(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2023年5月4日(04.05.2023)

WIPO | PCT

(10) 国際公開番号
WO 2023/073837 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号： PCT/JP2021/039692

(22) 国際出願日： 2021年10月27日(27.10.2021)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 富士通株式会社 (**FUJITSU LIMITED**) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

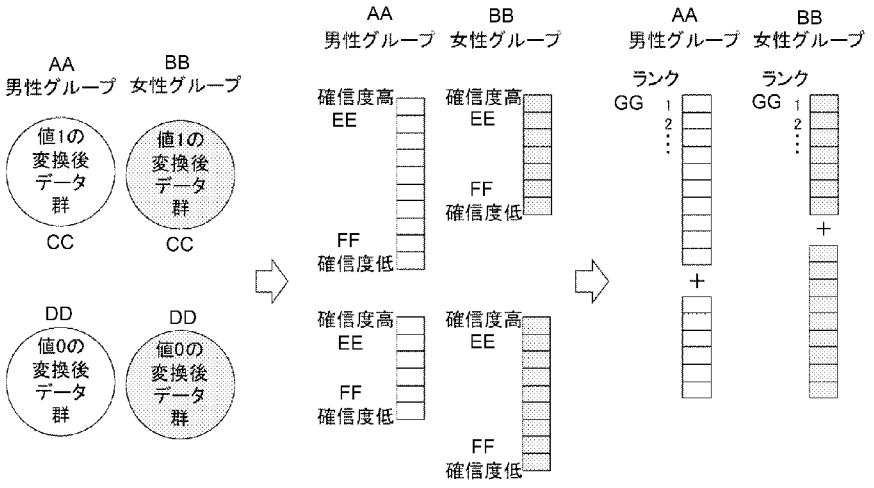
(72) 発明者: 新宮 理史 (**SHINGU, Masafumi**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 中島 淳，外 (**NAKAJIMA, Jun** et al.); 〒1600022 東京都新宿区新宿４丁目３番１７号 ＨＫ新宿ビル７階 太陽国際特許事務所 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) **Title:** DATA MODIFICATION PROGRAM, DEVICE, AND METHOD

(54) 発明の名称： データ修正プログラム、装置、及び方法



AA Male group
BB Female group
CC Value 1 converted data group
DD Value 0 converted data group
EE High degree of confidence
FF Low degree of confidence
GG Rank

(57) **Abstract:** In this data modification device: a machine learning model is generated on the basis of training data, in which components of a one-hot expression of a to-be-modified category variable contained in a converted dataset obtained by converting a to-be-modified category variable contained in a dataset to a one-hot expression is set as a target variable, and in which an attribute other than the to-be-modified category variable is set as an explanatory variable; and a bias of the value of each component in the to-be-modified category variable, due to a difference in the value of a protected attribute,

WO 2023/073837 A1

［続葉有］

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

一　国際調査報告（条約第21条(3)）

is modified in the converted dataset on the basis of the result of the converted data being ranked on the basis of a degree of confidence, which is an estimation result of the machine learning model when a portion other than the one-hot expression of the to-be-modified category variable is input, the ranking being by the order value of the components of the one-hot expression of the to-be-modified category variable and by the degree of confidence.

(57) 要約：データ修正装置は、データ集合に含まれる修正対象のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変更した変換後データ集合に含まれる修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の各成分を目的変数とし、修正対象のカテゴリ変数以外の属性を説明変数とする訓練データに基づいて機械学習モデルを生成し、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を入力した場合の機械学習モデルの推測結果である確信度に基づいて、変換後データを、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の成分の値順、かつ確信度順に並べてランク付けした結果に基づいて、変換後データ集合において、保護属性の値の相違による、修正対象のカテゴリ変数の各成分の値の偏りを修正する。

明　細　書

発明の名称 ： データ修正プログラム、装置、及び方法

技術分野

[0001] 開示の技術は、データ修正プログラム、データ修正装置、及びデータ修正方法に関する。

背景技術

[0002] 機械学習モデルの訓練に用いられた訓練データに含まれる特定の属性の値がバイアスとなり、その機械学習モデルによる判定結果が差別的なものとなる場合がある。例えば、人物の性別、年齢、出身地等の属性の値を説明変数とし、採用やテスト等についてのその人物の合否の結果を目的変数とする訓練データを用いて、人物の属性から合否の結果を予測する機械学習モデルを訓練する場合を想定する。この場合において、性別が女性であることが、合否の結果に対して不利な扱いを受けているという過去の履歴を訓練データとした場合、その訓練データを用いて訓練された機械学習モデルは、女性に不利な判定を下すような、差別的な予測を行うようになる。

[0003] 上記のように、機械学習モデルが差別的な予測を行うことを防止するために、訓練データにおいて、差別的な予測へのバイアスとなる要因を除去するＤＩＲ（Disparate Impact Remover）という技術が存在する。ＤＩＲでは、差別的な扱いからの保護の対象となる属性（上記の例では「性別」）と他の属性の値との相関関係を軽減し、データの偏りを抑制するように、他の属性の値を修正する。

先行技術文献

非特許文献

[0004] 非特許文献1：Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian, "Certifying and Removing Disparate Impact", Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2015, Pages 2

59-268.

発明の概要

発明が解決しようとする課題

[0005] しかしながら、ＤＩＲは、値に大小等の優劣が存在する数値を変数とする属性、いわゆる数値特徴量を対象に修正が行われるものである。したがって、ＤＩＲは、限られた数の数値又はカテゴリを変数に持つ属性、いわゆるカテゴリ変数の修正には適用することができないという問題がある。

[0006] 一つの側面として、開示の技術は、機械学習モデルの差別的な予測へのバイアスとなる要因を、カテゴリ変数を含むデータから除去することを目的とする。

課題を解決するための手段

[0007] 一つの態様として、開示の技術は、第１の複数のデータのそれぞれに含まれる第１種別のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変更した第２の複数のデータを生成する。また、開示の技術は、前記第２の複数のデータそれぞれに含まれる前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の第１の成分を目的変数とし、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を説明変数とする。開示の技術は、上記の目的変数及び説明変数を含む訓練データに基づいて生成された機械学習モデルに、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を入力する。開示の技術は、その場合の前記機械学習モデルの推測結果に基づいて前記第２の複数のデータのそれぞれをランク付けする。そして、開示の技術は、前記ランク付け処理の結果に基づいて、前記第２の複数のデータにおける前記第１種別のカテゴリ変数の各属性の偏りを修正することによって第３の複数のデータを生成する。

発明の効果

[0008] 一つの側面として、機械学習モデルの差別的な予測へのバイアスとなる要因を、カテゴリ変数を含むデータから除去することができる、という効果を

有する。

図面の簡単な説明

[0009] [図1]データ修正装置の機能ブロック図である。

[図2]データ集合の一例を示す図である。

[図3]変換後データ集合の一例を示す図である。

[図4]カテゴリ変数を含むデータ集合にＤＩＲを適用する場合の問題点を説明するための図である。

[図5]確信度の算出を説明するための図である。

[図6]ランクの設定を説明するための図である。

[図7]ランク毎の、修正対象の成分の値及び確信度の組み合わせと、そのランクの中央値の一例を示す図である。

[図8]Ｏｎｅ－ｈｏｔ表現の不整合の修正を説明するための図である。

[図9]Ｏｎｅ－ｈｏｔ表現の不整合の修正を説明するための図である。

[図10]データ修正装置として機能するコンピュータの概略構成を示すブロック図である。

[図11]データ修正処理の一例を示すフローチャートである。

発明を実施するための形態

[0010] 以下、図面を参照して、開示の技術に係る実施形態の一例を説明する。

[0011] 図１に示すように、データ修正装置１０には、複数のデータを含むデータ集合が入力される。そして、データ修正装置１０は、入力されたデータ集合に含まれる、機械学習モデルの差別的な予測へのバイアスとなる要因（以下、単に「バイアス」ともいう）を修正し、修正後データ集合を出力する。データ集合は、開示の技術の「第１の複数のデータ」の一例であり、修正後データ集合は、開示の技術の「第３の複数のデータ」の一例である。図２に、データ集合の一例を示す。図２の例では、各行（各レコード）が１つのデータであり、各データには、そのデータの識別情報である「ＩＤ」が付与されている。また、図２の例では、各データは、「性別」、「年齢」、「出身地」、「労働形態」等の属性のそれぞれについての値（変数）を特徴量として

有する。属性「年齢」の値は数値特徴量である。また、属性「性別」、「出身地」、及び「労働形態」の各々は、その値がカテゴリ変数である。

[0012] ここで、図２に示すようにカテゴリ変数を含むデータ集合から、ＤＩＲを適用してバイアスを除去する方法を考える。上述したように、ＤＩＲは、数値特徴量に対する修正を行うもので、カテゴリ変数には適用することができない。そのため、カテゴリ変数を含むデータ集合に対してＤＩＲを適用する単純な方法として、以下の２つが考えられる。１つ目は、カテゴリ変数には手を加えず、数値特徴量のみを修正する方法である。２つ目は、データ集合からカテゴリ変数列を除外し、数値特徴量列のみを残してＤＩＲを適用する方法である。しかし、１つ目の方法の場合、カテゴリ変数の中に潜在するバイアスが残るため、データ集合からバイアスを完全に除去することができない。また、２つ目の方法では、機械学習モデルでの予測に有用な情報がデータ集合から欠落することになり、そのデータ集合を用いて訓練される機械学習モデルの精度が大きく低下する可能性がある。

[0013] また、カテゴリ変数を含むデータ集合に対してＤＩＲを適用するために、カテゴリ変数を単純に数値へ置き換えて数値特徴量に変換することも考えられる。しかし、この場合、カテゴリ間に優劣がないにもかかわらず、変換後の数値には大小等の優劣が生じてしまうため、訓練される機械学習モデルの精度に影響を与えてしまう。

[0014] そこで、カテゴリ変数をＯｎｅ－ｈｏｔ表現に変換し、ＤＩＲを適用することが考えられる。Ｏｎｅ－ｈｏｔ表現とは、複数の成分のうち、１つの成分の値が１で、他の成分の値は０とする表現である。Ｏｎｅ－ｈｏｔ表現により、１列のカテゴリ変数は複数列の数値特徴量に変換される。図２の例において、カテゴリ変数である「労働形態」をＯｎｅ－ｈｏｔ表現に変換した例を図３に示す。「フルタイム」及び「パートタイム」というカテゴリを含むカテゴリ変数「労働形態」が、図３の太枠部に示すように、「フルタイム」及び「パートタイム」という複数の成分（列）を持つＯｎｅ－ｈｏｔ表現に変換される。この例では、各データにおいて、Ｏｎｅ－ｈｏｔ表現の成分

「フルタイム」及び「パートタイム」のいずれか一方の値が１で、他方の値が０となる。Ｏｎｅ－ｈｏｔ表現の成分（列）それぞれに着目すると、各成分は数値特徴量となっている。

[0015] また、ＤＩＲのアルゴリズムは、性別、国籍等、差別的扱いからの保護の対象となる属性（以下、「保護属性」という）の値によるグループ毎に、保護属性以外の属性の値である数値特徴量の値を基にランキングを生成する。そして、ＤＩＲのアルゴリズムは、保護属性以外の数値特徴量の値を、同一のランクに属する数値特徴量の値の中央値に修正する。Ｏｎｅ－ｈｏｔ表現の１つの成分である数値特徴量に着目した場合、取りうる値が０又は１のみであるため、上記のようなＤＩＲのアルゴリズムを適用すると、値が０のランクと、値が１のランクとの２ランクしか存在しないことになる。したがって、値が０のランクに属する数値特徴量の中央値は０、値が１のランクに属する数値特徴量の中央値は１であり、修正前後で値が変わらないため、バイアスが除去されない。

[0016] 図４を参照して、より詳細に説明する。図４に示すように、カテゴリ変数「労働形態」をＯｎｅ－ｈｏｔ表現に変換したデータ集合において、保護属性を「性別」、修正対象の数値特徴量を、カテゴリ変数「労働形態」のＯｎｅ－ｈｏｔ表現の成分の１つである「フルタイム」とする。ＤＩＲを実行する情報処理装置は、図４中のＡに示すように、保護属性「性別」の値が「男性」であるデータのグループ（以下、「男性グループ」ともいう）に含まれるデータの各々の「フルタイム」の値を抽出する。同様に、情報処理装置は、属性「性別」の値が「女性」であるデータのグループ（以下、「女性グループ」ともいう）に含まれるデータの各々の「フルタイム」の値を抽出する。情報処理装置は、各グループから抽出した値の重複のない集合を取得する。この場合、情報処理装置は、いずれのグループについても、（０，１）という集合を取得する。

[0017] また、情報処理装置は、例えば、各グループから取得された集合のサイズ、すなわち集合に含まれる値の数のうち、小さい方をランキングのサイズと

する。ここでは、いずれのグループの集合もサイズは２であるため、ランキングのサイズも２となる。情報処理装置は、グループ毎に、取得した集合に含まれる値の大小に応じて、その値にランクを設定する。例えば、図４中のＢに示すように、情報処理装置は、男性グループについて、値「１」にランク１を設定し、値「０」にランク２を設定する。情報処理装置は、女性グループについても、値「１」にランク１を設定し、値「０」にランク２を設定する。なお、集合の大きさがランキングのサイズより大きい場合は、２つ以上の値に同一のランクが設定される場合がある。

[0018] 情報処理装置は、図４中のＣに示すように、ランク毎に、そのランクが設定された値を「フルタイム」の値として持つデータを各グループから集め、同一のランク内のデータが持つ「フルタイム」の値の中央値を算出する。そして、情報処理装置は、図４中のＤに示すように、各データの「フルタイム」の値を、そのデータが属するランクについて算出した中央値に修正する。ここでは、ランク１に属する「フルタイム」の値は全て「１」であるため、中央値も「１」であり、ランク２に属する「フルタイム」の値は全て「０」であるため、中央値も「０」である。そのため、上記の方法では、修正前後で値が変更されないため、バイアスも除去されない。

[0019] そこで、本実施形態では、グループ毎に、修正対象の成分の値に基づくデータのランキングを作成する際に、修正対象の成分の値として同じ値を持つデータの中でさらに優劣をつけたうえで、ランキングを作成する。以下、本実施形態に係るデータ修正装置１０の機能的構成について詳述する。

[0020] データ修正装置１０は、機能的には、図１に示すように、変換部１２と、ランキング部１４と、修正部１６とを含む。

[0021] 変換部１２は、データ集合に含まれるデータのそれぞれに含まれる修正対象のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変換し、変換後データ集合を生成する。修正対象のカテゴリ変数は、開示の技術の「第１種別のカテゴリ変数」の一例であり、変換後データ集合は、開示の技術の「第２の複数のデータ」の一例である。例えば、変換部１２は、図２に示すようなデータ集合にお

いて、修正対象のカテゴリ変数を「労働形態」とする場合、属性「労働形態」のカテゴリ変数の値をＯｎｅ－ｈｏｔ表現に変換し、図３に示すような変換後データ集合を生成する。なお、図３に示す変換後データ集合の各行（各レコード）は、１つの変換後データに相当する。

[0022] ランキング部１４は、変換後データにおいて、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の各成分の値に優劣を付けるための代理値を算出する。代理値は、例えば、変換後データのうち、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分に対する、Ｏｎｅ－ｈｏｔ表現の各成分の値の尤もらしさを示す確信度としてよい。具体的には、図５に示すように、ランキング部１４は、変換後データ集合に含まれる変換後データのそれぞれに含まれる修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の各成分（図５の例では、斜線で示す列「成分１」及び「成分２」）を目的変数とする。また、ランキング部１４は、変換後データのうち修正対象のカテゴリ変数以外の部分（図５の例では、「属性１」～「属性６」）を説明変数とする。ランキング部１４は、上記の目的変数及び説明変数からなる訓練データに基づいて、機械学習モデルを生成する。ランキング部１４は、訓練データの全てを用いて機械学習モデルの機械学習を実行するようにしてもよい。また、ランキング部１４は、訓練データの一部（例えば、全体の４／５）を用いて機械学習モデルの機械学習を実行し、訓練データの残りの部分（例えば、全体の１／５）を用いて、生成された機械学習モデルを検証する交差検証を行ってもよい。

[0023] ランキング部１４は、生成した機械学習モデルに変換後データのうち修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分、すなわち訓練データの説明変数部分を入力し、機械学習モデルの推測結果として、各変換後データの各成分についての確信度を取得する。各成分の確信度はそれぞれ０．０～１．０の値であり、各変換後データについての各成分の確信度の合計は１．０となる。

[0024] ランキング部１４は、機械学習モデルの推測結果である確信度に基づいて、変換後データのそれぞれをランク付けする。具体的には、ランキング部１

４は、保護属性の値別に、変換後データを、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の成分の値の順に並べる。さらに、ランキング部１４は、修正対象の成分の値が同一である変換後データを確信度の高い順に並べてランクを設定する。なお、保護属性は、開示の技術の「第１の属性」の一例である。

[0025] 図６を参照して、より具体的に説明する。ランキング部１４は、保護属性の指定を受け付ける。例えば、受け付けた保護属性の指定が属性「性別」である場合、ランキング部１４は、男性グループ及び女性グループのそれぞれについて、修正対象の成分の値が「１」の変換後データ、及び、修正対象の成分の値が「０」の変換後データを抽出する。ランキング部１４は、グループ毎に、修正対象の成分の値が「１」の変換後データを確信度が高い順に並べる。また、同様に、ランキング部１４は、各グループについて、修正対象の成分の値が「０」の変換後データを確信度が高い順に並べる。ランキング部１４は、グループ毎に、確信度が高い順に並べた、修正対象の成分の値が「１」の変換後データの後に、確信度が高い順に並べた、修正対象の成分の値が「０」の変換後データを連結する。これにより、グループ毎に、修正対象の成分の値順、かつ確信度が高い順に変換後データが並べられる。

[0026] ランキング部１４は、各グループについて、修正対象の成分の値と確信度との組み合わせについて重複のない集合を取得し、集合のサイズ、すなわち集合に含まれる組み合わせの数のうち、小さい方をランキングのサイズとする。そして、ランキング部１４は、ランキングのサイズに応じて、修正対象の成分の値順、かつ確信度が高い順に並べた変換後データにランクを設定する。例えば、あるカテゴリ変数のＯｎｅ－ｈｏｔ表現の１成分が同じ値「１」である２つの変換後データについて、その成分についての確信度が、一方の変換後データは０．９９８、他方の変換後データは０．９４０であったとする。この場合、前者の変換後データの方が高いランクに設定される。

[0027] 修正部１６は、ランキング部１４によるランク付け処理の結果に基づいて、変換後データ集合における、修正対象のカテゴリ変数の各属性の偏りを修

正することによって修正後データ集合を生成する。具体的には、修正部１６は、ランキング部１４により設定されたランクが同一の変換後データの修正対象の成分の値を、ランクが同一の変換後データの集合における修正対象の成分の値の中央値に修正する。図７に、ランク毎の、修正対象の成分の値及び確信度の組み合わせと、そのランクの中央値の一例を示す。修正対象の成分の値が同一であっても、確信度の高低でランクが異なるため、同一のランクであっても、男性グループと女性グループとで値が異なる場合がある（図７の例では、ランク４５３）。この場合、いずれか一方のグループの変換後データの修正対象の成分の値が、元の値から変更されることになる。すなわち、カテゴリ変数に対するＤＩＲの適用が可能になる。なお、修正部１６は、中央値が０．５になる場合、１又は０の予め定めた方の値を中央値、すなわち修正後の値とすればよい。

[0028] さらに、修正部１６は、成分の値を中央値に変更した後の修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の整合性を保つように修正する。具体的には、図８に示すように、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値、すなわち「１」が複数存在する場合がある。この場合、修正部１６は、各成分の確信度に基づいて、複数の「１」のうち、いずれか１つの成分の値を「１」として残し、他の成分の値を「０」に修正する。例えば、修正部１６は、確信度が最も高い成分の値を「１」として残すようにしてよい。図８の例では、成分１及び成分２の値が「１」となっているが、成分１の方が成分２より確信度が高いため、修正部１６は、成分２の値を「１」から「０」に修正する（図８中の太枠部）。

[0029] また、図９に示すように、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値、すなわち「１」が存在しない場合がある。この場合、修正部１６は、各成分の確信度に基づいて、いずれか１つの成分の値を「１」に修正する。例えば、修正部１６は、確信度が最も高い成分の値を「１」に修正してよい。図９の例では、成分１、成分２、及び成分３のいずれも値が「０」であるため、修正部１６は、確信度が最も高い成分１の値を「

０」から「１」に修正する（図９中の太枠部）。修正部１６は、上記のようなＯｎｅ－ｈｏｔ表現の整合性を保つ修正も行ったうえで、修正後データ集合を出力する。

[0030] データ修正装置１０は、例えば図１０に示すコンピュータ４０で実現されてよい。コンピュータ４０は、ＣＰＵ（Central Processing Unit）４１と、一時記憶領域としてのメモリ４２と、不揮発性の記憶部４３とを備える。また、コンピュータ４０は、入力部、表示部等の入出力装置４４と、非一時的な記憶媒体４９に対するデータの読み込み及び書き込みを制御するＲ／Ｗ（Read/Write）部４５とを備える。また、コンピュータ４０は、インターネット等のネットワークに接続される通信Ｉ／Ｆ（Interface）４６を備える。ＣＰＵ４１、メモリ４２、記憶部４３、入出力装置４４、Ｒ／Ｗ部４５、及び通信Ｉ／Ｆ４６は、バス４７を介して互いに接続される。

[0031] 記憶部４３は、ＨＤＤ（Hard Disk Drive）、ＳＳＤ（Solid State Drive）、フラッシュメモリ等によって実現されてよい。記憶媒体としての記憶部４３には、コンピュータ４０を、データ修正装置１０として機能させるためのデータ修正プログラム５０が記憶される。データ修正プログラム５０は、変換プロセス５２と、ランキングプロセス５４と、修正プロセス５６とを有する。

[0032] ＣＰＵ４１は、データ修正プログラム５０を記憶部４３から読み出してメモリ４２に展開し、データ修正プログラム５０が有するプロセスを順次実行する。ＣＰＵ４１は、変換プロセス５２を実行することで、図１に示す変換部１２として動作する。また、ＣＰＵ４１は、ランキングプロセス５４を実行することで、図１に示すランキング部１４として動作する。また、ＣＰＵ４１は、修正プロセス５６を実行することで、図１に示す修正部１６として動作する。これにより、データ修正プログラム５０を実行したコンピュータ４０が、データ修正装置１０として機能することになる。なお、プログラムを実行するＣＰＵ４１はハードウェアである。

[0033] なお、データ修正プログラム５０により実現される機能は、例えば半導体

集積回路、より詳しくはＡＳＩＣ（Application Specific Integrated Circuit）等で実現することも可能である。

[0034] 次に、本実施形態に係るデータ修正装置１０の作用について説明する。データ修正装置１０にデータ集合が入力され、バイアスの除去が指示されると、データ修正装置１０において、図１１に示すデータ修正処理が実行される。なお、データ修正処理は、開示の技術のデータ修正方法の一例である。

[0035] ステップＳ１０で、変換部１２が、データ修正装置１０に入力されたデータ集合を取得する。また、ランキング部１４が、ユーザから保護属性の指定を受け付ける。次に、ステップＳ１２で、変換部１２が、取得したデータ集合に含まれる修正対象のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変換し、変換後データ集合を生成する。修正対象のカテゴリ変数は、ユーザから指定されたカテゴリ変数であってもよいし、データ集合に含まれる全てのカテゴリ変数を順次修正対象のカテゴリ変数として設定してもよい。

[0036] 次に、ステップＳ１４で、ランキング部１４が、変換後データ集合に含まれる修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の各成分を目的変数、修正対象のカテゴリ変数以外の属性を説明変数とする訓練データを生成する。ランキング部１４は、生成した訓練データに基づいて、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分に対する、Ｏｎｅ－ｈｏｔ表現の各成分の値の尤もらしさを示す確信度を推測する機械学習モデルを生成する。

[0037] 次に、ステップＳ１６で、ランキング部１４が、生成した機械学習モデルに、生成した訓練データの説明変数部分を入力する。そして、ランキング部１４が、機械学習モデルの推測結果として、各変換後データにおける、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の各成分についての確信度を取得する。

[0038] 次に、ステップＳ１８で、ランキング部１４が、上記ステップＳ１０で受け付けた保護属性の値別に、変換後データを、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の成分の値の順に並べる。さらに、ランキング部１４が、修正対象の成分の値が同一である変換後データを確信度の高い順に並べてラ

ンクを設定する。

[0039] 次に、ステップＳ２０で、修正部１６が、変換後データの修正対象の成分の値を、上記ステップＳ１８で設定されたランクが同一の変換後データの集合における修正対象の成分の値の中央値に修正する。次に、ステップＳ２２で、修正部１６が、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現に「１」が複数ある、又は「１」がないという不整合が存在する変換後データについて、各成分の確信度に基づいて、いずれか１つの成分の値が「１」となるように修正する。そして、修正部１６が、修正後データ集合を出力し、データ修正処理は終了する。

[0040] 以上説明したように、本実施形態に係るデータ修正装置は、データ集合に含まれる修正対象のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変更した変換後データ集合を生成する。また、データ修正装置は、変換後データ集合に含まれる修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現の各成分を目的変数とし、修正対象のカテゴリ変数以外の属性を説明変数とする訓練データに基づいて機械学習モデルを生成する。そして、データ修正装置は、変換後データ集合のうち、修正対象のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を入力した場合の機械学習モデルの推測結果に基づいて、変換後データのそれぞれをランク付けする。さらに、データ修正装置は、ランク付け処理の結果に基づいて、変換後データ集合において、保護属性の値の相違による、修正対象のカテゴリ変数の各成分の値の偏りを修正することによって、修正後データ集合を生成し、出力する。これにより、機械学習モデルの差別的な予測へのバイアスとなる要因を、カテゴリ変数を含むデータから除去することができる。また、本実施形態に係るデータ修正装置により生成された修正後データ集合を訓練データとして用いて機械学習モデルを生成することにより、機械学習モデルによる差別的な予測を抑制することができる。

[0041] なお、上記実施形態では、データ修正プログラムが記憶部に予め記憶（インストール）されている態様を説明したが、これに限定されない。開示の技術に係るプログラムは、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ、ＵＳＢメモリ等の

非一時的記憶媒体に記憶された形態で提供することも可能である。

符号の説明

[0042] １０　データ修正装置

１２　変換部

１４　ランキング部

１６　修正部

４０　コンピュータ

４１　ＣＰＵ

４２　メモリ

４３　記憶部

４４　入出力装置

４５　Ｒ／W部

４６　通信Ｉ／Ｆ

４７　バス

４９　記憶媒体

５０　データ修正プログラム

請求の範囲

[請求項1] 第１の複数のデータのそれぞれに含まれる第１種別のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変更した第２の複数のデータを生成し、

前記第２の複数のデータそれぞれに含まれる前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の第１の成分を目的変数とし、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を説明変数とする訓練データに基づいて生成された機械学習モデルに、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を入力した場合の前記機械学習モデルの推測結果に基づいて前記第２の複数のデータのそれぞれをランク付けし、

前記ランク付け処理の結果に基づいて、前記第２の複数のデータにおける前記第１種別のカテゴリ変数の各属性の偏りを修正することによって第３の複数のデータを生成する、

処理をコンピュータに実行させることを特徴とするデータ修正プログラム。

[請求項2] 前記ランク付け処理は、第１の属性の値別に、前記データを前記第１の成分の値の順に並べると共に、前記第１の成分の値が同一のデータを前記推測結果が示す値の順に並べてランクを設定することを含む、

請求項１に記載のデータ修正プログラム。

[請求項3] 前記偏りを修正する処理は、前記第１の属性の値別に設定されたランクが同一のデータの前記第１の成分の値を、前記ランクが同一のデータの集合における前記第１の成分の値の中央値に修正することを含む、

請求項２に記載のデータ修正プログラム。

[請求項4] 前記偏りを修正する処理は、修正後の前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の整合性を保つように修正することを含む、

請求項１～請求項３のいずれか１項に記載のデータ修正プログラム。

[請求項5] 前記Ｏｎｅ－ｈｏｔ表現の整合性を保つように修正する処理は、前記Ｏｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値が複数存在する場合には、前記推測結果に基づいて、複数のＯｎｅ－ｈｏｔの値のうち１つを残すように修正し、前記Ｏｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値が存在しない場合には、前記推測結果に基づいて、前記Ｏｎｅ－ｈｏｔ表現内のいずれかの成分の値をＯｎｅ－ｈｏｔの値に変更することを含む、

請求項４に記載のデータ修正プログラム。

[請求項6] 前記機械学習モデルの推測結果は、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分に対する、前記第１の成分の値の確信度である、

請求項１～請求項５のいずれか１項に記載のデータ修正プログラム。

[請求項7] 前記訓練データの全てを用いて前記機械学習モデルの機械学習を実行するか、又は、前記訓練データの一部を用いて前記機械学習モデルの機械学習を実行し、前記訓練データの残りを用いて、生成された機械学習モデルを検証する、

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～請求項６のいずれか１項に記載のデータ修正プログラム。

[請求項8] 第１の複数のデータのそれぞれに含まれる第１種別のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変更した第２の複数のデータを生成し、

前記第２の複数のデータそれぞれに含まれる前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の第１の成分を目的変数とし、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を説明変数とする訓練データに基づいて生成された機械学習モデルに、前記第２の複数のデータのうち前記第１種別のカテ

ゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を入力した場合の前記機械学習モデルの推測結果に基づいて前記第２の複数のデータのそれぞれをランク付けし、

前記ランク付け処理の結果に基づいて、前記第２の複数のデータにおける前記第１種別のカテゴリ変数の各属性の偏りを修正することによって第３の複数のデータを生成する、

処理を実行する制御部を含むことを特徴とするデータ修正装置。

[請求項9] 前記ランク付け処理は、第１の属性の値別に、前記データを前記第１の成分の値の順に並べると共に、前記第１の成分の値が同一のデータを前記推測結果が示す値の順に並べてランクを設定することを含む、

請求項８に記載のデータ修正装置。

[請求項10] 前記偏りを修正する処理は、前記第１の属性の値別に設定されたランクが同一のデータの前記第１の成分の値を、前記ランクが同一のデータの集合における前記第１の成分の値の中央値に修正することを含む、

請求項９に記載のデータ修正装置。

[請求項11] 前記偏りを修正する処理は、修正後の前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の整合性を保つように修正することを含む、

請求項８～請求項１０のいずれか１項に記載のデータ修正装置。

[請求項12] 前記Ｏｎｅ－ｈｏｔ表現の整合性を保つように修正する処理は、前記Ｏｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値が複数存在する場合には、前記推測結果に基づいて、複数のＯｎｅ－ｈｏｔの値のうち１つを残すように修正し、前記Ｏｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値が存在しない場合には、前記推測結果に基づいて、前記Ｏｎｅ－ｈｏｔ表現内のいずれかの成分の値をＯｎｅ－ｈｏｔの値に変更することを含む、

請求項１１に記載のデータ修正装置。

[請求項13] 前記機械学習モデルの推測結果は、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分に対する、前記第１の成分の値の確信度である、

請求項８～請求項１２のいずれか１項に記載のデータ修正装置。

[請求項14] 前記制御部は、前記訓練データの全てを用いて前記機械学習モデルの機械学習を実行するか、又は、前記訓練データの一部を用いて前記機械学習モデルの機械学習を実行し、前記訓練データの残りを用いて、生成された機械学習モデルを検証する、

ことを特徴とする請求項８～請求項１３のいずれか１項に記載のデータ修正装置。

[請求項15] 第１の複数のデータのそれぞれに含まれる第１種別のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変更した第２の複数のデータを生成し、

前記第２の複数のデータそれぞれに含まれる前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の第１の成分を目的変数とし、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を説明変数とする訓練データに基づいて生成された機械学習モデルに、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を入力した場合の前記機械学習モデルの推測結果に基づいて前記第２の複数のデータのそれぞれをランク付けし、

前記ランク付け処理の結果に基づいて、前記第２の複数のデータにおける前記第１種別のカテゴリ変数の各属性の偏りを修正することによって第３の複数のデータを生成する、

処理をコンピュータが実行することを特徴とするデータ修正方法。

[請求項16] 前記ランク付け処理は、第１の属性の値別に、前記データを前記第１の成分の値の順に並べると共に、前記第１の成分の値が同一のデータを前記推測結果が示す値の順に並べてランクを設定することを含む、

請求項１５に記載のデータ修正方法。

[請求項17] 前記偏りを修正する処理は、前記第１の属性の値別に設定されたランクが同一のデータの前記第１の成分の値を、前記ランクが同一のデータの集合における前記第１の成分の値の中央値に修正することを含む、

請求項１６に記載のデータ修正方法。

[請求項18] 前記偏りを修正する処理は、修正後の前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の整合性を保つように修正することを含む、

請求項１５～請求項１７のいずれか１項に記載のデータ修正方法。

[請求項19] 前記Ｏｎｅ－ｈｏｔ表現の整合性を保つように修正する処理は、前記Ｏｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値が複数存在する場合には、前記推測結果に基づいて、複数のＯｎｅ－ｈｏｔの値のうち１つを残すように修正し、前記Ｏｎｅ－ｈｏｔ表現内にＯｎｅ－ｈｏｔの値が存在しない場合には、前記推測結果に基づいて、前記Ｏｎｅ－ｈｏｔ表現内のいずれかの成分の値をＯｎｅ－ｈｏｔの値に変更することを含む、

請求項１８に記載のデータ修正方法。

[請求項20] 第１の複数のデータのそれぞれに含まれる第１種別のカテゴリ変数をＯｎｅ－ｈｏｔ表現に変更した第２の複数のデータを生成し、

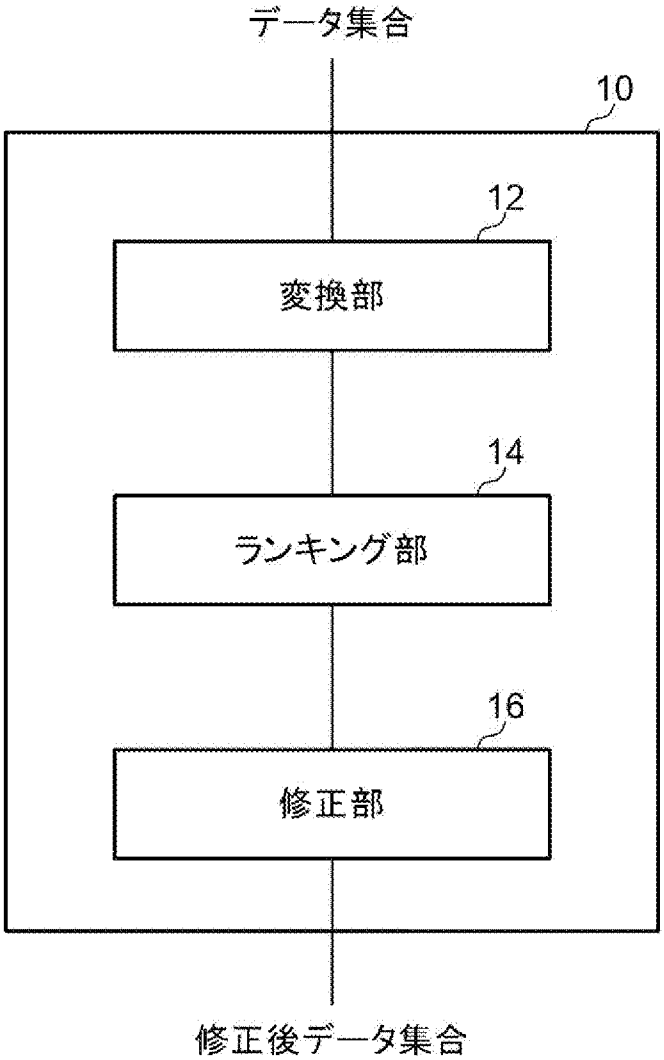
前記第２の複数のデータそれぞれに含まれる前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現の第１の成分を目的変数とし、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を説明変数とする訓練データに基づいて生成された機械学習モデルに、前記第２の複数のデータのうち前記第１種別のカテゴリ変数のＯｎｅ－ｈｏｔ表現以外の部分を入力した場合の前記機械学習モデルの推測結果に基づいて前記第２の複数のデータのそれぞれをランク付けし、

前記ランク付け処理の結果に基づいて、前記第２の複数のデータに

おける前記第１種別のカテゴリ変数の各属性の偏りを修正することによって第３の複数のデータを生成する、

処理をコンピュータに実行させることを特徴とするデータ修正プログラムを記憶した非一時的記憶媒体。

[図1]



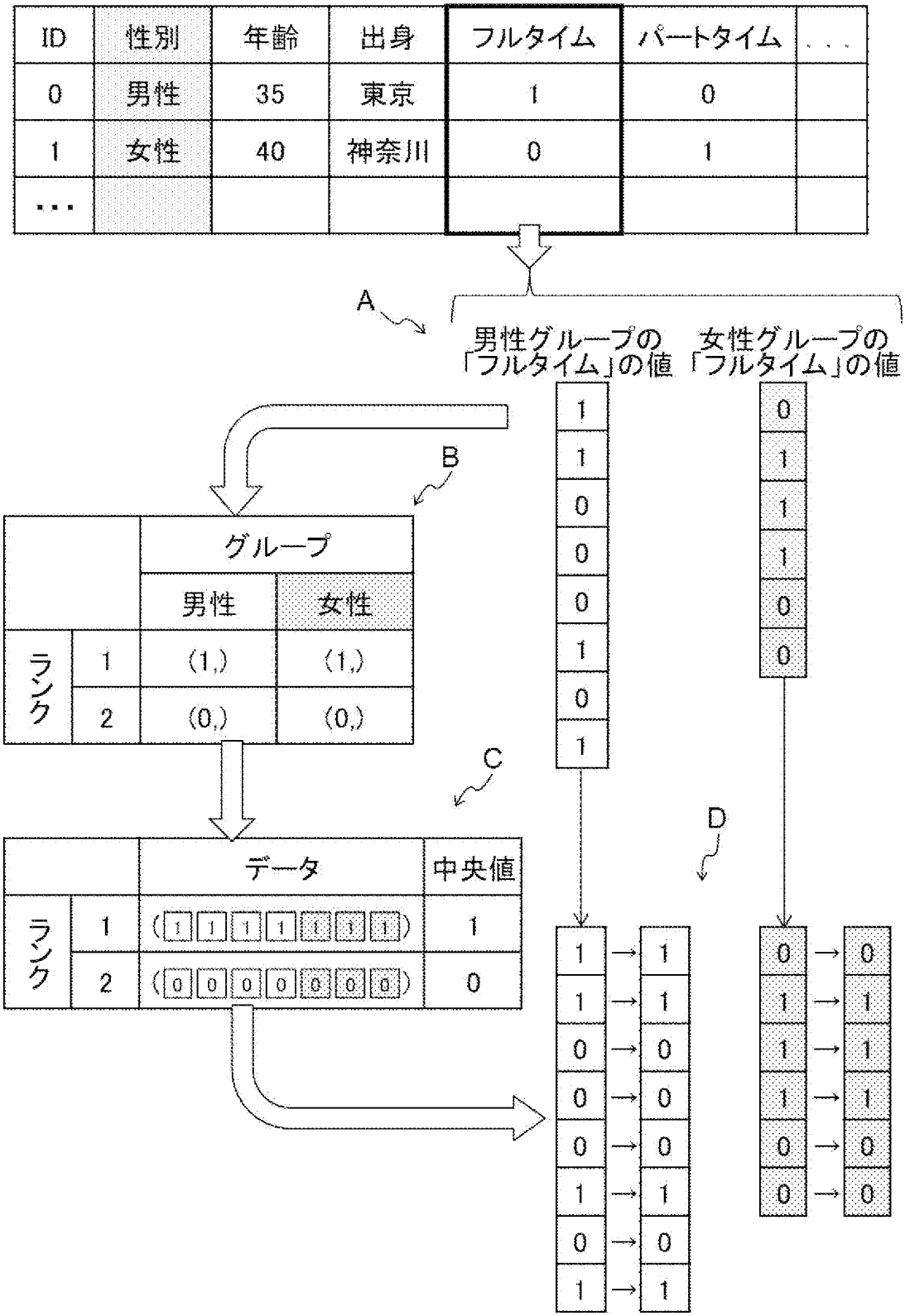
[図2]

ID	性別	年齢	出身地	労働形態	. . .
0	男性	35	東京	フルタイム	
1	女性	40	神奈川	パートタイム	
・・・					

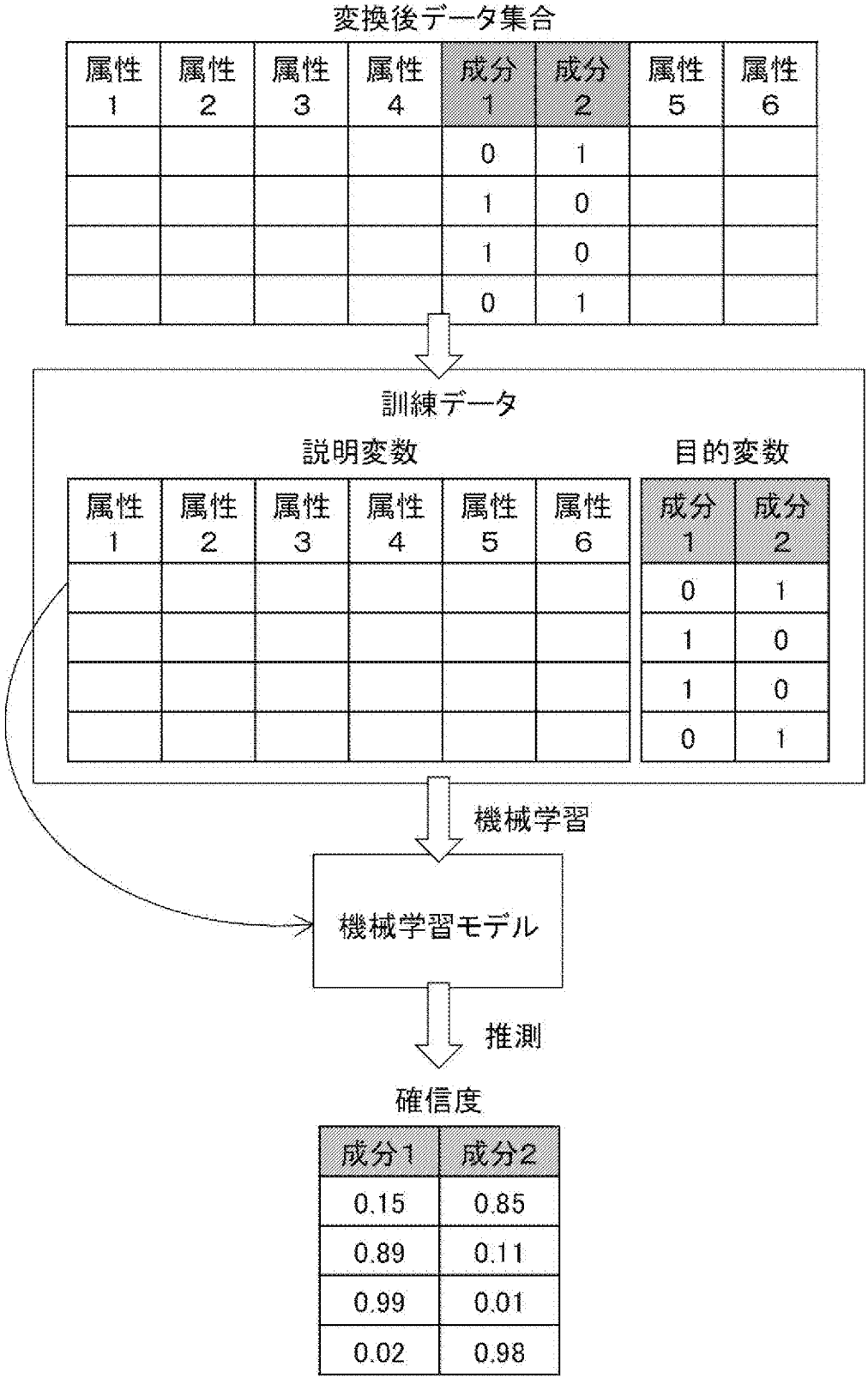
[図3]

ID	性別	年齢	出身	フルタイム	パートタイム	. . .
0	男性	35	東京	1	0	
1	女性	40	神奈川	0	1	
・・・						

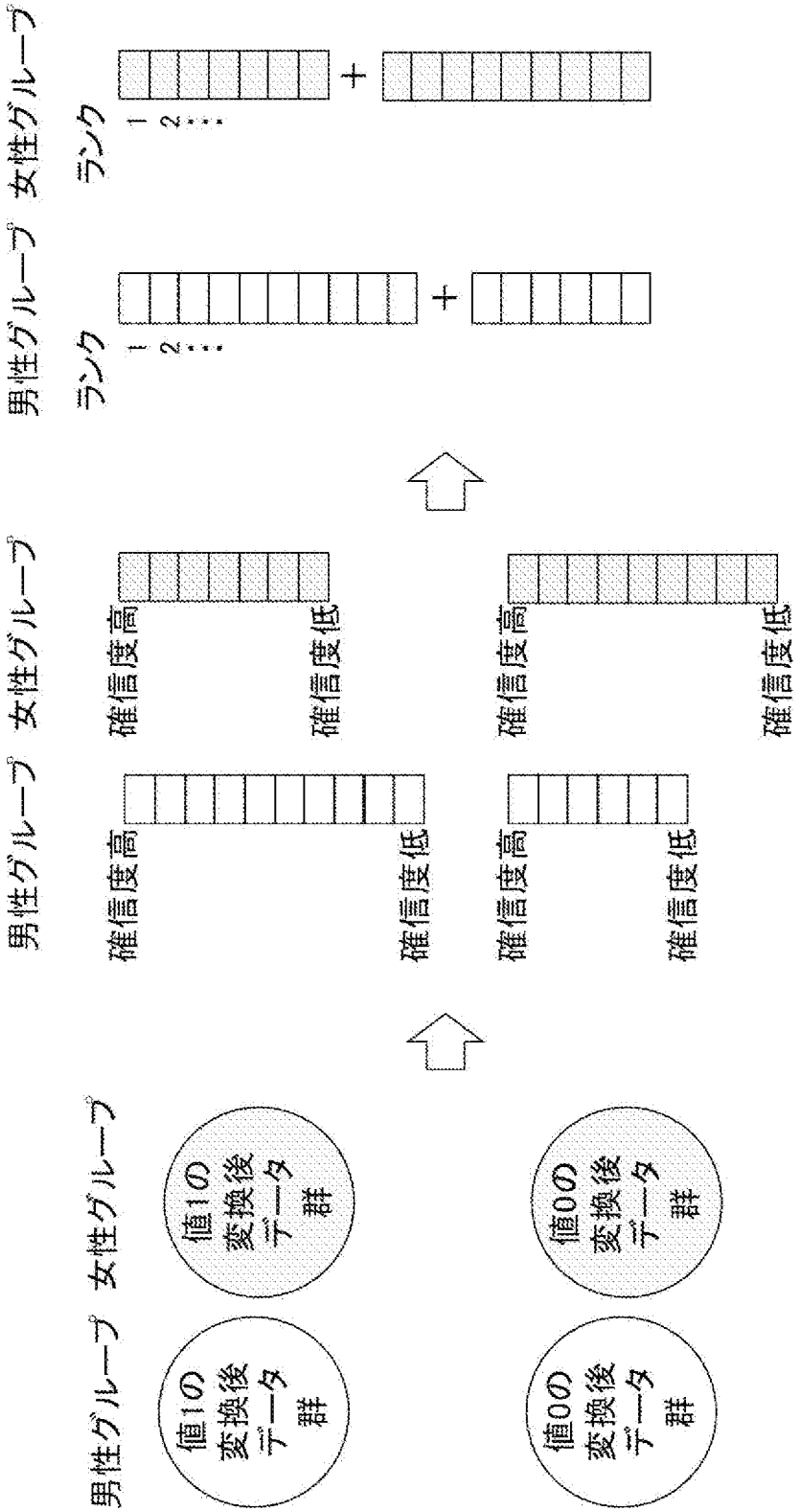
[図4]


ID
性別
年齢
出身
フルタイム
パートタイム
0
男性
35
東京
1
0
1
女性
40
神奈川
0
1
A
男性グループの「フルタイム」の値
女性グループの「フルタイム」の値
B
グループ
男性
女性
ランク
1
(1,)
(1,)
2
(0,)
(0,)
C
データ
中央値
(1 1 1 1 1 1 1)
1
(0 0 0 0 0 0 0)
0
D

[図5]


変換後データ集合
属性1
属性2
属性3
属性4
成分1
成分2
属性5
属性6
0 1
1 0
1 0
0 1
訓練データ
説明変数
目的変数
属性1
属性2
属性3
属性4
属性5
属性6
成分1
成分2
0 1
1 0
1 0
0 1
機械学習
機械学習モデル
推測
確信度
成分1
成分2
0.15 0.85
0.89 0.11
0.99 0.01
0.02 0.98

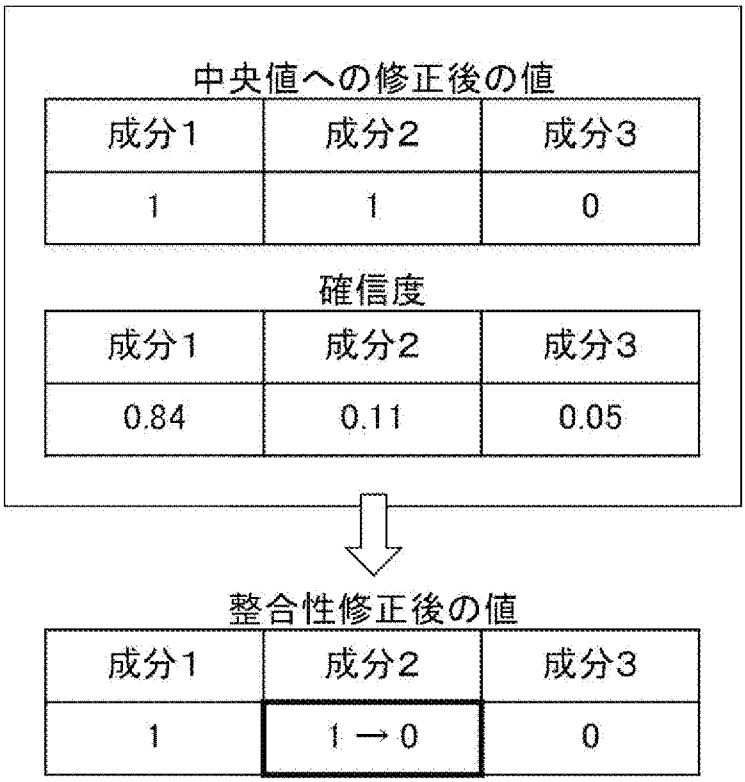
[図6]



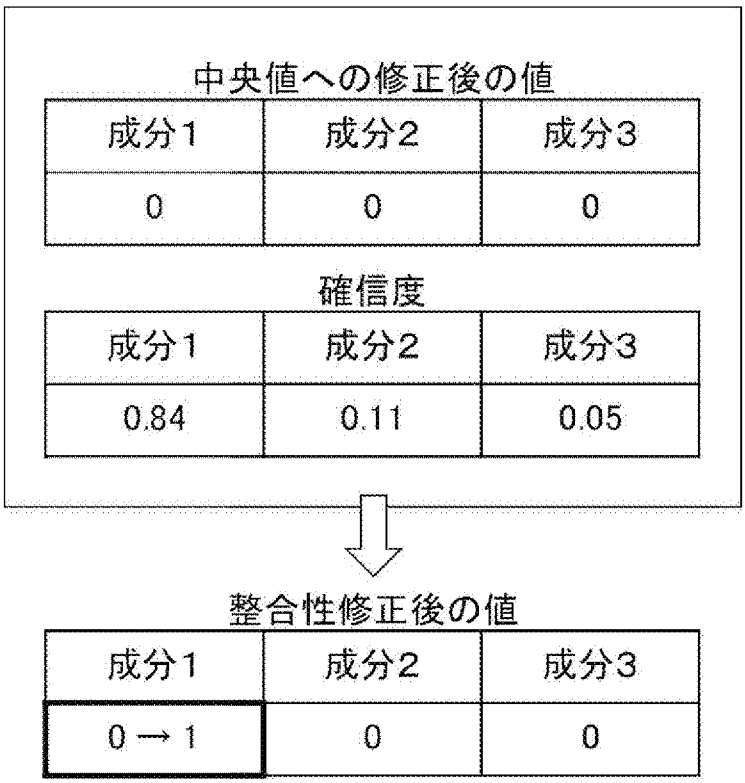
[図7]

		男性グループ	女性グループ	中央値 （修正後の値）
ランク	1	(1, 確信度0.99)	(1, 確信度0.98)	1
	2	(1, 確信度0.98)	(1, 確信度0.97)	1
	・・・			
	453	(1, 確信度0.56)	(0, 確信度0.76)	0
	・・・			

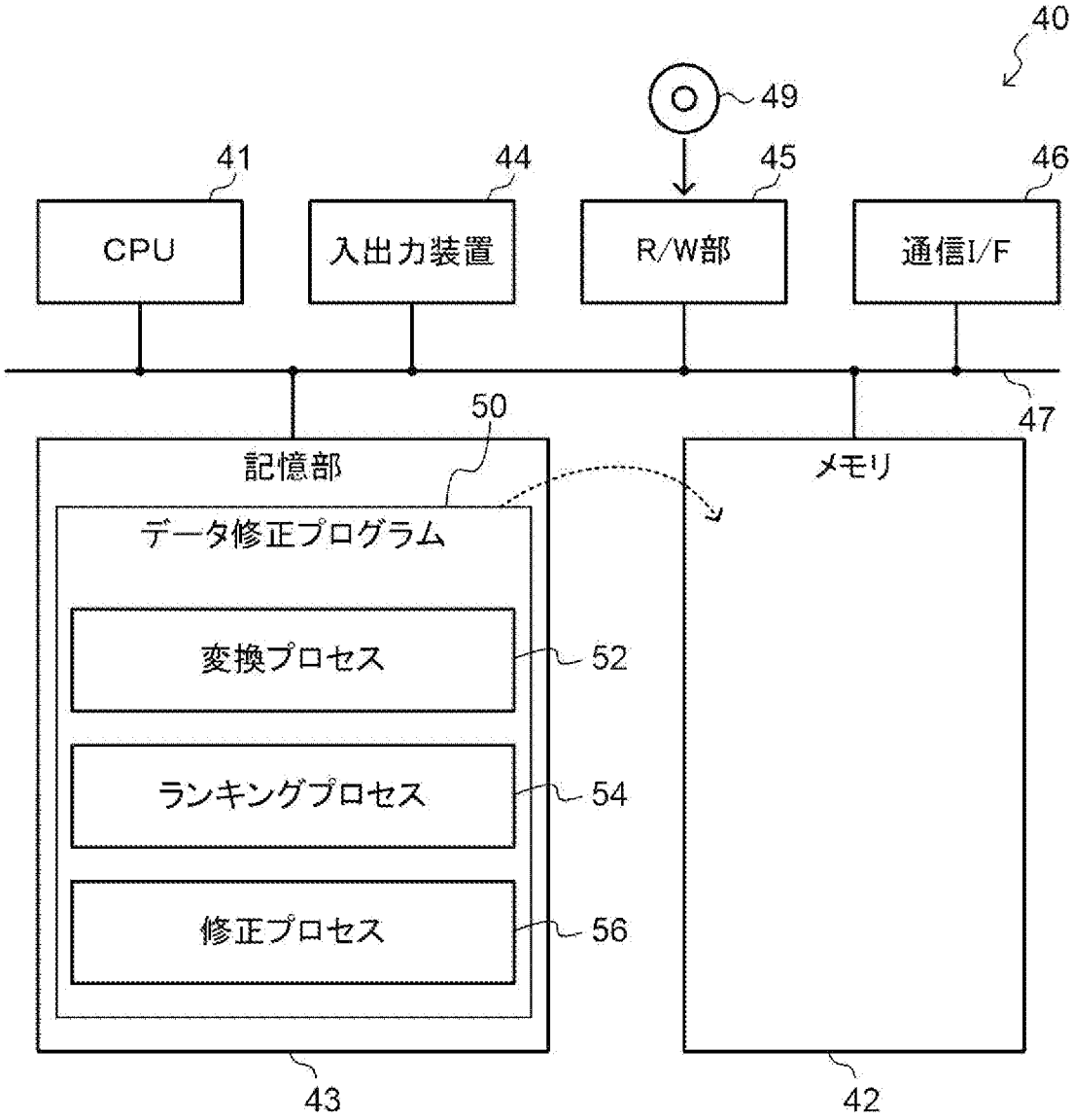
[図8]



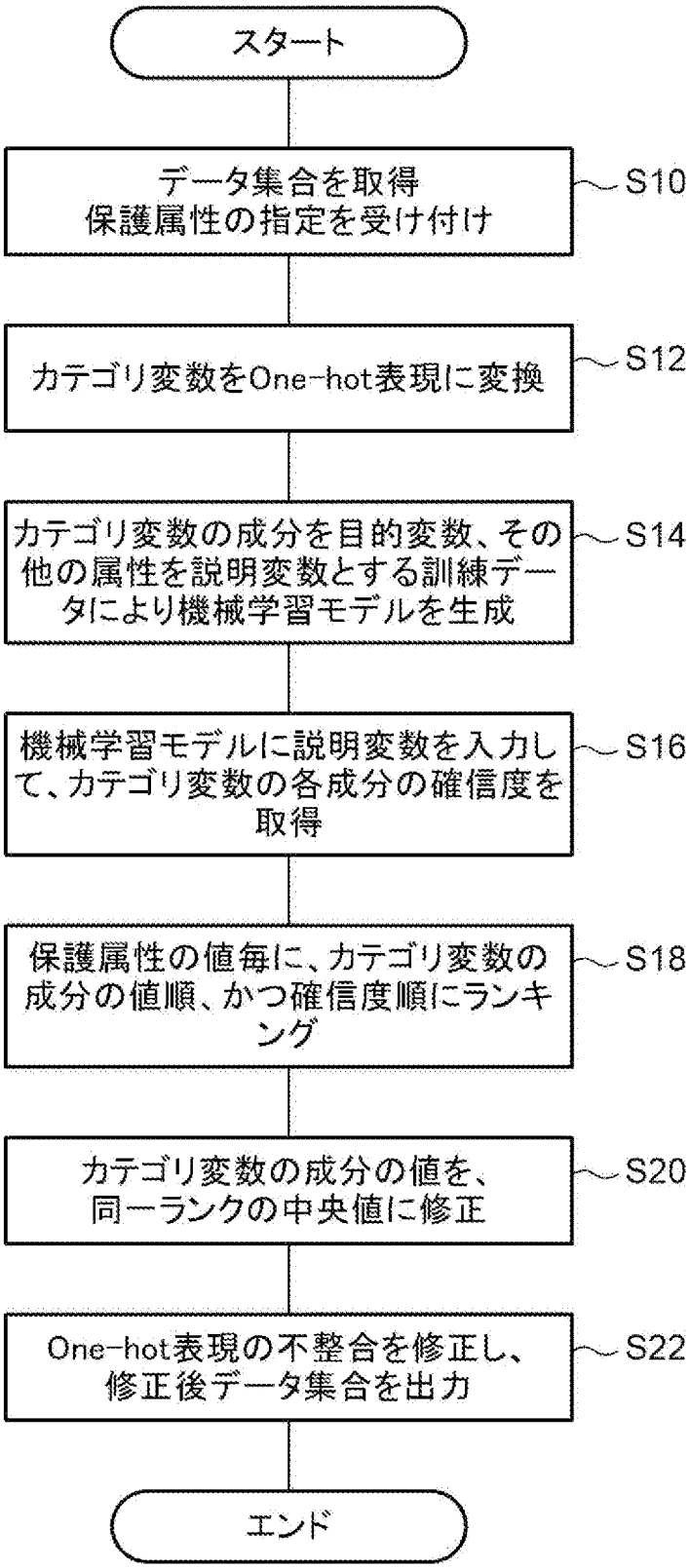
[図9]


中央値への修正後の値
成分1
成分2
成分3
0
0
0
確信度
成分1
成分2
成分3
0.84
0.11
0.05
整合性修正後の値
成分1
成分2
成分3
0 → 1
0
0

[図10]


40
49
41
44
45
46
CPU
入出力装置
R/W部
通信I/F
47
50
記憶部
メモリ
データ修正プログラム
変換プロセス
52
ランキングプロセス
54
修正プロセス
56
43
42

[図11]


スタート
データ集合を取得
保護属性の指定を受け付け
S10
カテゴリ変数をOne-hot表現に変換
S12
カテゴリ変数の成分を目的変数、その他の属性を説明変数とする訓練データにより機械学習モデルを生成
S14
機械学習モデルに説明変数を入力して、カテゴリ変数の各成分の確信度を取得
S16
保護属性の値毎に、カテゴリ変数の成分の値順、かつ確信度順にランキング
S18
カテゴリ変数の成分の値を、同一ランクの中央値に修正
S20
One-hot表現の不整合を修正し、修正後データ集合を出力
S22
エンド

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2021/039692

A. CLASSIFICATION OF SUBJECT MATTER

G06N 20/00(2019.01)i
FI: G06N20/00 130

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N20/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2022
Registered utility model specifications of Japan 1996-2022
Published registered utility model applications of Japan 1994-2022

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2021-152838 A (HITACHI LTD) 30 September 2021 (2021-09-30) paragraphs [0001]-[0010]	1-20
A	JP 2020-154828 A (FUJITSU LTD) 24 September 2020 (2020-09-24) paragraphs [0001]-[0021]	1-20
A	WO 2019/234802 A1 (NEC CORPORATION) 12 December 2019 (2019-12-12) paragraph [0032]	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 January 2022	**18 January 2022**
Name and mailing address of the ISA/JP **Japan Patent Office (ISA/JP)** **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2021/039692

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP 2021-152838 A	30 September 2021	(Family: none)	
JP 2020-154828 A	24 September 2020	US 2020/0302324 A1 paragraphs [0001]-[0018]	
WO 2019/234802 A1	12 December 2019	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2021/039692

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06N 20/00(2019.01)i
FI: G06N20/00 130

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

G06N20/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０２２年
日本国実用新案登録公報	１９９６－２０２２年
日本国登録実用新案公報	１９９４－２０２２年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2021-152838 A (株式会社日立製作所) 30.09.2021 (2021-09-30) 段落［０００１］－［００１０］	1-20
A	JP 2020-154828 A (富士通株式会社) 24.09.2020 (2020-09-24) 段落［０００１］－［００２１］	1-20
A	WO 2019/234802 A1 (NEC CORPORATION) 12.12.2019 (2019-12-12) 段落［００３２］	1-20

☐ Ｃ欄の続きにも文献が列挙されている。

☑ パテントファミリーに関する別紙を参照。

＊ 引用文献のカテゴリー
“A” 特に関連のある文献ではなく、一般的技術水準を示すもの
“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
“O” 口頭による開示、使用、展示等に言及する文献
“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの
“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
“Y” 特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
“&” 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
11.01.2022	18.01.2022

名称及びあて先

日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）

川▲崎▼　博章 5B 5284

電話番号 03-3581-1101 内線 3545

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2021/039692

引用文献	公表日	パテントファミリー文献	公表日
JP 2021-152838 A	30.09.2021	（ファミリーなし）	
JP 2020-154828 A	24.09.2020	US 2020/0302324 A1 段落［０００１］－［００１８］	
WO 2019/234802 A1	12.12.2019	（ファミリーなし）	