学習を行う。- 5 ターンにわたり勝利数、敗北数、引き分け数を記録する

第 5 章

結論

本論文では、IL-RBM を改良し、強化学習タスクへの応用方法を示した。

文献で提案された IL-RBM は与えられたデータが既学習か未学習かを判定することが可能であるが、強化学習への応用の際、与えられたデータが正の報酬に関連づくものか、負の報酬に関連づくものかを区別することが出来なかった。

そこで、IL-RBM に正と負の二種類のネットワークを持たせることで、IL-RBM が正の報酬と関連の強いデータと負の報酬と関連の強いデータを区別することが可能であることを示した。

IL-RBM が正のネットワークと負のネットワークを持つことで、既存手法では扱えなかった負のサブゴールを設定可能となり、負の報酬を避けるような長期的な戦略を獲得できることを示した。

また、提案した IL-RBM をもちいたエージェントに三目並べタスクを実際に解かせた。エージェントは正の報酬へ繋がる行動のデータセットと負の報酬へ繋がる行動のデータセットを採集し、それぞれを提案した正負のネットワークで学習することで、負のネットワークを持たない IL-RBM より高い勝率で勝つことができることを示した。また、負の報酬を避けるような長期的な戦略により、より顕著に敗北率が減少することを示した。