

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号  
特許第7529028号  
(P7529028)

(45)発行日 令和6年8月6日(2024.8.6)

(24)登録日 令和6年7月29日(2024.7.29)

(51)国際特許分類

F I

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

請求項の数 10 (全16頁)

|             |                             |          |                         |
|-------------|-----------------------------|----------|-------------------------|
| (21)出願番号    | 特願2022-545246(P2022-545246) | (73)特許権者 | 000004237               |
| (86)(22)出願日 | 令和2年8月31日(2020.8.31)        |          | 日本電気株式会社                |
| (86)国際出願番号  | PCT/JP2020/032848           |          | 東京都港区芝五丁目7番1号           |
| (87)国際公開番号  | WO2022/044314               | (74)代理人  | 100103090               |
| (87)国際公開日   | 令和4年3月3日(2022.3.3)          |          | 弁理士 岩壁 冬樹               |
| 審査請求日       | 令和5年2月15日(2023.2.15)        | (74)代理人  | 100124501               |
|             |                             |          | 弁理士 塩川 誠人               |
|             |                             | (72)発明者  | 江藤 力                    |
|             |                             |          | 東京都港区芝五丁目7番1号 日本電気株式会社内 |
|             |                             | 審査官      | 多賀 実                    |

最終頁に続く

(54)【発明の名称】 学習装置、学習方法および学習プログラム

(57)【特許請求の範囲】

【請求項1】

候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる前記候補特徴量の各重みを導出する第一逆強化学習実行部と、

前記各重みが導出された候補特徴量から一つの特徴量を選択した場合に、当該特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づく と推定される特徴量を選択する特徴量選択部と、

選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行部とを備えた

ことを特徴とする学習装置。

10

【請求項2】

特徴量選択部は、導出された候補特徴量の各重みを最適なパラメータとみなし、候補特徴量の中から、目的関数の部分最適性を最小にする特徴量を選択する

請求項1記載の学習装置。

【請求項3】

第二の目的関数の学習結果に基づいて、候補特徴量の中から、さらに特徴量を選択するか否か判定する判定部を備え、

特徴量選択部は、さらに特徴量を選択すると判定された場合、前記候補特徴量の中から、既に選択された特徴量以外の特徴量を新たに選択し、

第二逆強化学習実行部は、新たに選択された特徴量を加えて逆強化学習を実行すること

20

により、第二の目的関数を生成する

請求項 1 または請求項 2 記載の学習装置。

【請求項 4】

生成された第二の目的関数の情報量規準を計算する情報量規準計算部を備え、

判定部は、前記情報量規準に基づいて、候補特徴量の中からさらに特徴量を選択するか否かを判定する

請求項 3 記載の学習装置。

【請求項 5】

判定部は、情報量規準が単調増加する場合に、候補特徴量の中から、さらに特徴量を選択すると判定する

請求項 4 記載の学習装置。

【請求項 6】

情報量規準が最大になったときの第二の目的関数に含まれる特徴量および対応する特徴量の重みを出力する出力部を備えた

請求項 1 から請求項 5 のうちのいずれか 1 項に記載の学習装置。

【請求項 7】

出力部は、特徴量選択部によって選択された順に特徴量を出力する

請求項 6 記載の学習装置。

【請求項 8】

特徴量選択部が選択した特徴量をユーザに提示する特徴量提示部と、

提示された前記特徴量に対するユーザからの選択の指示を受け付ける指示受付部とを備え、

特徴量選択部は、理想的な報酬の結果により近づくことと推定される、予め定めた数の上位の特徴量を 1 つ以上選択し、

前記特徴量提示部は、選択された一つ以上の特徴量をユーザに提示し、

第二逆強化学習実行部は、ユーザにより選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する

請求項 1 から請求項 7 のうちのいずれか 1 項に記載の学習装置。

【請求項 9】

コンピュータが、候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる前記候補特徴量の各重みを導出し、

前記コンピュータが、前記各重みが導出された候補特徴量から一つの特徴量を選択した場合に、当該特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくことと推定される特徴量を選択し、

前記コンピュータが、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する

ことを特徴とする学習方法。

【請求項 10】

コンピュータに、

候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる前記候補特徴量の各重みを導出する第一逆強化学習実行処理、

前記各重みが導出された候補特徴量から一つの特徴量を選択した場合に、当該特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくことと推定される特徴量を選択する特徴量選択処理、および、

選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行処理を実行させる

ための学習プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

10

20

30

40

50

本発明は、逆強化学習を行う学習装置、学習方法および学習プログラムに関する。

【背景技術】

【0002】

機械学習の分野において、逆強化学習の技術が知られている。逆強化学習では、熟練者の意思決定履歴データを利用して、目的関数における特徴量ごとの重み（パラメータ）を学習する。

【0003】

他にも、機械学習の分野において、特徴量を自動で決定する技術が知られている。非特許文献1には、“Teaching Risk”に基づく特徴量選択の技術を開示する。非特許文献1に記載された方法では、目的関数における理想的なパラメータを仮定して学習過程のパラメータと比較し、二つのパラメータの差をより小さくする特徴量を重要な特徴量として選択する。

10

【先行技術文献】

【非特許文献】

【0004】

【文献】Luis Haug, et al., "Teaching Inverse Reinforcement Learners via Features and Demonstrations", Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 8473-8482, December 2018.

【発明の概要】

【発明が解決しようとする課題】

20

【0005】

逆強化学習を行う場合には、目的関数に含まれる特徴量をユーザが指定する必要がある。しかし、現実の問題に逆強化学習を適用する場合、様々なトレードオフ関係を考慮して目的関数の特徴量を設計する必要がある。そのため、逆強化学習を行う際の目的関数の特徴量設計は高コストになってしまうという問題がある。

【0006】

そのため、非特許文献1に記載された方法を用いて特徴量選択を行うことも考えられる。非特許文献1に記載された方法は、理想的なパラメータを仮定することを前提としているが、そもそも、このような理想的なパラメータを導出する方法自体が不明確である。そのため、非特許文献1に記載された方法そのままでは、逆強化学習の特徴量の選択に利用することは難しい。

30

【0007】

そこで、本発明は、逆強化学習で用いる目的関数の特徴量の選択を支援できる学習装置、学習方法および学習プログラムを提供することを目的とする。

【課題を解決するための手段】

【0008】

本発明による学習装置は、候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる候補特徴量の各重みを導出する第一逆強化学習実行部と、各重みが導出された候補特徴量から一つの特徴量を選択した場合に、その特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくと推定される特徴量を選択する特徴量選択部と、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行部とを備えたことを特徴とする。

40

【0009】

本発明による学習方法は、コンピュータが、候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる候補特徴量の各重みを導出し、コンピュータが、各重みが導出された候補特徴量から一つの特徴量を選択した場合に、その特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくと推定される特徴量を選択し、コンピュータが、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成することを特徴とする。

【0010】

50

本発明による学習プログラムは、コンピュータに、候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる候補特徴量の各重みを導出する第一逆強化学習実行処理、各重みが導出された候補特徴量から一つの特徴量を選択した場合に、その特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくことと推定される特徴量を選択する特徴量選択処理、および、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行処理を実行させることを特徴とする。

【発明の効果】

【0011】

本発明によれば、逆強化学習で用いる目的関数の特徴量の選択を支援できる。

10

【図面の簡単な説明】

【0012】

【図1】本発明による学習装置の第一の実施形態の構成例を示すブロック図である。

【図2】第一の実施形態の学習装置の動作例を示すフローチャートである。

【図3】本発明による学習装置の第二の実施形態の構成例を示すブロック図である。

【図4】ユーザに提示される特徴量の候補の例を示す説明図である。

【図5】第二の実施形態の学習装置の動作例を示すフローチャートである。

【図6】本発明による学習装置の概要を示すブロック図である。

【図7】少なくとも1つの実施形態に係るコンピュータの構成を示す概略ブロック図である。

20

【発明を実施するための形態】

【0013】

以下、本発明の実施形態を図面を参照して説明する。

【0014】

実施形態1.

図1は、本発明による学習装置の第一の実施形態の構成例を示すブロック図である。本実施形態の学習装置100は、対象者の行動から報酬（関数）を推定する逆強化学習を行う装置である。学習装置100は、記憶部10と、入力部20と、第一逆強化学習実行部30と、特徴量選択部40と、第二逆強化学習実行部50と、情報量規準計算部60と、判定部70と、出力部80とを備えている。

30

【0015】

記憶部10は、学習装置100が各種処理を行うために必要な情報を記憶する。記憶部10は、後述する第一逆強化学習実行部30および第二逆強化学習実行部50が学習に用いる熟練者の意思決定履歴データ（トラジェクトリと言うこともある。）や、目的関数の特徴量の候補を記憶していてもよい。さらに、記憶部10は、特徴量の候補と、その特徴量の内容を示す情報（ラベル）とを対応付けて記憶していてもよい。

【0016】

また、記憶部10は、後述する第一逆強化学習実行部30および第二逆強化学習実行部50を実現するための数理最適化ソルバを記憶していてもよい。なお、数理最適化ソルバの内容は任意であり、実行する環境や装置に応じて決定されればよい。記憶部10は、例えば、磁気ディスク等により実現される。

40

【0017】

入力部20は、学習装置100が各種処理を行うために必要な情報の入力を受け付ける。入力部20は、例えば、上述する意思決定履歴データの入力を受け付けてもよい。

【0018】

第一逆強化学習実行部30は、候補とする複数の特徴量（以下、候補特徴量と記す。）を用いて目的関数（以下、第一の目的関数と記す。）を設定する。具体的には、第一逆強化学習実行部30は、候補として想定されるすべての特徴量を候補特徴量として第一の目的関数を設定してもよい。そして、第一逆強化学習実行部30は、逆強化学習により、第一の目的関数に含まれる候補特徴量の各重み $w^*$ を導出する。

50

## 【 0 0 1 9 】

このようにして学習された第一の目的関数は、想定されるすべての特徴量を用いて報酬を表現していることから、複数の要因を想定した理想的な報酬の結果を表わしていると言える。また、以下の説明では、第一の目的関数を学習する際に用いた候補特徴量全体を含むリストを、特徴量リスト A と記すこともある。

## 【 0 0 2 0 】

特徴量選択部 40 は、各重み  $w^*$  が導出された候補特徴量から一つの特徴量を選択した場合に、その特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくと推定される特徴量を選択する。このような特徴量は、候補特徴量のうち最も報酬に影響を与え得る特徴量と言える。別の言い方をすると、特徴量選択部 40 は、上述する特徴量リスト A から特徴量の一つを選択する処理を行っているとも言える。

10

## 【 0 0 2 1 】

特徴量選択部 40 は、例えば、熟練者が最も重視すると判断している特徴量を、理想的な報酬の結果に最も近づくと推定される特徴量として選択してもよい。また、このような熟練者も意識していないような特徴量を選択できるようにするため、特徴量選択部 40 は、非特許文献 1 に記載された方法を用いて、候補特徴量の中から特徴量を選択してもよい。

## 【 0 0 2 2 】

以下、非特許文献 1 に記載された Teaching Risk の技術を利用して、候補特徴量の中から一の特徴量を選択する方法を説明する。非特許文献 1 に記載された Teaching Risk は、逆強化学習によって学習された目的関数の（潜在的な）部分最適性を示す値である。目的関数の部分最適性を説明するために、恣意的に選択された特徴量に基づいて逆強化学習により目的関数を最適化（学習）すると仮定する。この場合、逆強化学習により最適化（学習）された目的関数は部分最適ではあるが、全体最適ではない（潜在的な）可能性がある。特徴量が恣意的に選択されているため、選択されなかった特徴量による最適化（学習）を考慮することができないからである。

20

## 【 0 0 2 3 】

また別の仮定として、特徴量が未選択の目的関数を仮定する。この場合、その目的関数と全体最適である理想的な目的関数とは、特徴量を選択する場合と比べて最も異なる。そのため、特徴量が未選択の目的関数における Teaching Risk は最大の状態である。この状態で、Teaching Risk を小さくする特徴量を選択することは、理想の特徴量ベクトルと現実の特徴量ベクトルの差を小さくすることで、潜在的な部分最適性を小さくする特徴量を選択しているため、理想的な報酬の結果に近づくと推定される特徴量を選択することに対応する。

30

## 【 0 0 2 4 】

以下、Teaching Risk の定義を説明する。以下、理想の特徴量ベクトルと、現実の特徴量ベクトルの差を表現した情報を WorldView と記載する。WorldView は、行列で表わすことが可能である。スパース学習の場合、WorldView を示す行列  $A^L$  は、使用される特徴量に対する対角成分が 1、それ以外が 0 になる行列に対応する。すなわち、

現在の特徴量ベクトル =  $A^L \cdot$  理想の特徴量ベクトル  
である。

40

## 【 0 0 2 5 】

理想の重みを  $w^*$  とした場合、Teaching Risk ( $\rho(A^L; w^*)$ ) は、以下に例示する式 1 で表わすことができる。

## 【 0 0 2 6 】

## 【数 1】

$$\rho(A^L; w^*) := \max_{v \in \ker A^L, \|v\| \leq 1} \langle w^*, v \rangle \quad (\text{式 1})$$

## 【 0 0 2 7 】

50

式 1 において、左辺は、理想の重みと WorldView の核（カーネル）に属するベクトルとの内積の最大値を表わす。なお、行列の核（カーネル）は、その行列による線形変換で零ベクトルになるベクトル集合のことであり、Teaching Risk の場合、このベクトル集合と理想の重みのコサインに対応する。

【 0 0 2 8 】

そこで、特徴量選択部 4 0 は、導出された候補特徴量の各重み  $w^*$  を最適なパラメータとみなし、候補特徴量の中から、Teaching Risk を最小にする特徴量を選択してもよい。

【 0 0 2 9 】

以下の説明では、特徴量選択部 4 0 が選択した特徴量を、特徴量リスト B に追加するものとする。具体的には、特徴量選択部 4 0 は、選択した特徴量を上述する特徴量リスト A の中から除去し、特徴量リスト B に追加する。なお、初期状態において、特徴量リスト B を空集合に初期化しておけばよい。

10

【 0 0 3 0 】

第二逆強化学習実行部 5 0 は、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する。具体的には、第二逆強化学習実行部 5 0 は、選択された特徴量（具体的には、特徴量リスト B に追加された特徴量）を用いて目的関数（以下、第二の目的関数と記す。）を設定する。そして、第二逆強化学習実行部 5 0 は、逆強化学習により、第二の目的関数に含まれる特徴量の各重み  $w$  を導出する。なお、特徴量選択部 4 0 により新たに特徴量が選択された場合（具体的には、特徴量リスト B にさらに特徴量が追加された場合）、第二逆強化学習実行部 5 0 は、新たに選択された特徴量とすでに選択されている特徴量とを含む第二の目的関数を設定し、設定された第二の目的関数に含まれる特徴量の各重みを導出する。

20

【 0 0 3 1 】

情報量規準計算部 6 0 は、生成された第二の目的関数の情報量規準を計算する。情報量規準の計算方法は任意であり、例えば、A I C（Akaike's Information Criterion）、B I C（Bayesian Information Criterion）、F I C（Focused Information Criterion）など、任意の計算方法を用いることが可能である。どの計算方法を用いるかは、予め決めておけばよい。

【 0 0 3 2 】

判定部 7 0 は、第二の目的関数の学習結果に基づいて、候補特徴量の中から特徴量をさらに選択するか否かを判定する。判定部 7 0 は、例えば、第二の目的関数の学習回数や、実行時間等、予め定めた条件を満たすか否かに基づいて、候補特徴量の中から特徴量をさらに選択するか否かを判定してもよい。この条件は、例えば、ロボット制御等において搭載可能なセンサの数などに応じて定められてもよい。

30

【 0 0 3 3 】

また、判定部 7 0 は、情報量規準計算部 6 0 によって計算された情報量規準に基づいて特徴量をさらに選択するか否かを判定してもよい。具体的には、判定部 7 0 は、情報量規準が単調増加している場合に、特徴量をさらに選択すると判定する。

【 0 0 3 4 】

判定部 7 0 により、特徴量をさらに選択すると判定された場合、特徴量選択部 4 0 は、候補特徴量の中から、既に選択された特徴量以外で、さらに特徴量を選択し、第二逆強化学習実行部 5 0 は、新たに選択された特徴量を加えて逆強化学習を実行することにより、第二の目的関数を生成し、情報量規準計算部 6 0 は、生成された第二の目的関数の情報量規準を計算する。以降、これらの処理が繰り返される。

40

【 0 0 3 5 】

言い換えると、判定部 7 0 により、特徴量をさらに選択すると判定された場合、特徴量選択部 4 0 は、特徴量リスト A の中から、さらに特徴量を選択して特徴量リスト B に追加し、第二逆強化学習実行部 5 0 は、特徴量リスト B に含まれる特徴量を含む第二の目的関数の重みを導出する。

【 0 0 3 6 】

50

なお、判定部 70 が情報量規準を用いずに予め定めた条件を満たすか否かに基づいて、候補特徴量の中から特徴量をさらに選択するか否か判定する場合、学習装置 100 は、情報量規準計算部 60 を備えていなくてもよい。

【0037】

ただし、判定部 70 が情報量規準計算部 60 により計算された情報量規準を用いて、特徴量をさらに選択するか否か判定することで、特徴量の数とフィッティングのトレードオフを実現できる。すなわち、全ての特徴量を用いて目的関数を表現することで、既存のデータに対するフィッティングを高めることができる一方、過適合を生ずる恐れもある。一方、本実施形態では、情報量規準を用いることで、より好ましい特徴量で目的関数を表現しつつ、スパースな目的関数を実現することが可能になる。

10

【0038】

出力部 80 は、生成された第二の目的関数に関する情報を出力する。具体的には、出力部 80 は、生成された第二の目的関数に含まれる特徴量のセットと、その特徴量の重みとを出力する。出力部 80 は、例えば、情報量規準が最大になったときの特徴量のセットと、その特徴量の重みとを出力してもよい。

【0039】

なお、情報量規準が単調増加しているか否かで特徴量を選択するか否か判定される場合、判定部 70 が特徴量をさらに選択しないと判定したときの情報量規準は、一つ前の第二の目的関数の情報量規準よりも小さくなっていると考えられる。そこで、この場合、出力部 80 は、一つ前の第二の目的関数に関する情報を出力すればよい。

20

【0040】

また、出力部 80 は、特徴量選択部 40 が選択した順番に特徴量を出力してもよい。特徴量選択部 40 が選択した特徴量の順番とは、理想的な報酬の結果に近づく順番であることから、ユーザは、より報酬に影響を与えうる特徴量の順番を把握することが可能になる。また、出力部 80 は、特徴量の内容を表わす情報（ラベル）を合わせて出力してもよい。このように特徴量を出力することで、利用者の解釈性を高めることが可能になる。

【0041】

入力部 20 と、第一逆強化学習実行部 30 と、特徴量選択部 40 と、第二逆強化学習実行部 50 と、情報量規準計算部 60 と、判定部 70 と、出力部 80 とは、プログラム（学習プログラム）に従って動作するコンピュータのプロセッサ（例えば、CPU（Central Processing Unit）、GPU（Graphics Processing Unit））によって実現される。

30

【0042】

例えば、プログラムは、学習装置 100 が備える記憶部 10 に記憶され、プロセッサは、そのプログラムを読み込み、プログラムに従って、入力部 20、第一逆強化学習実行部 30、特徴量選択部 40、第二逆強化学習実行部 50、情報量規準計算部 60、判定部 70 および出力部 80 として動作してもよい。また、学習装置 100 の機能が SaaS（Software as a Service）形式で提供されてもよい。

【0043】

また、入力部 20 と、第一逆強化学習実行部 30 と、特徴量選択部 40 と、第二逆強化学習実行部 50 と、情報量規準計算部 60 と、判定部 70 と、出力部 80 とは、それぞれが専用のハードウェアで実現されていてもよい。また、各装置の各構成要素の一部又は全部は、汎用または専用の回路（circuitry）、プロセッサ等やこれらの組合せによって実現されもよい。これらは、単一のチップによって構成されてもよいし、バスを介して接続される複数のチップによって構成されてもよい。各装置の各構成要素の一部又は全部は、上述した回路等とプログラムとの組合せによって実現されてもよい。

40

【0044】

また、学習装置 100 の各構成要素の一部又は全部が複数の情報処理装置や回路等により実現される場合には、複数の情報処理装置や回路等は、集中配置されてもよいし、分散配置されてもよい。例えば、情報処理装置や回路等は、クライアントサーバシステム、クラウドコンピューティングシステム等、各々が通信ネットワークを介して接続される形態

50

として実現されてもよい。

【0045】

次に、本実施形態の学習装置100の動作を説明する。図2は、本実施形態の学習装置100の動作例を示す説明図である。図2では、Teaching Risk および特徴量リストを用いて、情報量規準に基づき、特徴量を選択する動作を説明する。

【0046】

初めに、第一逆強化学習実行部30は、特徴量リストAにすべての特徴量を格納し、特徴量リストBを空集合として初期化する(ステップS11)。次に、第一逆強化学習実行部30は、すべての特徴量を用いた逆強化学習で目的関数の重み $w^*$ を推定する(ステップS12)。

10

【0047】

以降、情報量規準が単調増加している間は、ステップS14からステップS17の処理が繰り返される。すなわち、判定部70は、情報量基準が単調増加していると判断したときに、ステップS14からステップS17の処理を繰り返し実行する制御を行う(ステップS13)。

【0048】

まず、特徴量選択部40は、特徴量リストAの中から、重み $w^*$ および特徴量リストBに格納された特徴量を用いたTeaching Riskが最小になる特徴量を1つ選択する(ステップS14)。そして、特徴量選択部40は、特徴量リストAの中から選択された特徴量を削除し、特徴量リストBへ追加する(ステップS15)。第二逆強化学習実行部50は、特徴量リストBに含まれる特徴量で逆強化学習を実行し(ステップS16)、情報量規準計算部60は、生成された目的関数の情報量規準を算出する(ステップS17)。

20

【0049】

情報量規準が単調増加しなくなると、出力部80は、生成された目的関数に関する情報を出力する(ステップS18)。

【0050】

以上のように、本実施形態では、第一逆強化学習実行部30が、候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる候補特徴量の各重みを導出し、特徴量選択部40が、各重みが導出された候補特徴量から、理想的な報酬の結果に最も近づくように推定される特徴量を選択する。そして、第二逆強化学習実行部50が、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する。よって、逆強化学習で用いる目的関数の特徴量の選択を支援できる。

30

【0051】

すなわち、本実施形態では、機械学習の過程で適切な特徴量が選択されるため、膨大な特徴量候補の中から、適切な特徴量を低コストで選択することが可能になる。

【0052】

実施形態2

次に、本発明の学習装置の第二の実施形態を説明する。第二の実施形態では、第二の目的関数の学習に用いる特徴量の候補をユーザに提示して選択させる態様を説明する。

【0053】

40

図3は、本発明による学習装置の第二の実施形態の構成例を示すブロック図である。本実施形態の学習装置200は、記憶部10と、入力部20と、第一逆強化学習実行部30と、特徴量選択部41と、特徴量提示部42と、指示受付部43と、第二逆強化学習実行部51と、情報量規準計算部60と、判定部70と、出力部80とを備えている。

【0054】

すなわち、本実施形態の学習装置200は、第一の実施形態の学習装置100と比較し、特徴量選択部40および第二逆強化学習実行部50の代わりに、特徴量選択部41と、特徴量提示部42と、指示受付部43と、第二逆強化学習実行部51とを備える点において異なる。それ以外の構成は、第一の実施形態と同様である。

【0055】

50

特徴量選択部 4 1 は、第一の実施形態の特徴量選択部 4 0 と同様、候補特徴量から特徴量を選択する。その際、本実施形態の特徴量選択部 4 1 は、理想的な報酬の結果により近づくと推定される、予め定めた数の上位の特徴量を 1 つ以上選択する。なお、選択される特徴量の数が 1 つの場合、特徴量選択部 4 1 の行う処理は、第一の実施形態の特徴量選択部 4 0 が行う処理と同一である。

【 0 0 5 6 】

特徴量提示部 4 2 は、特徴量選択部 4 1 が選択した特徴量をユーザに提示する。例えば、特徴量が複数選択されている場合、特徴量提示部 4 2 は、上位の特徴量から順に表示してもよい。また、特徴量のラベルが存在する場合、特徴量提示部 4 2 は、特徴量に対応するラベルを合わせて表示してもよい。

10

【 0 0 5 7 】

図 4 は、ユーザに提示される特徴量の候補の例を示す説明図である。図 4 に示す例では、特徴量提示部 4 2 が、第一の実施形態で例示する Teaching Task の逆数が横軸に、候補の特徴量が縦軸にそれぞれ設定されたグラフを、値の上位から 4 つ選択して表示していることを示す。

【 0 0 5 8 】

指示受付部 4 3 は、特徴量提示部 4 2 によって提示された特徴量の候補に対するユーザからの選択指示を受け付ける。指示受付部 4 3 は、例えば、ポインティングデバイスを通じてユーザから特徴量の選択指示を受け付けてもよい。なお、指示受付部 4 3 が受け付ける選択指示は、1 つの特徴量の選択であってもよく、複数の特徴量の選択であってもよい。また、該当する特徴量が存在しないとユーザに判断された場合、指示受付部 4 3 は、選択しないという指示を受け付けてもよい。

20

【 0 0 5 9 】

第二逆強化学習実行部 5 1 は、ユーザにより選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する。例えば、ユーザによって 1 つの特徴量が選択された場合、第二逆強化学習実行部 5 1 は、第一の実施形態の第二逆強化学習実行部 5 0 と同様の処理を行えばよい。また、例えば、複数の特徴量が選択された場合、第二逆強化学習実行部 5 1 は、複数の特徴量を（例えば、特徴量リスト B に）加えて、第二の目的関数を生成してもよい。なお、特徴量が選択されなかった場合、第二逆強化学習実行部 5 1 は、第二の目的関数を生成しなくてもよい。

30

【 0 0 6 0 】

入力部 2 0 と、第一逆強化学習実行部 3 0 と、特徴量選択部 4 1 と、特徴量提示部 4 2 と、指示受付部 4 3 と、第二逆強化学習実行部 5 1 と、情報量規準計算部 6 0 と、判定部 7 0 と、出力部 8 0 とは、プログラム（学習プログラム）に従って動作するコンピュータのプロセッサによって実現される。

【 0 0 6 1 】

次に、本実施形態の学習装置 2 0 0 の動作を説明する。図 5 は、本実施形態の学習装置 2 0 0 の動作例を示す説明図である。第一の目的関数を生成するまでのステップ S 1 1 からステップ S 1 2 の処理は、図 2 に例示する処理と同様である。以降、情報量規準が単調増加している間は、ステップ S 2 2 からステップ S 2 4 およびステップ S 1 5 からステップ S 1 7 の処理が繰り返される。すなわち、判定部 7 0 は、情報量基準が単調増加していると判断したときに、ステップ S 2 2 からステップ S 2 4 およびステップ S 1 5 からステップ S 1 7 の処理を繰り返し実行する制御を行う（ステップ S 2 1 ）。

40

【 0 0 6 2 】

特徴量選択部 4 1 は、Teaching Risk の小さい順に複数選択する（ステップ S 2 2 ）。特徴量提示部 4 2 は、特徴量選択部 4 1 が選択した特徴量をユーザに提示する（ステップ S 2 3 ）。そして、指示受付部 4 3 は、ユーザから特徴量の選択指示を受け付ける（ステップ S 2 4 ）。特徴量選択部 4 1 は、以降、図 2 に例示するステップ S 1 5 からステップ S 1 7 までの処理が行われる。その後、生成された目的関数に関する情報を出力するステップ S 1 8 の処理が行われる。

50

## 【 0 0 6 3 】

以上のように、本実施形態では、特徴量選択部 4 1 が、理想的な報酬の結果により近づく と推定される、予め定めた数の上位の特徴量を 1 つ以上選択し、特徴量提示部 4 2 が、選択された一つ以上の特徴量をユーザに提示する。そして、指示受付部 4 3 が、提示された特徴量に対するユーザからの選択の指示を受け付け、第二逆強化学習実行部 5 1 は、ユーザにより選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する。

## 【 0 0 6 4 】

よって、第一の実施形態の効果に加え、熟練者を含むユーザの知見を反映した学習を効率的に進めることが可能になる。

## 【 0 0 6 5 】

次に、本発明の概要を説明する。図 6 は、本発明による学習装置の概要を示すブロック図である。本発明による学習装置 9 0 は、候補とする複数の（具体的には、すべての）特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる候補特徴量の各重み（例えば、 $w^*$ ）を導出する第一逆強化学習実行部 9 1（例えば、第一逆強化学習実行部 3 0）と、各重み（例えば、 $w^*$ ）が導出された候補特徴量から一つの特徴量を選択した場合に、その特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づく と推定される特徴量を選択する特徴量選択部 9 2（例えば、特徴量選択部 4 0）と、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行部 9 3（例えば、第二逆強化学習実行部 5 0）とを備えている。

## 【 0 0 6 6 】

そのような構成により、逆強化学習で用いる目的関数の特徴量の選択を支援できる。

## 【 0 0 6 7 】

また、特徴量選択部 9 2 は、導出された候補特徴量の各重み（例えば、 $w^*$ ）を最適なパラメータとみなし、候補特徴量の中から、目的関数の部分最適性（例えば、Teaching Risk）を最小にする特徴量を選択してもよい。

## 【 0 0 6 8 】

また、学習装置 9 0 は、第二の目的関数の学習結果に基づいて、候補特徴量の中から、さらに特徴量を選択するか否か判定する判定部 9 4（例えば、判定部 7 0）を備えていてもよい。そして、特徴量選択部 9 2 は、さらに特徴量を選択すると判定された場合、候補特徴量の中から、既に選択された特徴量以外の特徴量を新たに選択し、第二逆強化学習実行部 9 3 は、新たに選択された特徴量を加えて逆強化学習を実行することにより、第二の目的関数を生成してもよい。

## 【 0 0 6 9 】

また、学習装置 9 0 は、生成された第二の目的関数の情報量規準を計算する情報量規準計算部（例えば、情報量規準計算部 6 0）を備えていてもよい。そして、判定部 9 4 は、情報量規準に基づいて、候補特徴量の中からさらに特徴量を選択するか否か判定してもよい。そのような構成により、特徴量の数とフィッティングのトレードオフを実現できる。

## 【 0 0 7 0 】

具体的には、判定部 9 4 は、情報量規準が単調増加する場合に、候補特徴量の中から、さらに特徴量を選択すると判定してもよい。

## 【 0 0 7 1 】

また、学習装置 9 0 は、情報量規準が最大になったときの第二の目的関数に含まれる特徴量および対応する特徴量の重みを出力する出力部 9 5（例えば、出力部 8 0）を備えていてもよい。

## 【 0 0 7 2 】

さらに、出力部 9 5 は、特徴量選択部 9 2 によって選択された順に特徴量を出力してもよい。

## 【 0 0 7 3 】

また、学習装置 9 0（例えば、学習装置 2 0 0）は、特徴量選択部 9 2 が選択した特徴量をユーザに提示する特徴量提示部（例えば、特徴量提示部 4 2）と、提示された特徴量

10

20

30

40

50

に対するユーザからの選択の指示を受け付ける指示受付部（例えば、指示受付部 43）とを備えていてもよい。そして、特徴量選択部 92 は、理想的な報酬の結果により近づく と推定される、予め定めた数の上位の特徴量を 1 つ以上選択し、特徴量提示部は、選択された一つ以上の特徴量をユーザに提示し、第二逆強化学習実行部 93 は、ユーザにより選択された特徴量を用いた逆強化学習により、第二の目的関数を生成してもよい。

【0074】

図 7 は、少なくとも 1 つの実施形態に係るコンピュータの構成を示す概略ブロック図である。コンピュータ 1000 は、プロセッサ 1001、主記憶装置 1002、補助記憶装置 1003、インタフェース 1004 を備える。

【0075】

上述の学習装置 90 は、コンピュータ 1000 に実装される。そして、上述した各処理部の動作は、プログラム（学習プログラム）の形式で補助記憶装置 1003 に記憶されている。プロセッサ 1001 は、プログラムを補助記憶装置 1003 から読み出して主記憶装置 1002 に展開し、当該プログラムに従って上記処理を実行する。

【0076】

なお、少なくとも 1 つの実施形態において、補助記憶装置 1003 は、一時的でない有形の媒体の一例である。一時的でない有形の媒体の他の例としては、インタフェース 1004 を介して接続される磁気ディスク、光磁気ディスク、CD-ROM（Compact Disc Read-only memory）、DVD-ROM（Read-only memory）、半導体メモリ等が挙げられる。また、このプログラムが通信回線によってコンピュータ 1000 に配信される場合、配信を受けたコンピュータ 1000 が当該プログラムを主記憶装置 1002 に展開し、上記処理を実行してもよい。

【0077】

また、当該プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、当該プログラムは、前述した機能を補助記憶装置 1003 に既に記憶されている他のプログラムとの組み合わせで実現するもの、いわゆる差分ファイル（差分プログラム）であってもよい。

【0078】

上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。

【0079】

（付記 1）候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる前記候補特徴量の各重みを導出する第一逆強化学習実行部と、前記各重みが導出された候補特徴量から一つの特徴量を選択した場合に、当該特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づく と推定される特徴量を選択する特徴量選択部と、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行部とを備えたことを特徴とする学習装置。

【0080】

（付記 2）特徴量選択部は、導出された候補特徴量の各重みを最適なパラメータとみなし、候補特徴量の中から、目的関数の部分最適性を最小にする特徴量を選択する付記 1 記載の学習装置。

【0081】

（付記 3）第二の目的関数の学習結果に基づいて、候補特徴量の中から、さらに特徴量を選択するか否か判定する判定部を備え、特徴量選択部は、さらに特徴量を選択すると判定された場合、前記候補特徴量の中から、既に選択された特徴量以外の特徴量を新たに選択し、第二逆強化学習実行部は、新たに選択された特徴量を加えて逆強化学習を実行することにより、第二の目的関数を生成する付記 1 または付記 2 記載の学習装置。

【0082】

（付記 4）生成された第二の目的関数の情報量規準を計算する情報量規準計算部を備え、判定部は、前記情報量規準に基づいて、候補特徴量の中からさらに特徴量を選択するか

10

20

30

40

50

否か判定する付記 3 記載の学習装置。

【 0 0 8 3 】

( 付記 5 ) 判定部は、情報量規準が単調増加する場合に、候補特徴量の中から、さらに特徴量を選択すると判定する付記 3 記載の学習装置。

【 0 0 8 4 】

( 付記 6 ) 情報量規準が最大になったときの第二の目的関数に含まれる特徴量および対応する特徴量の重みを出力する出力部を備えた付記 1 から付記 5 のうちのいずれか 1 つに記載の学習装置。

【 0 0 8 5 】

( 付記 7 ) 出力部は、特徴量選択部によって選択された順に特徴量を出力する付記 6 記載の学習装置。

10

【 0 0 8 6 】

( 付記 8 ) 特徴量選択部が選択した特徴量をユーザに提示する特徴量提示部と、提示された前記特徴量に対するユーザからの選択の指示を受け付ける指示受付部とを備え、特徴量選択部は、理想的な報酬の結果により近づくことと推定される、予め定めた数の上位の特徴量を 1 つ以上選択し、前記特徴量提示部は、選択された一つ以上の特徴量をユーザに提示し、第二逆強化学習実行部は、ユーザにより選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する付記 1 から付記 7 のうちのいずれか 1 つに記載の学習装置。

【 0 0 8 7 】

( 付記 9 ) 候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる前記候補特徴量の各重みを導出し、前記各重みが導出された候補特徴量から一つの特徴量を選択した場合に、当該特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくことと推定される特徴量を選択し、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成することを特徴とする学習方法。

20

【 0 0 8 8 】

( 付記 1 0 ) 導出された候補特徴量の各重みを最適なパラメータとみなし、候補特徴量の中から、目的関数の部分最適性を最小にする特徴量を選択する付記 9 記載の学習方法。

【 0 0 8 9 】

( 付記 1 1 ) コンピュータに、候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる前記候補特徴量の各重みを導出する第一逆強化学習実行処理、前記各重みが導出された候補特徴量から一つの特徴量を選択した場合に、当該特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくことと推定される特徴量を選択する特徴量選択処理、および、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行処理を実行させるための学習プログラムを記憶するプログラム記憶媒体。

30

【 0 0 9 0 】

( 付記 1 2 ) コンピュータに、特徴量選択処理で、導出された候補特徴量の各重みを最適なパラメータとみなし、候補特徴量の中から、目的関数の部分最適性を最小にする特徴量を選択させるための学習プログラムを記憶する付記 1 1 記載のプログラム記憶媒体。

【 0 0 9 1 】

( 付記 1 3 ) コンピュータに、候補とする複数の特徴量である候補特徴量を用いた逆強化学習により、第一の目的関数に含まれる前記候補特徴量の各重みを導出する第一逆強化学習実行処理、前記各重みが導出された候補特徴量から一つの特徴量を選択した場合に、当該特徴量を用いて表現される報酬が、理想的な報酬の結果に最も近づくことと推定される特徴量を選択する特徴量選択処理、および、選択された特徴量を用いた逆強化学習により、第二の目的関数を生成する第二逆強化学習実行処理を実行させるための学習プログラム。

40

【 0 0 9 2 】

( 付記 1 4 ) コンピュータに、特徴量選択処理で、導出された候補特徴量の各重みを最適なパラメータとみなし、候補特徴量の中から、目的関数の部分最適性を最小にする特徴量を選択させる付記 1 2 記載の学習プログラム。

50

【 0 0 9 3 】

以上、実施形態を参照して本願発明を説明したが、本願発明は上記実施形態に限定されるものではない。本願発明の構成や詳細には、本願発明のスコープ内で当業者が理解し得る様々な変更をすることができる。

【符号の説明】

【 0 0 9 4 】

- 1 0 記憶部
- 2 0 入力部
- 3 0 第一逆強化学習実行部
- 4 0 , 4 1 特徴量選択部 10
- 4 2 特徴量提示部
- 4 3 指示受付部
- 5 0 , 5 1 第二逆強化学習実行部
- 6 0 情報量規準計算部
- 7 0 判定部
- 8 0 出力部
- 1 0 0 , 2 0 0 学習装置

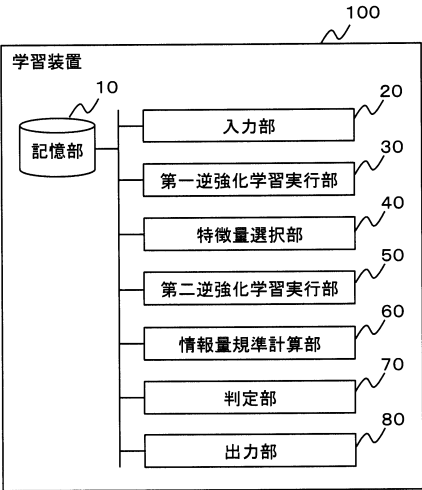
20

30

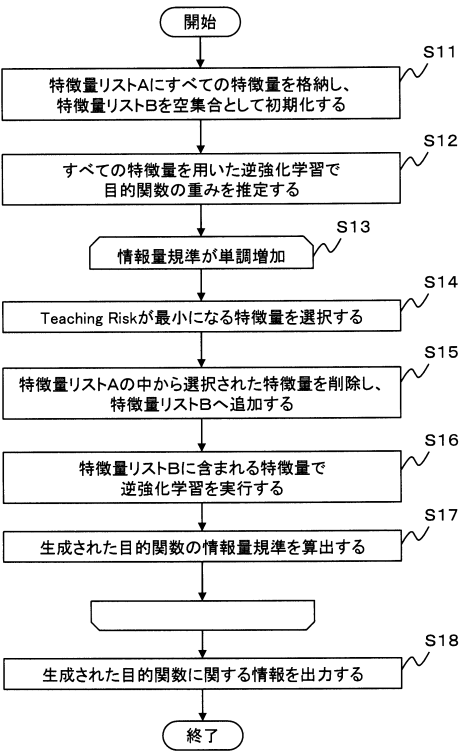
40

50

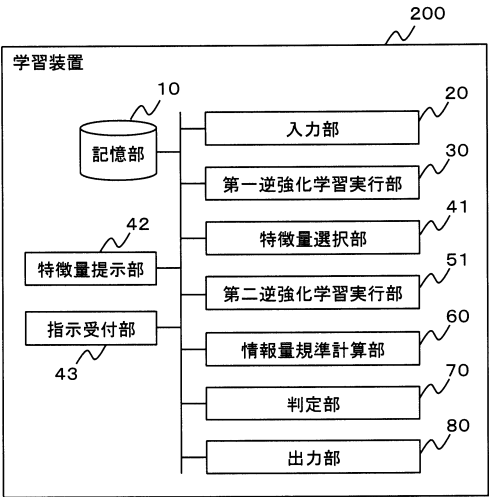
【図面】  
【図 1】



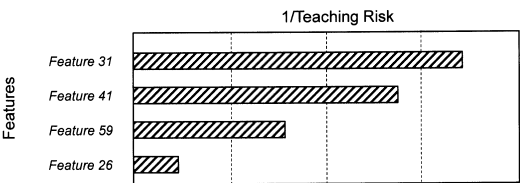
【図 2】



【図 3】



【図 4】



10

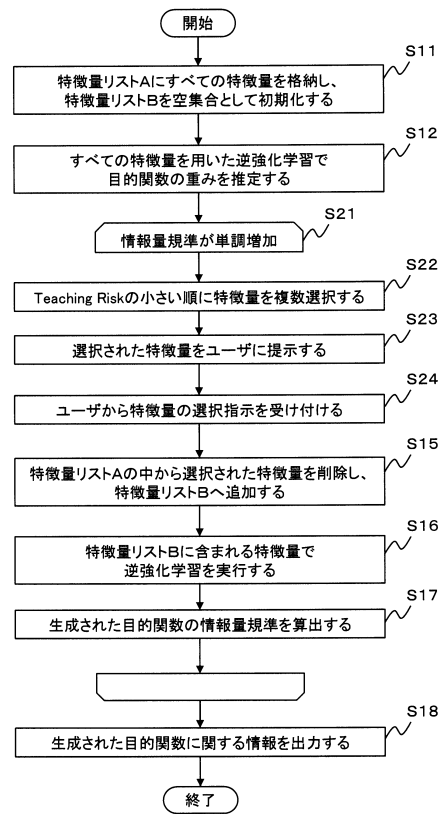
20

30

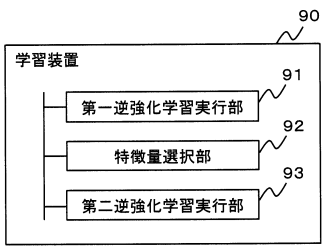
40

50

【図 5】



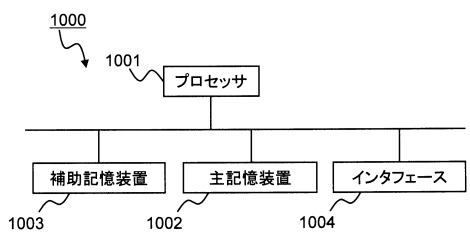
【図 6】



10

20

【図 7】



30

40

50

---

フロントページの続き

- (56)参考文献 中国特許出願公開第 1 1 1 4 0 1 5 5 6 ( C N , A )  
特開 2 0 2 0 - 1 1 9 3 8 4 ( J P , A )  
HAUG, Luis et al. , "Teaching Inverse Reinforcement Learners via Features and Demonstrations" , arXiv.org [online] , arXiv:1810.08926v4 , 米国 , Cornell University , 2019年 , pp.1-13 , [検索日 2020.11.04], インターネット: URL: <https://arxiv.org/pdf/1810.08926v4>
- (58)調査した分野 (Int.Cl. , D B 名)  
G 0 6 N 3 / 0 0 - 2 0 / 0 0  
G 0 6 F 1 8 / 2 1 1 - 1 8 / 2 1 1 5