

(19)日本国特許庁(JP)

(12)公開特許公報(A)

(11)特許出願公開番号

特開2022-134563
(P2022-134563A)

(43)公開日

令和4年9月15日(2022. 9. 15)

(51)Int. Cl.	F I	テーマコード (参考)
G 0 6 N 5/04 (2006. 01)	G 0 6 N 5/04	5 L 0 4 9
G 0 6 N 7/00 (2006. 01)	G 0 6 N 7/00	
G 0 6 Q 10/04 (2012. 01)	G 0 6 Q 10/04	

審査請求 未請求 請求項の数 7 O L (全 19 頁)

(21)出願番号	特願2021-33756(P2021-33756)	(71)出願人	000005223 富士通株式会社 神奈川県川崎市中原区上小田中4丁目1番1号
(22)出願日	令和3年3月3日(2021. 3. 3)	(74)代理人	100092978 弁理士 真田 有
		(74)代理人	100189201 弁理士 横田 功
		(72)発明者	渡邊 俊一 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
		Fターム(参考)	5L049 AA04

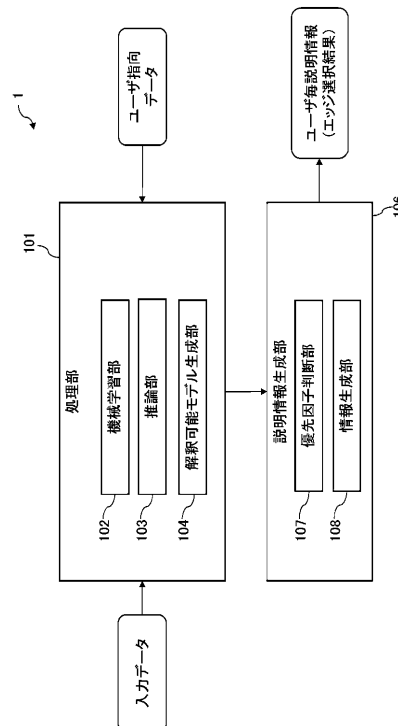
(54)【発明の名称】説明情報出力プログラム、説明情報出力方法および説明情報出力装置

(57)【要約】

【課題】機械学習モデルの予測結果について適切な説明情報を出力できるようにする。

【解決手段】それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予測結果と複数のデータとに基づいて生成された説明用のモデルを用いて、予測結果に対する複数のデータのそれぞれの寄与度を算出し、複数の変数のうち特定の変数を選択し、複数のデータのそれぞれの特定の変数の値と、寄与度とに基づいて、複数のデータのうちの特定のデータを決定し、特定のデータを前記予測結果の説明情報として出力することで、機械学習モデルの予測結果について適切な説明情報を出力する。

【選択図】図1



【特許請求の範囲】**【請求項 1】**

それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予測結果と前記複数のデータとに基づいて生成された説明用のモデルを用いて、前記予測結果に対する前記複数のデータのそれぞれの寄与度を算出し、

前記複数の変数のうち特定の変数を選択し、前記複数のデータのそれぞれの前記特定の変数の値と、前記寄与度とに基づいて、前記複数のデータのうち特定のデータを決定し、前記特定のデータを前記予測結果の説明情報として出力する、
処理をコンピュータに実行させることを特徴とする説明情報出力プログラム。

【請求項 2】

前記特定の変数を選択する処理は、ユーザに応じた優先変数に基づいて、前記複数の変数のうち前記特定の変数を選択する処理を含む、

ことを特徴とする請求項 1 記載の説明情報出力プログラム。

【請求項 3】

前記特定のデータを決定する処理は、前記特定の変数が数値データである場合、前記複数のデータのうち、前記寄与度が寄与度閾値よりも大きい複数の提示候補データの中から、前記特定の変数の値と前記寄与度とを用いて前記特定のデータを決定する処理を含む、
ことを特徴とする請求項 1 に記載の説明情報出力プログラム。

【請求項 4】

前記特定のデータを決定する処理は、前記複数の提示候補データから抽出した前記特定の変数の値を正規化した値と、前記複数の提示候補データそれぞれの前記寄与度を正規化した値との和が最も大きいデータを、前記特定のデータに決定する処理を含む、

ことを特徴とする請求項 3 に記載の説明情報出力プログラム。

【請求項 5】

前記特定のデータを決定する処理は、前記特定の変数がカテゴリカルデータである場合、前記複数のデータのうち、前記特定の変数の値が特定のカテゴリを示す複数の提示候補データであって前記寄与度が寄与度閾値よりも大きい前記複数の提示候補データの中から、前記寄与度が最も高いデータを、前記特定のデータに決定する

処理を前記コンピュータに実行させることを特徴とする請求項 1 に記載の説明情報出力プログラム。

【請求項 6】

それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予測結果と前記複数のデータとに基づいて生成された説明用のモデルを用いて、前記予測結果に対する前記複数のデータのそれぞれの寄与度を算出し、

前記複数の変数のうち特定の変数を選択し、前記複数のデータのそれぞれの前記特定の変数の値と、前記寄与度とに基づいて、前記複数のデータのうち特定のデータを決定し、前記特定のデータを前記予測結果の説明情報として出力する、
処理をコンピュータが実行することを特徴とする説明情報出力方法。

【請求項 7】

それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予測結果と前記複数のデータとに基づいて生成された説明用のモデルを用いて、前記予測結果に対する前記複数のデータのそれぞれの寄与度を算出し、

前記複数の変数のうち特定の変数を選択し、前記複数のデータのそれぞれの前記特定の変数の値と、前記寄与度とに基づいて、前記複数のデータのうち特定のデータを決定し、前記特定のデータを前記予測結果の説明情報として出力する、
処理部を有することを特徴とする説明情報出力装置。

【発明の詳細な説明】**【技術分野】****【0001】**

本発明は、説明情報出力プログラム、説明情報出力方法および説明情報出力装置に関する

10

20

30

40

る。

【背景技術】

【0002】

DL (Deep Learning) を中心とした複雑な機械学習の予測モデルはブラックボックス性が高く、ビジネスにおける実運用の大きな障害となっている。

【0003】

そこで、近年、機械学習モデルにおいて、予測結果や推定結果に至るプロセスについて人に説明可能な状態とする技術として、XAI (Explainable Artificial Intelligence : 説明可能なAI) が知られている。説明可能な状態を、ユーザが納得できる状態といってもよい。

10

【0004】

XAIにおいて、あるサンプルの予測について機械学習モデルがどのような根拠でその予測を行なったかを解釈するツールとして、例えば、LIMEやSHAPが知られている。

また、LIMEやSHAP等のモデルが何を重視しているか明らかにする従来技術として、因子単位における寄与度算出法が知られている。

寄与度算出法においては、複数の因子のそれぞれにモデルの出力結果への寄与度を算出し、寄与度の高い因子を提示する。

図9はSHAPによる出力例を示す図である。

【0005】

この図9に示す例においては、モデルのとある出力に対する複数の因子の各寄与度(重要度)をグラフで表している。この図9に示す例においては因子“LSTAT”が予測タスクに対して一番貢献したことがわかる。

20

【先行技術文献】

【特許文献】

【0006】

【特許文献1】特開2010-165166号公報

【特許文献2】特開2019-82883号公報

【発明の概要】

【発明が解決しようとする課題】

【0007】

しかしながら、XAIの実運用においては、単一の説明でユーザ全体が説明可能な状態になるのは稀である。

図10は、XAIにおいて説明可能な状態になるための条件を説明するための図である。

【0008】

この図10においては、図9に示した例において、「とあるモデル出力に関して因子“LSTAT”の寄与度が最も高い」との説明情報の複数ユーザによる受け止められ方を示す。

【0009】

「とあるモデル出力に因子“LSTAT”の寄与度が高い」ことは、属人性が高く、A氏にとって当たり前(比較的普遍)であるとする。このような場合には、当該説明情報はA氏にとって何ら有益な説明になっていない。XAIにおける説明情報は、当たりまえの壁を越えることが求められる。

40

【0010】

一方、上記の「とあるモデル出力に関して因子“LSTAT”の寄与度が最も高い」との説明情報は、B氏の知識知見とは違うとする。このような場合には、上記の説明情報はB氏に理解できない。XAIにおける説明情報は、知識知見の壁を越えることも求められる。

【0011】

また、上記の「とあるモデル出力に関して因子“LSTAT”の寄与度が最も高い」との説明情報は、C氏の知識知見に合うとする。このような場合には、上記の説明情報は、C氏

50

にとって自身の感覚と似ていて納得できるものとして受け入れられる。

【 0 0 1 2 】

このように、説明可能な状態は、属人性が高くユーザ毎に異なる場合がある。そのため、X A Iにおける従来の情報提示手法においては、説明情報が各ユーザの知識知見に触れた説明・因子になっていない場合には、説明情報がユーザに説明可能な状態とならないおそれがある。

【 0 0 1 3 】

X A Iにおける従来の情報提示手法においては、複数の属性を含むデータ（変数）ごとにモデルの出力結果への寄与度を算出し、寄与度の高いデータを提示する。しかしながら、解釈時に重要視される属性は時と場合による（人依存も含む）ため、適切な因子提示とならない場合がある。

1つの側面では、本発明は、機械学習モデルの予測結果について適切な説明情報を出力できるようにすることを目的とする。

【課題を解決するための手段】

【 0 0 1 4 】

このため、この説明情報出力プログラムは、それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予測結果と前記複数のデータとに基づいて生成された説明用のモデルを用いて、前記予測結果に対する前記複数のデータのそれぞれの寄与度を算出し、前記複数の変数のうち特定の変数を選択し、前記複数のデータのそれぞれの前記特定の変数の値と、前記寄与度とに基づいて、前記複数のデータのうち特定のデータを決定し、前記特定のデータを前記予測結果の説明情報として出力する処理をコンピュータに実行させる。

【発明の効果】

【 0 0 1 5 】

一実施形態によれば、機械学習モデルの予測結果について適切な説明情報を出力できる。

【図面の簡単な説明】

【 0 0 1 6 】

【図 1】実施形態の一例としての情報処理装置の構成を模式的に示す図である。

【図 2】実施形態の一例としての情報処理装置のハードウェア構成を例示する図である。

【図 3】実施形態の一例としての情報処理装置における入力データを例示する図である。

【図 4】実施形態の一例としての情報処理装置におけるユーザ指向データ入力画面を例示する図である。

【図 5】実施形態の一例としての情報処理装置の解釈可能モデル生成部が生成する寄与度を説明するための図である。

【図 6】実施形態の一例としての情報処理装置における、カテゴリカルデータ時の提示エッジの選択方法を説明するための図である。

【図 7】実施形態の一例としての情報処理装置における、数値データ時の提示エッジの選択方法を説明するための図である。

【図 8】実施形態の一例としての情報処理装置における処理を説明するためのフローチャートである。

【図 9】S H A Pによる出力例を示す図である。

【図 10】X A Iにおいて説明可能な状態になるための条件を説明するための図である。

【発明を実施するための形態】

【 0 0 1 7 】

以下、図面を参照して本説明情報出力プログラム、説明情報出力方法および説明情報出力装置にかかる実施の形態を説明する。ただし、以下に示す実施形態はあくまでも例示に過ぎず、実施形態で明示しない種々の変形例や技術の適用を排除する意図はない。すなわち、本実施形態を、その趣旨を逸脱しない範囲で種々変形して実施することができる。また、各図は、図中に示す構成要素のみを備えるという趣旨ではなく、他の機能等を含むこ

10

20

30

40

50

とができる。

【0018】

(A)構成

図1は実施形態の一例としての情報処理装置1の構成を模式的に示す図である。

情報処理装置1は、入力データに対して機械学習モデルによる推論を行なうとともに、当該推論の根拠を示す説明情報をユーザに提示するXAIを実現する説明情報出力装置(コンピュータ)である。

図2は実施形態の一例としての情報処理装置1のハードウェア構成を例示する図である。

【0019】

情報処理装置1は、例えば、プロセッサ11、メモリ12、記憶装置13、グラフィック処理装置14、入力インタフェース15、光学ドライブ装置16、機器接続インタフェース17およびネットワークインタフェース18を構成要素として有する。これらの構成要素11~18は、バス19を介して相互に通信可能に構成される。

【0020】

プロセッサ(処理部)11は、情報処理装置1全体を制御する。プロセッサ11は、マルチプロセッサであってもよい。プロセッサ11は、例えばCPU(Central Processing Unit)、MPU(Micro Processing Unit)、DSP(Digital Signal Processor)、ASIC(Application Specific Integrated Circuit)、PLD(Programmable Logic Device)、FPGA(Field Programmable Gate Array)のいずれか一つであってもよい。また、プロセッサ11は、CPU、MPU、DSP、ASIC、PLD、FPGAのうちの2種類以上の要素の組み合わせであってもよい。

【0021】

そして、プロセッサ11が情報処理装置1用の制御プログラム(説明情報出力プログラム:図示省略)を実行することにより、図1に例示する、処理部101および説明情報生成部106としての機能が実現される。これにより、情報処理装置1は、説明情報出力装置として機能する。

【0022】

なお、情報処理装置1は、例えばコンピュータ読み取り可能な非一時的な記録媒体に記録されたプログラム{説明情報出力プログラム、OS(Operating System)プログラム}を実行することにより、処理部101および説明情報生成部106としての機能を実現する。

【0023】

情報処理装置1に実行させる処理内容を記述したプログラムは、様々な記録媒体に記録しておくことができる。例えば、情報処理装置1に実行させるプログラムを記憶装置13に格納しておくことができる。プロセッサ11は、記憶装置13内のプログラムの少なくとも一部をメモリ12にロードし、ロードしたプログラムを実行する。

【0024】

また、情報処理装置1(プロセッサ11)に実行させるプログラムを、光ディスク16a、メモリ装置17a、メモ리카ード17c等の非一時的な可搬型記録媒体に記録しておくこともできる。可搬型記録媒体に格納されたプログラムは、例えばプロセッサ11からの制御により、記憶装置13にインストールされた後、実行可能になる。また、プロセッサ11が、可搬型記録媒体から直接プログラムを読み出して実行することもできる。

【0025】

メモリ12は、ROM(Read Only Memory)およびRAM(Random Access Memory)を含む記憶メモリである。メモリ12のRAMは情報処理装置1の主記憶装置として使用される。RAMには、プロセッサ11に実行させるプログラムの少なくとも一部が一時的に格納される。また、メモリ12には、プロセッサ11による処理に必要な各種データが格納される。

【0026】

10

20

30

40

50

記憶装置 1 3 は、ハードディスクドライブ (Hard Disk Drive : H D D)、S S D (Solid State Drive)、ストレージクラスメモリ (Storage Class Memory : S C M)等の記憶装置であって、種々のデータを格納するものである。記憶装置 1 3 は、情報処理装置 1 の補助記憶装置として使用される。記憶装置 1 3 には、O S プログラム、制御プログラムおよび各種データが格納される。制御プログラムには説明情報出力プログラムが含まれる。

【 0 0 2 7 】

なお、補助記憶装置としては、S C M やフラッシュメモリ等の半導体記憶装置を使用することもできる。また、複数の記憶装置 1 3 を用いて R A I D (Redundant Arrays of Inexpensive Disks) を構成してもよい。

また、記憶装置 1 3 には、上述した処理部 1 0 1 および説明情報生成部 1 0 6 が各処理を実行する際に生成される各種データを格納してもよい。

10

【 0 0 2 8 】

グラフィック処理装置 1 4 には、モニタ 1 4 a が接続されている。グラフィック処理装置 1 4 は、プロセッサ 1 1 からの命令に従って、画像をモニタ 1 4 a の画面に表示させる。モニタ 1 4 a としては、C R T (Cathode Ray Tube) を用いた表示装置や液晶表示装置等が挙げられる。

【 0 0 2 9 】

入力インタフェース 1 5 には、キーボード 1 5 a およびマウス 1 5 b が接続されている。入力インタフェース 1 5 は、キーボード 1 5 a やマウス 1 5 b から送られてくる信号をプロセッサ 1 1 に送信する。なお、マウス 1 5 b は、ポインティングデバイスの一例であり、他のポインティングデバイスを使用することもできる。他のポインティングデバイスとしては、タッチパネル、タブレット、タッチパッド、トラックボール等が挙げられる。

20

【 0 0 3 0 】

光学ドライブ装置 1 6 は、レーザ光等を利用して、光ディスク 1 6 a に記録されたデータの読み取りを行なう。光ディスク 1 6 a は、光の反射によって読み取り可能にデータを記録された可搬型の非一時的な記録媒体である。光ディスク 1 6 a には、D V D (Digital Versatile Disc)、D V D - R A M、C D - R O M (Compact Disc Read Only Memory)、C D - R (Recordable) / R W (ReWritable) 等が挙げられる。

【 0 0 3 1 】

機器接続インタフェース 1 7 は、情報処理装置 1 に周辺機器を接続するための通信インタフェースである。例えば、機器接続インタフェース 1 7 には、メモリ装置 1 7 a やメモリリーダライタ 1 7 b を接続することができる。メモリ装置 1 7 a は、機器接続インタフェース 1 7 との通信機能を搭載した非一時的な記録媒体、例えば U S B (Universal Serial Bus) メモリである。メモリリーダライタ 1 7 b は、メモリカード 1 7 c へのデータの書き込み、またはメモリカード 1 7 c からのデータの読み出しを行なう。メモリカード 1 7 c は、カード型の非一時的な記録媒体である。

30

【 0 0 3 2 】

ネットワークインタフェース 1 8 は、ネットワークに接続される。ネットワークインタフェース 1 8 は、ネットワークを介してデータの送受信を行なう。ネットワークには他の情報処理装置や通信機器等が接続されてもよい。

40

図 1 に例示する情報処理装置 1 は、処理部 1 0 1 および説明情報生成部 1 0 6 を備える。

処理部 1 0 1 は、入力データを受け付け、この入力データに対して機械学習モデルを用いた処理を行なう。

図 3 は実施形態の一例としての情報処理装置 1 における入力データを例示する図である。

【 0 0 3 3 】

本実施形態においては、情報処理装置 1 が、健康経営の領域において用いられ、出勤データから将来の休職の有無 (休職するか否か) の二値分類を機械学習モデルを用いて行なうタスクを例に説明する。すなわち、機械学習モデルに、予測対象者の出勤データを入力

50

して、当該予測対象者が将来的に休職するか否かの予測を行なわせる。

図3に例示する入力データは、社員の出勤データを示すテンソルデータであり、日付、出欠、出張の有無（出張）および残業時間（残業）を備える。

【0034】

図3に例示する入力データにおいて、日付、出欠、出張および残業の組み合わせをエッジという場合がある。また、エッジを構成する変数（出張、残業）を因子という場合がある。エッジは複合的な因子関係であるといえる。

入力データは、複数の変数（因子：出張、残業）を含む複数のデータ（出勤データ、エッジ）に相当する。

【0035】

また、処理部101には、ユーザ指向データが入力される。ユーザ指向データは、エッジを構成する複数種類の因子のうち、機械学習モデルによる予測に寄与するとユーザが考える（着目する）因子を示す情報である。

ユーザ指向データは、例えば、図4に例示する入力画面等を用いて予めユーザに入力させてもよい。

図4は実施形態の一例としての情報処理装置1におけるユーザ指向データ入力画面を例示する図である。

【0036】

ユーザ指向データ入力画面は、本情報処理装置1に備えられたモニタ14a等に表示される。ユーザは、このユーザ指向データ入力画面を介して、自身のユーザ指向データを入力する。

【0037】

この図4に例示するユーザ指向データ入力画面には、「以下に示す出勤管理項目の中から、あなたが休職の有無に最も影響を与えると考えるものを1つ選択してください。」とのメッセージとともに、“出張の有無”および“残業時間”の2つの選択肢がトグルスイッチとして表示されている。これらの“出張の有無”および“残業時間”は、出張および残業の各因子に対応する。

【0038】

例えば、企業の人事や健康推進室に所属する複数のユーザ（A氏、B氏）は、それぞれ、このユーザ指向データ入力画面において、自身が休職の有無に最も影響を与えると考えるものを選択し、キーボード15aやマウス15bを操作して選択入力を行なう。

【0039】

ユーザが選択入力した内容は、ユーザ指向データとして、ユーザ毎に記憶装置13やメモリ12等の所定の記憶領域に記憶される。以下、ユーザ指向データを優先因子といってもよい。ユーザ指向データは、ユーザに応じた優先変数に相当する。

【0040】

図4に示した例においては、ユーザ指向データとして1つの因子を選択する例を示したが、これに限定されるものではない。ユーザは、ユーザ指向データとして2つ以上の因子を選択してもよく、適宜変更して実施することができる。ユーザが選択するユーザ指向データの数を優先因子数（N）という場合がある。優先因子数（N）はシステム管理者等が予め任意に設定してもよい。本実施形態においては、便宜上、優先因子数N=1の例について示す。

処理部101は、図1に示すように、機械学習部102、推論部103および解釈可能モデル生成部104を備える。

機械学習部102は、訓練データセットを用いて機械学習モデルを作成する。訓練データセットは、入力データと正解データ（教師データ）とを備える。機械学習モデルとしては、例えば、サポートベクターマシン、ニューラルネットワーク、勾配ブースティングツリーなどが使用されてもよい。

以下に示す例においては、主に表やグラフをテンソルで表現したものを入力データとして、機械学習部102が、既知の機械学習技術であるDeep Tensor（登録商標）のアルゴ

10

20

30

40

50

リズムを用いて、機械学習モデルを作成する例について示す。

【0041】

機械学習部102は、表やグラフをテンソル形式で取り扱う。表やグラフをテンソルデータといってもよい。機械学習部102は、入力データをテンソル分解によってコアテンソルに変換し、当該コアテンソルをニューラルネットワークへ入力して、入力データに対する推論結果を出力する機械学習モデルを機械学習によって作成する。

【0042】

なお、機械学習部102は、訓練データに基づいて、テンソル分解に用いるパラメータとニューラルネットワークのパラメータとを最適化することによって機械学習モデルを生成する。機械学習部102は、例えば、勾配降下法を用いて、訓練データに対する機械学習モデルの推論結果と正解データとの誤差を定義した損失関数を小さくする方向に、テンソル分解のパラメータとニューラルネットワークのパラメータをと更新することによって、パラメータの最適化を行なう。

10

【0043】

ニューラルネットワークは、ハードウェア回路であってもよいし、プロセッサ11によりコンピュータプログラム上で仮想的に構築される階層間を接続するソフトウェアによる仮想的なネットワークであってもよい。

【0044】

推論部103は、機械学習部102が作成した機械学習モデルを用いて推論を行なう。推論部103は、推論の対象のテンソルデータを機械学習モデルへ入力し、機械学習モデルが出力する推論結果を得る。

20

【0045】

解釈可能モデル生成部104は、機械学習部102が作成した機械学習モデルの推論結果を説明するため、入力データに含まれる各変数の寄与を説明しやすい機械学習モデルを局所的に近似した解釈可能モデルを生成する。

【0046】

解釈可能モデル生成部104は、例えば、LIMEやSHAP等の、機械学習モデルの予測（推論）について、どのような根拠でその予測を行なったかを解釈する既知の手法を用いて解釈可能モデルを生成する。解釈可能モデル生成部104は、例えば、機械学習モデルで推論を行なった説明対象のデータの推論結果と説明対象のデータの近傍データの推論結果とに対する重回帰分析に基づいて、機械学習モデルを局所的に近似した線形回帰モデルを生成しても良い。ここで、近傍データは説明対象のデータと類似しているデータであり、説明対象のデータとの差分が閾値以下のデータである。解釈可能モデル生成部104は、コアテンソルの類似度が閾値以上のデータを近傍データと決定としてよい。また、類似度は、テンソル間の距離（例えば、二乗誤差）で定義することができる。線形回帰モデルを、説明対象のデータのコアテンソルとその推論結果とに基づいて生成することで、解釈可能モデル生成部104は、コアテンソル類似度に基づいて解釈可能モデルを生成するといえる。

30

【0047】

解釈可能モデルは、生成した線形回帰モデルのパラメータに基づいて、説明対象のデータに含まれる各変数の推論結果に対する寄与度を特定する。解釈可能モデル生成部104は、例えば、それぞれが複数の変数（因子）を含む複数のデータ（出勤データ、エッジ）の入力に応じて機械学習モデルが出力した予測結果と複数のデータとに基づいて解釈可能モデルを生成しても良い。

40

この場合、解釈可能モデル生成部104は、解釈可能モデルに基づき、エッジ単位で寄与度を算出しても良い。線形回帰モデル偏回帰係数を寄与度としてもよい。

図5は実施形態の一例としての情報処理装置1の解釈可能モデル生成部104が生成する寄与度を説明するための図である。

【0048】

この図5に示す例においては、例えば、日付5/27で特定されるエッジに寄与度“0.6”

50

が設定されている。なお、寄与度の算出は既知の手法で実現することができ、その説明は省略する。

【 0 0 4 9 】

説明情報生成部 1 0 6 は、処理部 1 0 1 によって入力データに対して行なわれた推定の根拠を示す情報を生成する。説明情報生成部 1 0 6 は、処理部 1 0 1 による推定をユーザが理解・納得できるような説明情報（ユーザ毎説明情報）を生成する。

【 0 0 5 0 】

本情報処理装置 1 においては、個々のユーザの知識知見によりユーザ指向データがユーザ毎に異なることに鑑み、X A I における機械学習モデルによる予測（推論）の根拠を説明する情報（説明情報）をユーザ毎にカスタマイズして提示する。本情報処理装置 1 において、説明情報生成部 1 0 6 は、ユーザに対する説明情報として、入力データの中から選択したエッジを用いる。すなわち、説明情報生成部 1 0 6 は、推定に寄与（貢献）した情報（エッジ）を選択し、説明情報として提示する。

説明情報生成部 1 0 6 は、図 1 に示すように、優先因子判断部 1 0 7 および情報生成部 1 0 8 を備える。

【 0 0 5 1 】

優先因子判断部 1 0 7 は、個々のユーザの優先因子（ユーザ指向データ）を確認する。優先因子判断部 1 0 7 は、ユーザがユーザ指向データ入力画面を介して入力したユーザ指向データを記憶装置 1 3 やメモリ 1 2 等の所定の記憶領域から読み出し、その内容を確認する。

情報生成部 1 0 8 は、X A I における機械学習モデルによる予測（推論）の根拠を説明する情報（説明情報）をユーザ毎に生成する。

【 0 0 5 2 】

情報生成部 1 0 8 は、例えば、複数の入力データの中から、ユーザ毎に選択したエッジを説明情報として用いる。以下、ユーザに説明情報として提示するエッジを提示エッジという場合がある。提示エッジは、機械学習モデルが出力した予測結果の説明情報に相当する。

【 0 0 5 3 】

例えば、情報生成部 1 0 8 は、入力データを構成する複数のエッジの中から一部のエッジを提示エッジとして選択し、この提示エッジのみを説明情報としてユーザに提示してもよい。また、情報生成部 1 0 8 は、提示エッジとして複数のエッジを選択し、これらの複数の提示エッジに優先順位を設定し、説明情報として、複数の提示エッジを、優先順位の高いものから順に優先的に並べてユーザに提示してもよい。

情報生成部 1 0 8 は、各ユーザのユーザ指向データに基づいて提示エッジの選択を行なう。

【 0 0 5 4 】

また、情報生成部 1 0 8 は、提示エッジの選択に際して、解釈可能モデル生成部 1 0 4 によって設定されたエッジ毎の寄与度を、予め設定された寄与度閾値と比較する。情報生成部 1 0 8 は、寄与度が寄与度閾値よりも大きいエッジを提示エッジの候補（提示候補エッジ）として選択する。

なお、寄与度閾値は、システム管理者等が予め設定してもよい。また、寄与度閾値は、運用に伴う過去の寄与度に基づき統計的に決定してもよい。

【 0 0 5 5 】

情報生成部 1 0 8 は、優先因子判断部 1 0 7 により判定された優先因子のデータ形式に応じた手法で説明情報を生成する。情報生成部 1 0 8 は、優先因子のデータ形式が数値データである場合とカテゴリカルデータである場合とで異なる手法を用いて提示データを生成する。

【 0 0 5 6 】

ここで、数値データである優先因子とは、数値で表される因子であり、本実施形態においては、「残業時間」という数値で表される“残業”が相当する。一方、カテゴリカルデ

10

20

30

40

50

ータである優先因子とは、カテゴリの集まりの中のいずれかに分類されることで表される因子であり、本実施形態においては、「終日（あり）」か「なし」かのいずれかに分類される“出張”が相当する。

図6は実施形態の一例としての情報処理装置1における、カテゴリカルデータ時の提示エッジの選択方法を説明するための図である。

【0057】

情報生成部108は、優先因子のデータ形式がカテゴリカルデータである場合には、入力データ（エッジ群）の中から、寄与度が寄与度閾値よりも大きい1つ以上のエッジ（提示候補エッジ）を抽出する。そして、情報生成部108は、これらの抽出した提示候補エッジの中から優先因子に値が発生している（活性している）1つ以上のエッジ（カテゴリカルデータ発生エッジ）を特定する。提示候補エッジは提示候補データに相当する。情報生成部108は、抽出したこれらの提示候補エッジの中から寄与度が最も高いエッジを提示エッジとして選択する。

10

【0058】

情報生成部108は、複数の変数（因子）のうち特定の変数（優先因子と同じ変数）を選択し、複数のデータ（エッジ）のそれぞれの特定の変数の値と、寄与度とに基づいて、複数のエッジのうち特定のデータ（提示エッジ）を決定する。

すなわち、情報生成部108は、複数のデータ（エッジ群）のうち、特定の変数の値が優先因子と同じ特定のカテゴリを示す複数の提示候補データであって寄与度が寄与度閾値よりも大きい複数の提示候補データ（提示候補エッジ）を決定する。そして、情報生成部108は、これらの複数の提示候補エッジの中から、寄与度が最も高いデータを提示エッジ（特定のデータ）に決定する。

20

【0059】

以下に示す例においては、寄与度閾値が0.45である例を示す。B氏の優先因子（ユーザ指向データ）はカテゴリカルデータである“出張”であるものとする。

図6に示す例においては、日付が5/24および5/26の2つのエッジが、寄与度が寄与度閾値（0.45）よりも大きい提示候補エッジに相当する。

情報生成部108は、これらの2つの提示候補エッジの中から、カテゴリカルデータ発生エッジに相当する、日付が5/26のエッジ（符号P1参照）を提示エッジとして選択する。

30

【0060】

なお、情報生成部108は、寄与度が寄与度閾値よりも大きい複数の提示エッジを選択し、これらの複数の提示エッジを、寄与度が大きいものから順に並べてユーザに提示してもよい。

図7は実施形態の一例としての情報処理装置1における、数値データ時の提示エッジの選択方法を説明するための図である。

【0061】

情報生成部108は、優先因子のデータ形式が数値データである場合には、入力データ（エッジ群）の中から、寄与度が寄与度閾値よりも大きい1つ以上の提示候補エッジ（提示候補データ）を抽出する。そして、情報生成部108は、これらの抽出した提示候補エッジの中から優先因子に数値が発生している（活性している）1つ以上のエッジ（数値データ発生エッジ）を特定する。そして、情報生成部108は、各数値データ発生エッジにおける優先因子の各データと各寄与度との2つのパラメータを用いて提示エッジの選択を行なう。

40

以下に示す例において、A氏の優先因子（ユーザ指向データ）は数値データである“残業”であるものとする。

【0062】

図7に示す例においては、日付が5/24、5/25および5/26の3つのエッジがいずれも寄与度の値が寄与度閾値（0.45）よりも大きい提示候補エッジに相当する。

【0063】

50

情報生成部 108 は、これらの 3 つの提示候補エッジの中から、優先因子の残業の各値と、各寄与度の値をそれぞれ取得し、各エッジの残業の値と寄与度の値とをそれぞれ正規化し、エッジ毎にこれらの正規化した残業の値と寄与度の値との和（合計値）を算出する。

【0064】

図 7 に示す例において、情報生成部 108 は、提示候補エッジの各残業の値を、平均 0、分散 1 とする標準化を行なうことで、日付 5/24、5/25 および 5/26 の各残業の値 3、1 および 1 を、1.4、-0.7 および -0.7 にそれぞれ変換する。

【0065】

また、情報生成部 108 は、提示候補エッジの各寄与度の値を、平均 0、分散 1 とする標準化を行なうことで、日付 5/24、5/25 および 5/26 の各寄与度の値 0.6、0.5 および 0.7 を、0、-1.2 および 1.2 にそれぞれ変換する。

なお、情報生成部 108 は、各残業値や各寄与度の値の標準化（正規化）には、例えば、アフィン変換を用いてもよい。

【0066】

そして、情報生成部 108 は、提示候補エッジ毎に標準化後の残業および寄与度の値を合計して和を求める。図 7 に示す例においては、例えば、日付 5/24 の標準化後の残業および寄与度の各値は 1.4 と 0 であり、これらの合計値は 1.4 である。

【0067】

情報生成部 108 は、3 つの提示候補エッジの中から、合計値が最も高い、日付が 5/24 の提示候補エッジ（符号 P 2 参照）を選択する。情報生成部 108 は、選択した提示候補エッジを提示エッジとしてユーザに提示する（符号 P 3 参照）。

【0068】

なお、情報生成部 108 は、複数の提示候補エッジを選択し、これらの複数の提示候補エッジを、合計値が大きいものから順に並べてユーザに提示してもよい。

【0069】

このように、情報生成部 108 は、優先因子のデータ形式が数値データである場合には、寄与度の大小だけではなく、優先因子の数値データの大小も考慮して提示エッジの選択を行なうのである。

【0070】

（B）動作

上述の如く構成された実施形態の一例としての情報処理装置 1 における処理を、図 8 に示すフローチャート（ステップ S 1 ~ S 13）に従って説明する。

ステップ S 1 において、処理部 101 は、推論の対象の入力データを取得する。ステップ S 2 において、推論部 103 は、入力データを機械学習部 102 が作成した機械学習モデルへ入力し、機械学習モデルが出力する推論結果を得る。

ステップ S 3 において、解釈可能モデル生成部 104 は、機械学習部 102 が生成したコアテンソル（特徴量）に基づき、解釈可能モデルを生成する。

【0071】

解釈可能モデル生成部 104 は、コアテンソルの類似度に基づいて近傍データを決定する。解釈可能モデル生成部 104 は、近傍データとその推論結果とを用いて線形回帰モデルを作成する。解釈可能モデルは、生成した線形回帰モデルのパラメータに基づいて、説明対象のデータに含まれる各変数の推論結果に対する寄与度を特定する。解釈可能モデル生成部 104 は、例えば、それぞれが複数の変数（因子）を含む複数のデータ（出勤データ、エッジ）の入力に応じて機械学習モデルが出力した予測結果と複数のデータとに基づいて解釈可能モデルを生成する。

ステップ S 4 において、処理部 101 は、例えば、ユーザが予め入力等した寄与度閾値を設定する。

【0072】

また、本情報処理装置 1 の各ユーザは、ユーザ指向データ入力画面を用いて入力する。

ステップ S 5 において、処理部 1 0 1 は、本情報処理装置 1 を用いる複数のユーザ（対象ユーザ）毎にユーザ指向データ（優先因子）を取得する。

ステップ S 6 において、情報生成部 1 0 8 は、全ユーザに対して説明情報の生成を行なったかを確認する。

【 0 0 7 3 】

確認の結果、全ユーザに対する説明情報の生成を行っていない場合には（ステップ S 6 の N O ルート参照）、ステップ S 7 に移行する。以下のステップ S 7 ~ S 1 2 の処理は、ユーザ毎に実施される。

【 0 0 7 4 】

ステップ S 7 において、優先因子判断部 1 0 7 はユーザの優先因子を確認する。説明情報生成部 1 0 6 は、入力データにおいて、ユーザの優先因子が活性している 1 つ以上のエッジ（エッジ群）を取得する。

10

【 0 0 7 5 】

ステップ S 8 において、情報生成部 1 0 8 は、ステップ S 7 において取得したエッジ群において、寄与度が寄与度閾値を越えるエッジ（提示候補エッジ）があるかを確認する。確認の結果、寄与度が寄与度閾値よりも大きいエッジがない場合には（ステップ S 8 の N O ルート参照）、ステップ S 9 に移行する。ステップ S 9 において、情報生成部 1 0 8 は、解釈可能モデル生成部 1 0 4 が L I M E や S H A P 等の既知のツールを用いて生成した解釈可能モデルの出力結果を変更せずに、そのまま説明情報として設定する。その後、処理はステップ S 6 に戻る。

20

【 0 0 7 6 】

また、ステップ S 8 における確認の結果、寄与度が寄与度閾値よりも大きいエッジがある場合には（ステップ S 8 の Y E S ルート参照）、ステップ S 1 0 に移行する。ステップ S 1 0 において、情報生成部 1 0 8 は、優先因子のデータ形式が数値データであるか確認する。

確認の結果、優先因子のデータ形式が数値データである場合には（ステップ S 1 0 の Y E S ルート参照）、ステップ S 1 2 に移行する。

【 0 0 7 7 】

ステップ S 1 2 において、情報生成部 1 0 8 は、複数の提示候補エッジにおける優先因子の各数値と各寄与度との 2 つのパラメータに基づいて、提示エッジの選択または複数の提示候補エッジの表示順序の変更を行なう。その後、処理はステップ S 6 に戻る。

30

【 0 0 7 8 】

また、ステップ S 1 0 における確認の結果、優先因子のデータ形式が数値データでない場合、すなわち、カテゴリカルデータである場合には（ステップ S 1 0 の N O ルート参照）、ステップ S 1 1 に移行する。

【 0 0 7 9 】

ステップ S 1 1 において、情報生成部 1 0 8 は、複数の提示候補エッジの中から寄与度が最も高いエッジを提示エッジとして選択する。または、情報生成部 1 0 8 は、寄与度の値が大きいものから順となるように複数の提示候補エッジの表示順序の変更を行なう。その後、処理はステップ S 6 に戻る。

40

【 0 0 8 0 】

また、ステップ S 6 における確認の結果、全ユーザに対する説明情報の生成を行なった場合には（ステップ S 6 の Y E S ルート参照）、ステップ S 1 3 に移行する。ステップ S 1 3 において、説明情報生成部 1 0 6 は、ユーザ毎に選択した提示エッジや、ユーザの優先因子に合わせた順序で並べた複数の提示エッジを説明情報として出力する。その後、処理を終了する。

【 0 0 8 1 】

（ C ）効果

このように、本発明の一実施形態としての情報処理装置 1 によれば、情報生成部 1 0 8 が、各ユーザの優先因子に応じて、ユーザ毎に応じて選択もしくは並べた説明情報（提示

50

エッジ)を生成する。これにより、ユーザのそれぞれにおいて、機械学習モデルによる予測に対する理解・納得性を高めることができ、X A Iシステム全体の信頼性を向上させることができる。

また、X A Iシステム全体の信頼性が向上することで、ユーザのX A I出力に対する需要幅を拡張することができる。

【0082】

ユーザの優先因子(ユーザ指向データ)がカテゴリカルデータである場合には、情報生成部108が、カテゴリカルデータ発生エッジを提示候補エッジとして抽出し、抽出したこれらの提示候補エッジの中から寄与度が最も高いエッジを提示エッジとして選択する。これにより、ユーザに提示する説明情報には、当該ユーザが機械学習モデルによる予測に寄与すると考える因子が反映されたエッジが含まれることになり、ユーザの知識知見に合った説明情報を生成することができる。従って、ユーザに説明可能な状態とすることができる。

10

【0083】

また、ユーザの優先因子(ユーザ指向データ)が数値データである場合には、情報生成部108が、数値データ発生エッジを提示候補エッジとして抽出し、抽出したこれらの提示候補エッジの中からユーザの優先因子の値と寄与度との2つのパラメータを踏まえた提示エッジを選択する。これにより、ユーザに提示する説明情報には、当該ユーザが機械学習モデルによる予測に寄与すると考える因子が反映されたエッジが含まれることになり、ユーザの知識知見に合った説明情報を生成することができる。従って、ユーザに説明可能な状態とすることができる。

20

【0084】

情報生成部108は、複数の提示候補エッジの中から、優先因子の残業の各値と、各寄与度の値をそれぞれ取得し、各エッジの残業の値と寄与度の値とをそれぞれ正規化し、エッジ毎にこれらの正規化した残業の値と寄与度の値とを総和を算出する。これにより、ユーザの優先因子の値と寄与度との2つのパラメータを踏まえた提示エッジを容易に選択することができる。

【0085】

(D)その他

本発明は上述した実施形態に限定されるものではなく、本発明の趣旨を逸脱しない範囲で種々変形して実施することができる。

30

【0086】

例えば、上述した実施形態においては、健康経営の領域において用いられ、出勤データから将来の休職の有無(休職するか否か)の二値分類を機械学習モデルを用いて行なうタスクを例示したが、これに限定されるものではなく、種々変形して実施することができる。

また、上述した実施形態においては、優先因子数 $N = 1$ の例について示したが、これに限定されるものではなく、優先因子数が複数であってもよい。

【0087】

上述した実施形態においては、入力データが表として表される出勤データの例を示したが、これに限定されるものではない。入力データは、例えば、グラフ等の表以外のフォーマットを有するものであってもよい。また、入力データは、テンソルデータとして取り扱うことができる表やグラフ以外の種類のデータであってもよく、種々変形して実施することができる。

40

また、上述した開示により本実施形態を当業者によって実施・製造することが可能である。

【0088】

(E)付記

(付記1)

それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予

50

測結果と前記複数のデータとに基づいて生成された説明用のモデルを用いて、前記予測結果に対する前記複数のデータのそれぞれの寄与度を算出し、

前記複数の変数のうち特定の変数を選択し、前記複数のデータのそれぞれの前記特定の変数の値と、前記寄与度とに基づいて、前記複数のデータのうち特定のデータを決定し、前記特定のデータを前記予測結果の説明情報として出力する、

処理をコンピュータに実行させることを特徴とする説明情報出力プログラム。

【0089】

(付記2)

前記特定の変数を選択する処理は、ユーザに応じた優先変数に基づいて、前記複数の変数のうち前記特定の変数を選択する処理を含む、

ことを特徴とする付記1記載の説明情報出力プログラム。

【0090】

(付記3)

前記特定のデータを決定する処理は、前記特定の変数が数値データである場合、前記複数のデータのうち、前記寄与度が寄与度閾値よりも大きい複数の提示候補データの中から、前記特定の変数の値と前記寄与度とを用いて前記特定のデータを決定する処理を含む、

ことを特徴とする付記1に記載の説明情報出力プログラム。

【0091】

(付記4)

前記特定のデータを決定する処理は、前記複数の提示候補データから抽出した前記特定の変数の値を正規化した値と、前記複数の提示候補データそれぞれの前記寄与度を正規化した値との和が最も大きいデータを、前記特定のデータに決定する処理を含む、

ことを特徴とする付記3に記載の説明情報出力プログラム。

【0092】

(付記5)

前記特定のデータを決定する処理は、前記特定の変数がカテゴリカルデータである場合、前記複数のデータのうち、前記特定の変数の値が特定のカテゴリを示す複数の提示候補データであって前記寄与度が寄与度閾値よりも大きい前記複数の提示候補データの中から、前記寄与度が最も高いデータを、前記特定のデータに決定する

処理を前記コンピュータに実行させることを特徴とする付記1に記載の説明情報出力プログラム。

【0093】

(付記6)

それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予測結果と前記複数のデータとに基づいて生成された説明用のモデルを用いて、前記予測結果に対する前記複数のデータのそれぞれの寄与度を算出し、

前記複数の変数のうち特定の変数を選択し、前記複数のデータのそれぞれの前記特定の変数の値と、前記寄与度とに基づいて、前記複数のデータのうち特定のデータを決定し、

前記特定のデータを前記予測結果の説明情報として出力する、

処理をコンピュータが実行することを特徴とする説明情報出力方法。

【0094】

(付記7)

前記特定の変数を選択する処理は、ユーザに応じた優先変数に基づいて、前記複数の変数のうち前記特定の変数を選択する処理を含む

ことを特徴とする、付記6記載の説明情報出力方法。

【0095】

(付記8)

前記特定のデータを決定する処理は、前記特定の変数が数値データである場合、前記複数のデータのうち、前記寄与度が寄与度閾値よりも大きい複数の提示候補データの中から、前記特定の変数の値と前記寄与度とを用いて前記特定のデータを決定する処理を含む、

10

20

30

40

50

ことを特徴とする、付記 6 に記載の説明情報出力方法。

【 0 0 9 6 】

(付記 9)

前記特定のデータを決定する処理は、前記複数の提示候補データから抽出した前記特定の変数の値を正規化した値と、前記複数の提示候補データそれぞれの前記寄与度を正規化した値との和が最も大きいデータを、前記特定のデータに決定する処理を含む、

ことを特徴とする、付記 8 に記載の説明情報出力方法。

【 0 0 9 7 】

(付記 10)

前記特定のデータを決定する処理は、前記特定の変数がカテゴリカルデータである場合、前記複数のデータのうち、前記特定の変数の値が特定のカテゴリを示す複数の提示候補データであって前記寄与度が寄与度閾値よりも大きい前記複数の提示候補データの中から、前記寄与度が最も高いデータを、前記特定のデータに決定する

処理を前記コンピュータが実行することを特徴とする、付記 6 に記載の説明情報出力方法。

【 0 0 9 8 】

(付記 11)

それぞれが複数の変数を含む複数のデータの入力に応じて機械学習モデルが出力した予測結果と前記複数のデータとに基づいて生成された説明用のモデルを用いて、前記予測結果に対する前記複数のデータのそれぞれの寄与度を算出し、

前記複数の変数のうち特定の変数を選択し、前記複数のデータのそれぞれの前記特定の変数の値と、前記寄与度とに基づいて、前記複数のデータのうち特定のデータを決定し、前記特定のデータを前記予測結果の説明情報として出力する、

処理部を有することを特徴とする説明情報出力装置。

【 0 0 9 9 】

(付記 12)

前記処理部が、

前記特定の変数を選択処理として、ユーザに応じた優先変数に基づいて、前記複数の変数のうち前記特定の変数を選択する

ことを特徴とする付記 11 に記載の説明情報出力装置。

【 0 1 0 0 】

(付記 13)

前記処理部が、

前記特定のデータを決定する処理として、前記特定の変数が数値データである場合、前記複数のデータのうち、前記寄与度が寄与度閾値よりも大きい複数の提示候補データの中から、前記特定の変数の値と前記寄与度とを用いて前記特定のデータを決定する

ことを特徴とする、付記 11 に記載の説明情報出力装置。

【 0 1 0 1 】

(付記 14)

前記処理部が、

前記特定のデータを決定する処理として、前記複数の提示候補データから抽出した前記特定の変数の値を正規化した値と、前記複数の提示候補データそれぞれの前記寄与度を正規化した値との和が最も大きいデータを、前記特定のデータに決定する

ことを特徴とする、付記 13 に記載の説明情報出力装置。

【 0 1 0 2 】

(付記 15)

前記処理部が、

前記特定のデータを決定する処理として、前記特定の変数がカテゴリカルデータである場合、前記複数のデータのうち、前記特定の変数の値が特定のカテゴリを示す複数の提示候補データであって前記寄与度が寄与度閾値よりも大きい前記複数の提示候補データの中

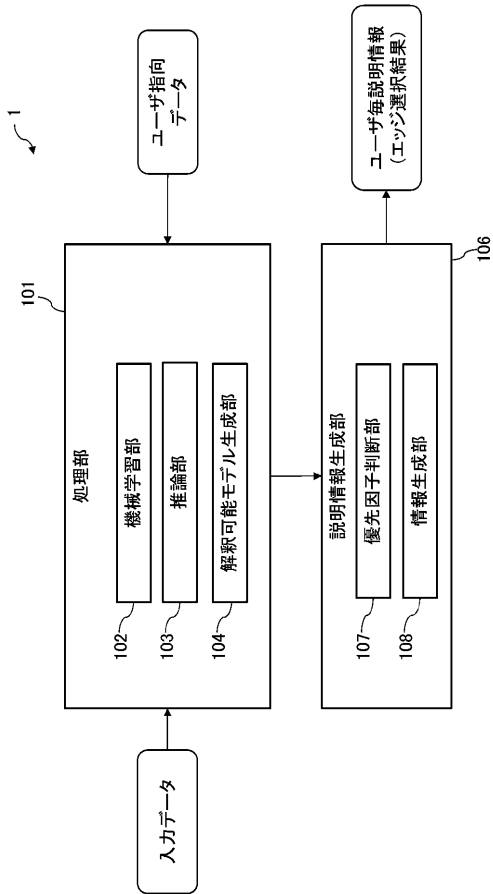
から、前記寄与度が最も高いデータを、前記特定のデータに決定することを特徴とする、付記 1 1 に記載の説明情報出力装置。

【符号の説明】

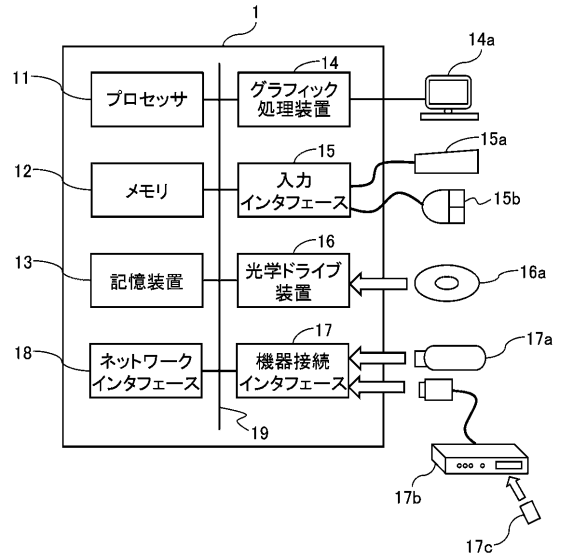
【 0 1 0 3 】

- 1 情報処理装置
- 1 1 プロセッサ
- 1 2 メモリ
- 1 3 記憶装置
- 1 4 グラフィック処理装置
- 1 4 a モニタ 10
- 1 5 入力インタフェース
- 1 5 a キーボード
- 1 5 b マウス
- 1 6 光学ドライブ装置
- 1 6 a 光ディスク
- 1 7 機器接続インタフェース
- 1 7 a メモリ装置
- 1 7 b メモリリーダーライタ
- 1 7 c メモリカード
- 1 8 ネットワークインタフェース 20
- 1 8 a ネットワーク
- 1 9 バス
- 1 0 1 処理部
- 1 0 2 機械学習部
- 1 0 3 推論部
- 1 0 4 解釈可能モデル生成部
- 1 0 6 説明情報生成部
- 1 0 7 優先因子判断部
- 1 0 8 情報生成部

【図1】



【図2】



【図3】

日付	出欠	出張	残業
5/27	出勤	なし	2時間
5/28	出勤	終日	1時間

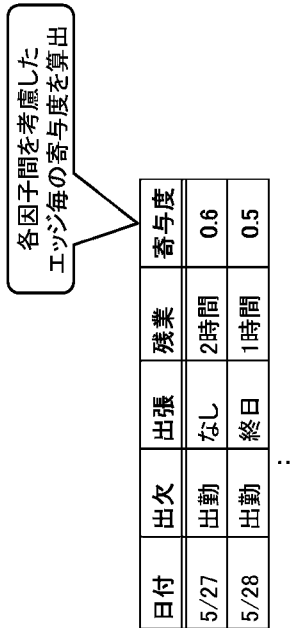
⋮

【図4】

以下に示す出勤管理項目の中から、あなたが休職の有無に最も影響を与えると考えるものを1つ選択してください。

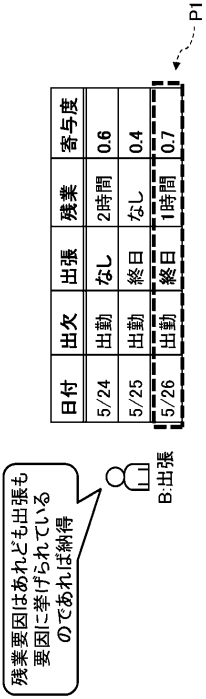
◎ 出張の有無
● 残業時間

【図5】



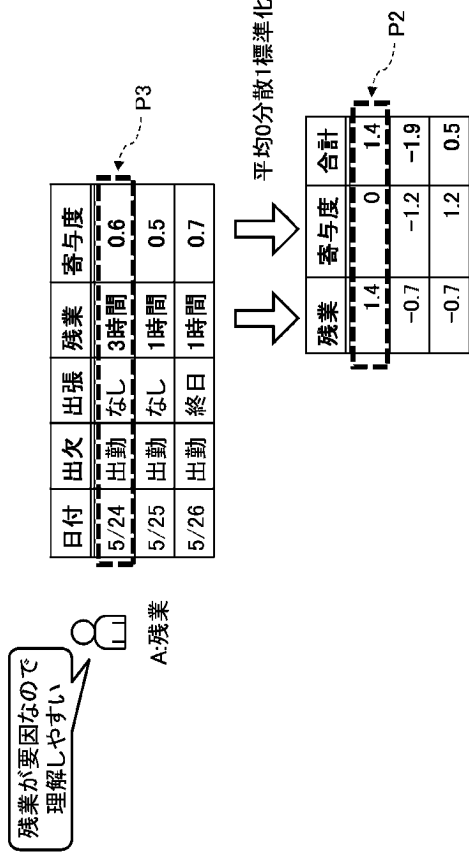
【 図 6 】

カテゴリカルデータ時(出張)の提示エッジの選択方法

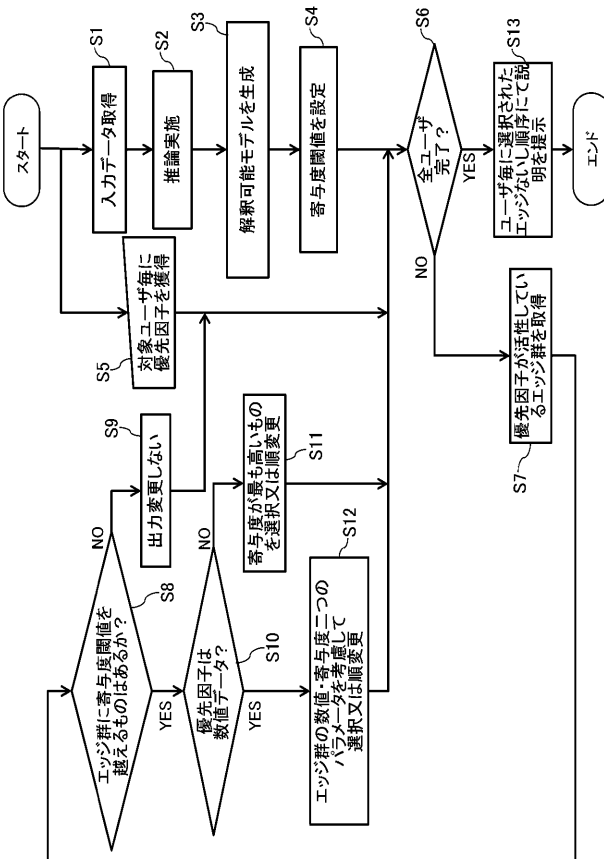


【 図 7 】

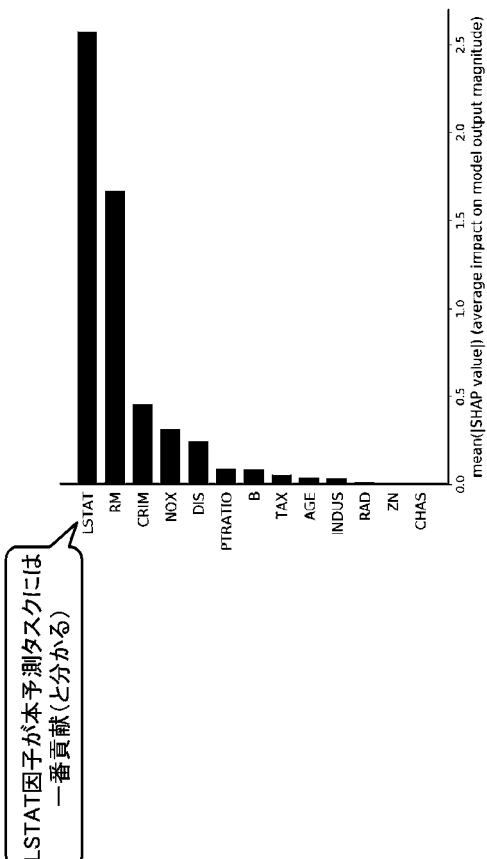
数値データ時(出張)の提示エッジの選択方法



【 図 8 】



【 図 9 】



【図 10】

