

(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
2024年4月18日(18.04.2024)



(10) 国際公開番号

WO 2024/079802 A1

(51) 国際特許分類:
G06F 21/62 (2013.01) G06N 20/00 (2019.01)

(21) 国際出願番号: PCT/JP2022/037938

(22) 国際出願日: 2022年10月11日(11.10.2022)

(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: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(72) 発明者: 芝原 俊樹 (SHIBAHARA, Toshiki); 〒1808585 東京都武蔵野市緑町3丁目9-1 1 1 N T T 知的財産センタ内 Tokyo (JP).
三浦 堯之 (MIURA, Takayuki); 〒1808585 東京都武蔵野市緑町3丁目9-1 1 1 N T T 知

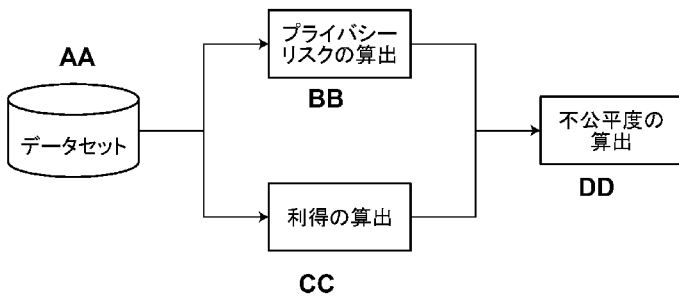
的財産センタ内 Tokyo (JP). 紀伊 真昇 (KII, Masanobu); 〒1808585 東京都武蔵野市緑町3丁目9-1 1 1 N T T 知的財産センタ内 Tokyo (JP). 市川 敦謙 (ICHIKAWA, Atsunori); 〒1808585 東京都武蔵野市緑町3丁目9-1 1 1 N T T 知的財産センタ内 Tokyo (JP).

(74) 代理人: 弁理士法人酒井国際特許事務所 (SAKAI INTERNATIONAL PATENT OFFICE); 〒1000013 東京都千代田区霞が関3丁目8番1号 虎ノ門ダイビルイースト 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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR,

(54) Title: EVALUATION DEVICE, EVALUATION METHOD, AND EVALUATION PROGRAM

(54) 発明の名称: 評価装置、評価方法、および、評価プログラム



AA Data set
BB Calculate privacy risk
CC Calculate gain
DD Calculate inequality

(57) Abstract: This evaluation device calculates a privacy risk of each piece of data included in a data set used for training a machine-learning model. In addition, the evaluation device calculates a gain that a user has obtained by providing the data to the data set. For example, the evaluation device calculates how much the accuracy of the machine-learning model is improved by using the data for training. In addition, the evaluation device calculates, as an inequality degree, the difference between the gain estimated from the privacy risk of the data and a user gain obtained by providing the data, and evaluates the inequality degree of the machine-learning model by using the calculated inequality of each of the users.

(57) 要約: 評価装置は、機械学習モデルの学習に用いられたデータセットに含まれるデータそれぞれのプライバシーリスクを算出する。また、評価装置は、データセットにデータを提供したユーザが、データの提供により得る利得を算出する。例えば、評価装置は、当該データを学習に用いることでどの程度機械学習モデルの精度が向上するかを算出する。そして、評価装置は、データのプライバシーリスクから想定される利得と当該データの提供により得るユーザの利得との差分を、ユーザの不公平度として算出し、算出したユーザそれぞれの不公平度により機械学習モデルの不公平度を評価する。

LS, LU, LY, MA, MD, 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, 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, SC, 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, ME, 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))

明 細 書

発明の名称： 評価装置、評価方法、および、評価プログラム

技術分野

[0001] 本発明は、プライバシーリスクに関する公平性を評価するための、評価装置、評価方法、および、評価プログラムに関する。

背景技術

[0002] Deep Neural Network (DNN) に代表される機械学習技術は、教師データを記憶しやすいという特性からプライバシーリスクがあることが指摘されている。具体的には、特定のデータが教師データに含まれていたか否かを、学習済みモデルの出力から推定できることが示されている。したがって、医療データやウェブの閲覧履歴など、ユーザが他人に知られたくないデータを扱う場合はプライバシーリスクへの配慮が必要である。

[0003] また、機械学習において、教師データとして使用するデータについて、ユーザのプライバシーリスクと、ユーザがデータを提供することの利得（例えば、提供したデータが、学習済みモデルの精度向上にどれくらい役立っているか等）とのバランスを考慮する必要がある。例えば、あるユーザのプライバシーリスクと、当該ユーザがデータを提供することの利得とが釣り合えば、公平であると考えることができる。

先行技術文献

非特許文献

[0004] 非特許文献1：Dwork, C. et al.: Fairness Through Awareness, Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214 - 226 (2012).

非特許文献2：John, P. G., Vijaykeerthy, D. and Saha, D.: Verifying Individual Fairness in Machine Learning Models, Proceedings of the 2020 Conference on Uncertainty in Artificial Intelligence, PMLR, pp. 749 - 758 (2020).

発明の概要

発明が解決しようとする課題

[0005] 従来、機械学習の分野において、分類問題における個人間の公平性を評価する技術（非特許文献 1，2 参照）は存在するが、上記のようなプライバシーリスクに関する公平性を評価する技術は存在しなかった。そこで本発明は、プライバシーリスクに関する公平性を評価することを課題とする。

課題を解決するための手段

[0006] 前記した課題を解決するため、本発明は、機械学習モデルの学習に用いられたデータセットに含まれるデータそれぞれのプライバシーリスクを算出するプライバシーリスク算出部と、前記データセットにデータを提供したユーザが、前記データの提供により得る利得を算出する利得算出部と、前記データのプライバシーリスクと前記データの提供により得るユーザの利得との差分を、前記ユーザの不公平度として算出し、算出した前記ユーザそれぞれの不公平度により前記機械学習モデルの不公平度を評価する不公平度算出部と、前記機械学習モデルの不公平度の評価結果を出力する出力処理部とを備えることを特徴とする。

発明の効果

[0007] 本発明によれば、プライバシーリスクに関する公平性を評価することができる。

図面の簡単な説明

[0008] [図1]図 1 は、評価装置の概要を説明するための図である。

[図2]図 2 は、評価装置の構成例を示す図である。

[図3]図 3 は、評価装置が実行する処理手順の例を示すフローチャートである。

[図4]図 4 は、評価装置の適用例を説明するためのフローチャートである。

[図5]図 5 は、評価プログラムを実行するコンピュータを示す図である。

発明を実施するための形態

[0009] 以下、図面を参照しながら、本発明を実施するための形態（実施形態）について説明する。本発明は、本実施形態に限定されない。

[0010] [概要]

本実施形態の評価装置は、機械学習モデルの構築に用いられるデータセットにデータを提供したユーザがプライバシーリスクに見合う利得が得られているか（公平性）を評価する。

[0011] 例えば、評価装置は、図1に示すように、機械学習モデルの構築に用いられたデータセットからデータを抽出すると、当該データのプライバシーリスクを算出する。また、評価装置は、ユーザがデータセットに当該データを提供することにより得られる利得を算出する。

[0012] そして、評価装置は、ユーザのプライバシーリスクから想定される利得と当該ユーザがデータを提供することにより実際に得られる利得との差分に基づき、各ユーザの不公平度を算出する。ここで、評価装置は、ユーザのプライバシーリスクから想定される利得と当該ユーザがデータを提供することにより実際に得られる利得との差分が大きいほど、不公平度を高く算出する。これにより評価装置は、データセットにデータを提供した各ユーザがプライバシーリスクに見合う利得を得られているか（不公平度）を算出する。そして、評価装置は、算出した各ユーザの不公平度により機械学習モデルの不公平度を評価する。

[0013] [構成例]

次に、図2を用いて、評価装置10の構成例を説明する。評価装置10は、例えば、入出力部11、記憶部12、および、制御部13を備える。

[0014] 入出力部11は、各種データの入出力を司るインターフェースである。入出力部11は、例えば、機械学習モデルの構築に用いるデータセットの入力を受け付ける。入力されたデータセットは記憶部12に格納される。

[0015] 記憶部12は、制御部13が各種処理を実行する際に参照されるデータ、プログラム等を記憶する。記憶部12は、RAM（Random Access Memory）、フラッシュメモリ（Flash Memory）等の半導体メモリ素子、または、ハ

ードディスク、光ディスク等の記憶装置によって実現される。例えば、記憶部 12 は、入出力部 11 で受け付けたデータセット等を記憶する。また、例えば、記憶部 12 は、データセットの各データがどのユーザにより提供されたかを示す情報を記憶してもよい。

[0016] 制御部 13 は、評価装置 10 全体の制御を司る。制御部 13 の機能は、例えば、CPU (Central Processing Unit) が、記憶部 12 に記憶されるプログラムを実行することにより実現される。制御部 13 は、例えば、プライバシーリスク算出部 131 と、利得算出部 132 と、不公平度算出部 133 と、出力処理部 134 とを備える。

[0017] プライバシーリスク算出部 131 は、データセットに含まれるデータそれぞれのプライバシーリスクを算出する。例えば、プライバシーリスク算出部 131 は、以下の式 (1) に基づき、差分プライバシーのパラメータ ϵ の下界 (LB: Lower Bound) を算出することによりプライバシーリスクを算出する。

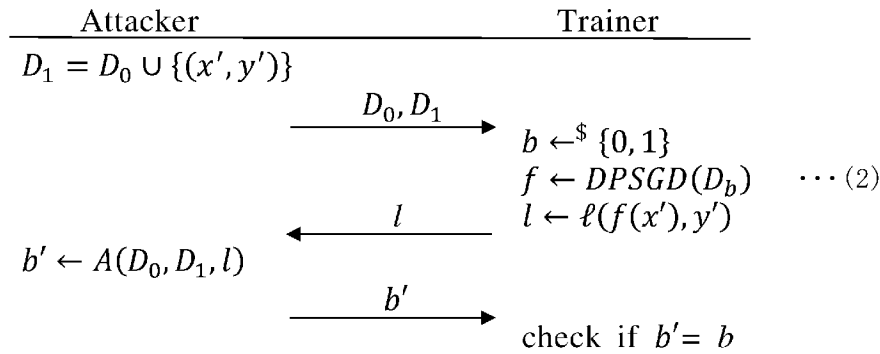
[0018] [数1]

$$\epsilon_{LB} = \max \left(\log \frac{1 - \delta - FPR_{UB}}{FNR_{UB}}, \log \frac{1 - \delta - FNR_{UB}}{FPR_{UB}} \right) \quad \dots (1)$$

[0019] なお、式 (1) における誤検知率の上界 (FPR_{UB}) と見逃し率の上界 (FNR_{UB}) は、評価対象のデータが、モデルの学習に用いられていたか否かを当てるゲームを多数回 (例えば、1000 回程度) 繰り返したときの、誤検知率 (FPR) と見逃し率 (FNR) を用いて算出される (式 (2) 参照)。また、誤検知率 (FPR) と見逃し率 (FNR) それぞれの上界は、Clopper-Pearson method で算出できる。

[0020]

[数2]



[0021] また、プライバシーリスク算出部 1 3 1 は、例えば、メンバーシップ推定を多数回行ったときの成功率等によりプライバシーリスクを算出してもよい。

[0022] 利得算出部 1 3 2 は、ユーザがデータセットにデータを提供することにより得る利得を算出する。ユーザがデータセットにデータを提供することにより得る利益は、例えば、当該データを機械学習モデルの学習に使用した場合に当該機械学習モデルの精度が向上する度合い等である。

[0023] 例えば、利得算出部 1 3 2 は、ユーザの提供したデータを学習に使用した場合に機械学習モデルの精度が向上する度合いを以下のようにして算出する。まず、利得算出部 1 3 2 は、ユーザの提供したデータを学習に使用したシャドウモデル（第 1 のシャドウモデル）と当該データを学習に使用しなかったシャドウモデル（第 2 のシャドウモデル）とをそれぞれ n 個構築する。

[0024] 次に、利得算出部 1 3 2 は、 n 個の第 1 のシャドウモデルのうち、正解データを出力した第 1 のシャドウモデルの数を c_{in} とする。そして、 c_{in} を n で割った値 (c_{in}/n) を第 1 のシャドウモデルの精度とする。また、利得算出部 1 3 2 は、 n 個の第 2 のシャドウモデルのうち、正解データを出力した第 2 のシャドウモデルの数を c_{out} とする。そして、 c_{out} を n で割った値 (c_{out}/n) を第 2 のシャドウモデルの精度とする。

[0025] そして、利得算出部 1 3 2 は、第 1 のシャドウモデルの精度 (c_{in}/n) と第 2 のシャドウモデルの精度 (c_{out}/n) との差を、ユーザがデータを利得 (g) として算出する（式 (3) 参照）。

[0026] [数3]

$$g = \frac{c_{in}}{n} - \frac{c_{out}}{n} \quad \dots (3)$$

[0027] なお、利得算出部 1 3 2 は、上記の方法以外にも、ユーザが保有する他のデータがシャドウモデルの精度をどの程度改善するかにより利得を算出してもよい。また、ユーザの利得は、ユーザが提供するデータと引き換えに提供されるサービスや金銭的報酬でもよい。

[0028] 不公平度算出部 1 3 3 は、データセットにデータを提供したユーザそれぞれの不公平度を算出する。そして、不公平度算出部 1 3 3 は、算出したユーザそれぞれの不公平度により、機械学習モデルの不公平度を評価する。例えば、不公平度算出部 1 3 3 は、プライバシーリスク算出部 1 3 1 により算出されたデータのプライバシーリスクから想定されるユーザの利得と、利得算出部 1 3 2 により算出された当該データの提供により得られたユーザの利得との差分を、当該ユーザの不公平度として算出する。そして、不公平度算出部 1 3 3 は、算出された各ユーザの不公平度のうち、その最大値を、機械学習モデルの不公平度とする。

[0029] 例えば、全ユーザーを $U = \{u_i\}_{i=1}^{n_1}$ とする。この全ユーザのリスクを $R = \{r_i\}_{i=1}^{n_1}$ 、利得を $G = \{g_i\}_{i=1}^{n_1}$ とする。また、平均=0、分散=1に正規化したリスクを R' 、利得を G' とする。そして、不公平度算出部 1 3 3 は、ユーザごとに、正規化されたリスク r' と正規化された利得 g' との差を算出し、その最大値を機械学習モデルの不公平度 (δ') とする (式 (4) 参照)。

[0030] [数4]

$$\hat{\delta} = \max_{u_i \in U} |g'_i - r'_i| \quad \dots (4)$$

[0031] なお、不公平度算出部 1 3 3 は、各ユーザの不公平度にはずれ値が含まれていた場合、そのはずれ値を除外した不公平度の中での最大値を機械学習モデルの不公平度としてもよい。

[0032] 出力処理部 1 3 4 は、制御部 1 3 による処理結果を出力する。例えば、出力処理部 1 3 4 は、不公平度算出部 1 3 3 による機械学習モデルの不公平度

の評価結果を出力する。

[0033] このような評価装置 10 によれば、機械学習モデルの不公平度を評価することができる。

[0034] [処理手順の例]

次に、図 3 を用いて、評価装置 10 が実行する処理手順の例を説明する。
まず、評価装置 10 のプライバシーリスク算出部 131 は、データセットに含まれるデータそれぞれのプライバシーリスクを算出する (S1)。

[0035] 次に、利得算出部 132 は、ユーザがデータセットにデータを提供することにより得た利得を算出する (S2)。例えば、利得算出部 132 は、ユーザが提供したデータを学習に使用した場合にシャドウモデルの精度がどの程度改善するかを算出する。

[0036] 次に、不公平度算出部 133 は、機械学習モデルの不公平度を評価する (S3)。例えば、不公平度算出部 133 は、S1 で算出されたデータのプライバシーリスクから想定される利得と、S3 で算出された当該データの提供により得られたユーザの実際の利得との差分を、当該ユーザの不公平度として算出する。そして、不公平度算出部 133 は、算出した各ユーザの不公平度のうち、その最大値を、機械学習モデルの不公平度とする。そして、出力処理部 134 は、S3 で得られた機械学習モデルの不公平度の評価結果を出力する (S4)。

[0037] 評価装置 10 が上記の処理を行うことで、機械学習モデルの不公平度を評価することができる。

[0038] [適用例]

次に、図 4 を用いて、評価装置 10 の適用例を説明する。例えば、評価装置 10 の管理者はプライバシーリスクの算出対象となる NN (ニューラルネットワーク) を設計する (S11)。例えば、DP-SGD (Differentially Private Stochastic Gradient Descent) により、Differential Privacy を満たす NN のトレーニングを行う。

[0039] 次に、管理者は、ユーザの不公平度の評価に用いるデータセットと不公平

度の評価対象のユーザの選定を行う（S 1 2）。例えば、管理者は、ユーザの多様性を考慮し100人程度のユーザを選定する。

[0040] その後、評価装置 1 0 は、S 1 2 で選定されたデータセットを用いて、S 1 2 で選定された各ユーザの不公平度を算出し、算出した各ユーザの不公平度により機械学習モデル（NN）の不公平度を評価する（S 1 3）。例えば、評価装置 1 0 は、S 1 1 で設計されたNNに対するデータセットの各データのプライバシーリスクを算出する。また、評価装置 1 0 はS 1 2 で選定されたユーザ（データセットにデータを提供したユーザ）の利得を算出する。そして、評価装置 1 0 は、算出した各データのプライバシーリスクとデータを提供した各ユーザの利得に基づき、各ユーザの不公平度を算出する。その後、評価装置 1 0 は、算出した各ユーザの不公平度のうち、その最大値を機械学習モデル（NN）の不公平度とする。

[0041] [システム構成等]

また、図示した各部の各構成要素は機能概念的なものであり、必ずしも物理的に図示のように構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部又は一部を、各種の負荷や使用状況等に応じて、任意の単位で機能的又は物理的に分散・統合して構成することができる。さらに、各装置にて行われる各処理機能は、その全部又は任意の一部が、CPU及び当該CPUにて実行されるプログラムにて実現され、あるいは、ワイヤードロジックによるハードウェアとして実現され得る。

[0042] また、前記した実施形態において説明した処理のうち、自動的に行われるものとして説明した処理の全部又は一部を手動的に行うこともでき、あるいは、手動的に行われるものとして説明した処理の全部又は一部を公知の方法で自動的に行うこともできる。この他、上記文書中や図面中で示した処理手順、制御手順、具体的名称、各種のデータやパラメータを含む情報については、特記する場合を除いて任意に変更することができる。

[0043] [プログラム]

前記した評価装置１０は、パッケージソフトウェアやオンラインソフトウェアとしてプログラム（評価プログラム）を所望のコンピュータにインストールさせることによって実装できる。例えば、上記のプログラムを情報処理装置に実行させることにより、情報処理装置を評価装置１０として機能させることができる。ここで言う情報処理装置にはスマートフォン、携帯電話機やＰＨＳ（Personal Handyphone System）等の移動体通信端末、さらには、ＰＤＡ（Personal Digital Assistant）等の端末等がその範疇に含まれる。

[0044] 図５は、評価プログラムを実行するコンピュータの一例を示す図である。コンピュータ１０００は、例えば、メモリ１０１０、ＣＰＵ１０２０を有する。また、コンピュータ１０００は、ハードディスクドライバインタフェース１０３０、ディスクドライバインタフェース１０４０、シリアルポートインタフェース１０５０、ビデオアダプタ１０６０、ネットワークインタフェース１０７０を有する。これらの各部は、バス１０８０によって接続される。

[0045] メモリ１０１０は、ＲＯＭ（Read Only Memory）１０１１及びＲＡＭ（Random Access Memory）１０１２を含む。ＲＯＭ１０１１は、例えば、ＢＩＯＳ（Basic Input Output System）等のブートプログラムを記憶する。ハードディスクドライバインタフェース１０３０は、ハードディスクドライブ１０９０に接続される。ディスクドライバインタフェース１０４０は、ディスクドライブ１１００に接続される。例えば磁気ディスクや光ディスク等の着脱可能な記憶媒体が、ディスクドライブ１１００に挿入される。シリアルポートインタフェース１０５０は、例えばマウス１１１０、キーボード１１２０に接続される。ビデオアダプタ１０６０は、例えばディスプレイ１１３０に接続される。

[0046] ハードディスクドライブ１０９０は、例えば、ＯＳ１０９１、アプリケーションプログラム１０９２、プログラムモジュール１０９３、プログラムデータ１０９４を記憶する。すなわち、上記の評価装置１０が実行する各処理

を規定するプログラムは、コンピュータにより実行可能なコードが記述されたプログラムモジュール1093として実装される。プログラムモジュール1093は、例えばハードディスクドライブ1090に記憶される。例えば、評価装置10における機能構成と同様の処理を実行するためのプログラムモジュール1093が、ハードディスクドライブ1090に記憶される。なお、ハードディスクドライブ1090は、SSD (Solid State Drive) により代替されてもよい。

[0047] また、上述した実施形態の処理で用いられるデータは、プログラムデータ1094として、例えばメモリ1010やハードディスクドライブ1090に記憶される。そして、CPU1020が、メモリ1010やハードディスクドライブ1090に記憶されたプログラムモジュール1093やプログラムデータ1094を必要に応じてRAM1012に読み出して実行する。

[0048] なお、プログラムモジュール1093やプログラムデータ1094は、ハードディスクドライブ1090に記憶される場合に限らず、例えば着脱可能な記憶媒体に記憶され、ディスクドライブ1100等を介してCPU1020によって読み出されてもよい。あるいは、プログラムモジュール1093及びプログラムデータ1094は、ネットワーク (LAN (Local Area Network)、WAN (Wide Area Network) 等) を介して接続される他のコンピュータに記憶されてもよい。そして、プログラムモジュール1093及びプログラムデータ1094は、他のコンピュータから、ネットワークインタフェース1070を介してCPU1020によって読み出されてもよい。

符号の説明

- [0049] 10 評価装置
 11 入出力部
 12 記憶部
 13 制御部
 131 プライバシーリスク算出部
 132 利得算出部

1 3 3 不公平度算出部

1 3 4 出力処理部

請求の範囲

- [請求項1] 機械学習モデルの学習に用いられたデータセットに含まれるデータそれぞれのプライバシーリスクを算出するプライバシーリスク算出部と、
- 前記データセットにデータを提供したユーザが、前記データの提供により得る利得を算出する利得算出部と、
- 前記データのプライバシーリスクと前記データの提供により得るユーザの利得との差分を、前記ユーザの不公平度として算出し、算出した前記ユーザそれぞれの不公平度により前記機械学習モデルの不公平度を評価する不公平度算出部と、
- 前記機械学習モデルの不公平度の評価結果を出力する出力処理部とを備えることを特徴とする評価装置。
- [請求項2] 前記不公平度算出部は、
- 前記データのプライバシーリスクと前記ユーザの利得それぞれを正規化し、正規化した前記データのプライバシーリスクと前記ユーザの利得との差分を、前記ユーザの不公平度とすることを特徴とする請求項1に記載の評価装置。
- [請求項3] 前記不公平度算出部は、
- 算出した前記ユーザそれぞれの不公平度のうち、その最大値を前記機械学習モデルの評価結果とすることを特徴とする請求項1に記載の評価装置。
- [請求項4] 前記データの提供により得る利得は、
- 前記データを機械学習モデルの学習に用いた場合に、前記機械学習モデルの精度が向上する度合いである
- ことを特徴とする請求項1に記載の評価装置。
- [請求項5] 前記利得算出部は、
- 前記データを学習に用いたシャドウモデルと前記データを学習に用いなかったシャドウモデルを構築し、前記データを学習に用いたシャ

ドウモデルの精度と前記データを学習に用いなかったシャドウモデルの精度との差を算出することにより、前記データを学習に用いた場合に、前記機械学習モデルの精度が向上する度合いを算出する

ことを特徴とする請求項4に記載の評価装置。

[請求項6]

前記データを提供することにより得る利得は、

前記データの提供と引き換えに前記ユーザに提供されるサービスまたは金銭的な報酬である

ことを特徴とする請求項1に記載の評価装置。

[請求項7]

評価装置により実行される評価方法であって、

機械学習モデルの学習に用いられたデータセットに含まれるデータそれぞれのプライバシーリスクを算出する工程と、

前記データセットにデータを提供したユーザが、前記データの提供により得る利得を算出する工程と、

前記データのプライバシーリスクと前記データの提供により得るユーザの利得との差分を、前記ユーザの不公平度として算出し、算出した前記ユーザそれぞれの不公平度により前記機械学習モデルの不公平度を評価する工程と、

前記機械学習モデルの不公平度の評価結果を出力する工程と

を含むことを特徴とする評価方法。

[請求項8]

機械学習モデルの学習に用いられたデータセットに含まれるデータそれぞれのプライバシーリスクを算出する工程と、

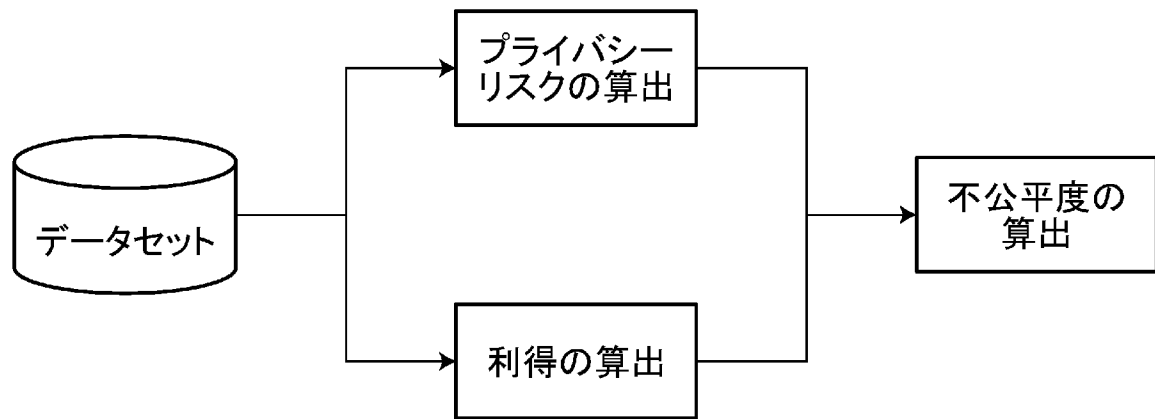
前記データセットにデータを提供したユーザが、前記データの提供により得る利得を算出する工程と、

前記データのプライバシーリスクと前記データの提供により得るユーザの利得との差分を、前記ユーザの不公平度として算出し、算出した前記ユーザそれぞれの不公平度により前記機械学習モデルの不公平度を評価する工程と、

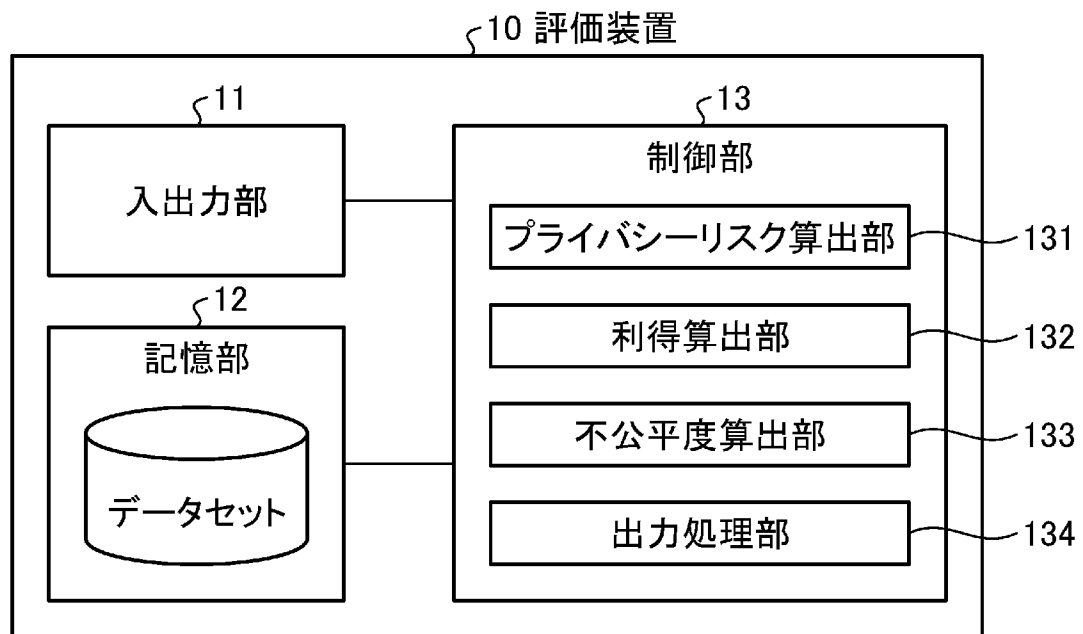
前記機械学習モデルのの評価結果を出力する工程と

をコンピュータに実行させるための評価プログラム。

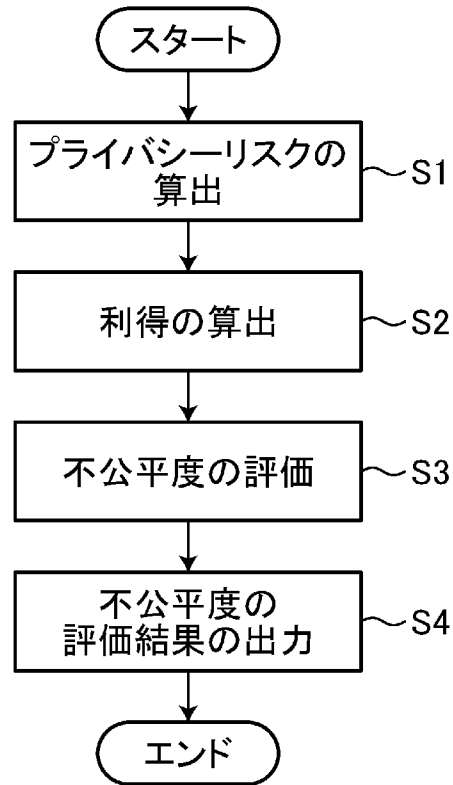
[図1]



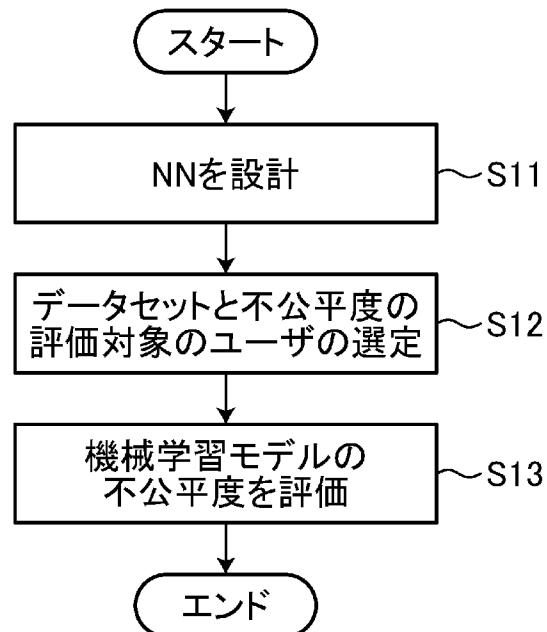
[図2]



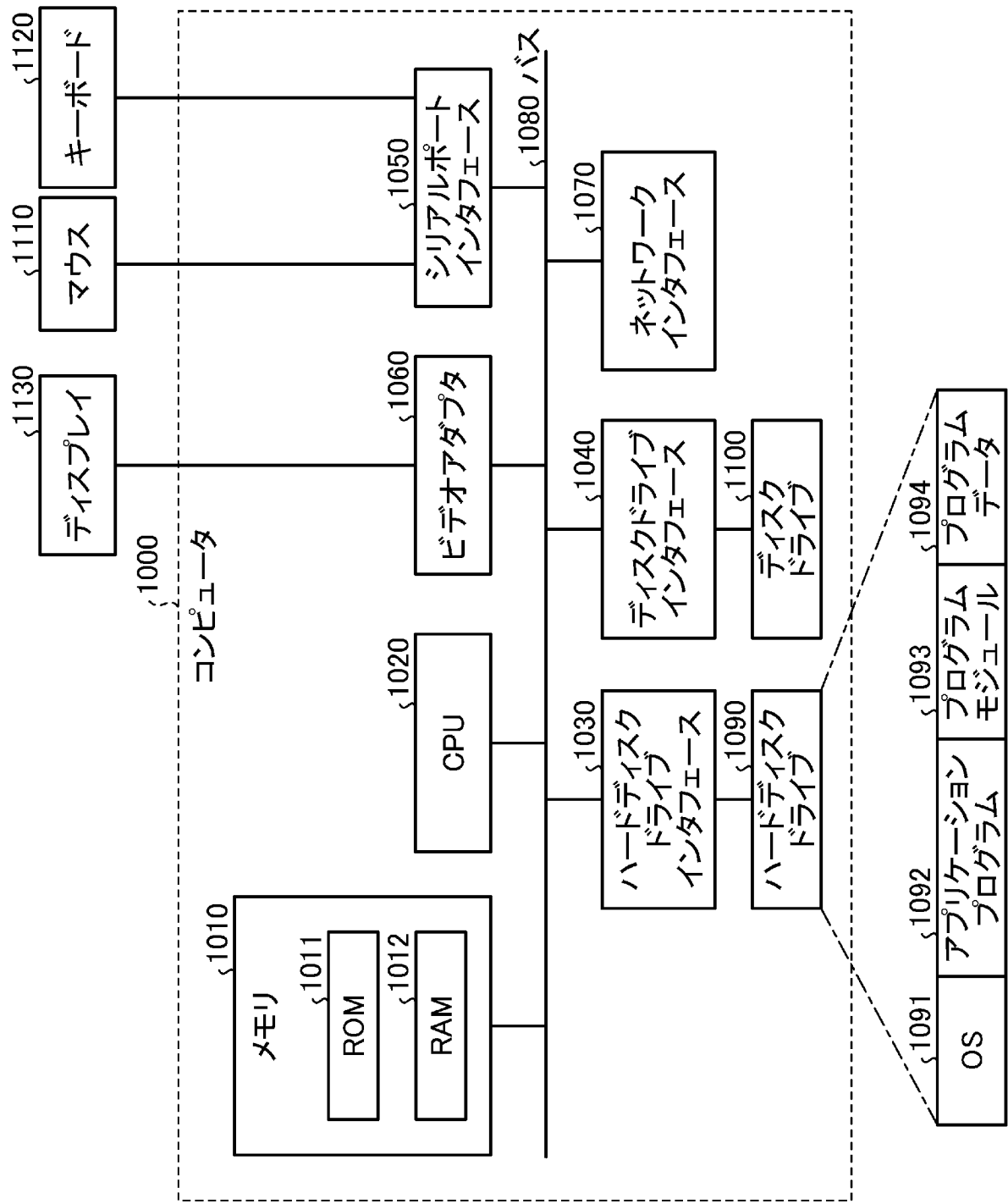
[図3]



[図4]



[図5]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2022/037938

A. CLASSIFICATION OF SUBJECT MATTER**G06F 21/62**(2013.01)i; **G06N 20/00**(2019.01)i

FI: G06F21/62 345; G06N20/00

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)

G06F21/62; 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	芝原 俊樹 ほか, 差分プライバシーなベイジアンニューラルネットワークのプライバシーリスク, コンピュータセキュリティシンポジウム2021論文集, October 2021, pp. 245-252 pp. 247-248, 2.4, The lower bound on privacy risk, (SHIBAHARA, Toshiki et al. Privacy Risk of Differentially Private Bayesian Neural Network.), non-official translation (Proceedings of Computer Security Symposium 2021)	1-8
A	JP 2015-130022 A (KDDI CORP) 16 July 2015 (2015-07-16) claim 1	1-8

☐ 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

09 November 2022

Date of mailing of the international search report

22 November 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/JP2022/037938

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP 2015-130022 A	16 July 2015	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06F 21/62(2013.01)i; G06N 20/00(2019.01)i FI: G06F21/62 345; G06N20/00										
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）） G06F21/62; 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										
<table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2022年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2022年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2022年</td> </tr> </table>			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
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	芝原 俊樹 ほか、差分プライバシーなベイジアンニューラルネットワークのプライバシーリスク、コンピュータセキュリティシンポジウム2021論文集, 2021.10, pp.245-252 pp.247-248の2.4 プライバシーリスクの下界	1-8								
A	JP 2015-130022 A (KDDI株式会社) 16.07.2015 (2015 - 07 - 16) 請求項1	1-8								
☐ 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。										
<table border="0"> <tr> <td> * 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技术水準を示すもの “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 </td> <td> “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 </td> </tr> </table>			* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技术水準を示すもの “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						
* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技术水準を示すもの “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 09.11.2022	国際調査報告の発送日 22.11.2022									
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） 局 成矢 5S 1594 電話番号 03-3581-1101 内線 3546									

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/JP2022/037938

引用文献	公表日	パテントファミリー文献	公表日
JP 2015-130022 A	16.07.2015	(ファミリーなし)	