(19)日本国特許庁(JP)

# (12)特許公報(B2)

(11)特許番号
特許第7363662号
(P7363662)

(45)発行日　令和5年10月18日(2023.10.18)

(24)登録日　令和5年10月10日(2023.10.10)

(51)国際特許分類
*Ｇ０６Ｆ　21/62　(2013.01)*

ＦＩ
Ｇ０６Ｆ　21/62　３５４

請求項の数　6　（全20頁）

(21)出願番号　特願2020-79550(P2020-79550)
(22)出願日　令和2年4月28日(2020.4.28)
(65)公開番号　特開2021-174390(P2021-174390A)
(43)公開日　令和3年11月1日(2021.11.1)
審査請求日　令和5年1月12日(2023.1.12)

(73)特許権者　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
(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　福岡 尊
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
(72)発明者　山岡 裕司
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
審査官　宮司 卓佳

最終頁に続く

(54)【発明の名称】　生成方法，情報処理装置及び生成プログラム

(57)【特許請求の範囲】
【請求項１】
　複数の項目値を含む複数の個人情報を受け付け、
　前記複数の項目値それぞれに対応付けられた第１のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成し、
　前記複数の項目値それぞれに対応付けられた前記第１のパラメータとは異なる第２のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成し、
　前記第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する、
処理をコンピュータが実行することを特徴とする生成方法。
【請求項２】
　前記第１及び第２のパラメータについての多様性に関する指標が最大化されるように、当該第１及び第２のパラメータを決定する、
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に記載の生成方法。
【請求項３】
　前記第１及び第２のパラメータは、前記複数の項目値間の優先順序である、
ことを特徴とする、請求項１又は２に記載の生成方法。

【請求項４】
　前記第１及び第２のパラメータは、前記複数の項目値それぞれの重み付け値である、
ことを特徴とする、請求項１又は２に記載の生成方法。
【請求項５】
　複数の項目値を含む複数の個人情報を受け付ける受付処理部と、
　前記複数の項目値それぞれに対応付けられた第１のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成する第１匿名情報生成部と、
　前記複数の項目値それぞれに対応付けられた前記第１のパラメータとは異なる第２のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成する第２匿名情報生成部と、
　前記第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する結合情報生成部と、
を備えることを特徴とする情報処理装置。
【請求項６】
　複数の項目値を含む複数の個人情報を受け付け、
　前記複数の項目値それぞれに対応付けられた第１のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成し、
　前記複数の項目値それぞれに対応付けられた前記第１のパラメータとは異なる第２のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成し、
　前記第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する、
処理をコンピュータに実行させることを特徴とする生成プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
　本発明は、生成方法，情報処理装置及び生成プログラムに関する。
【背景技術】
【０００２】
　顧客から預かったデータをもとに、人工知能（ＡＩ）製品を展開するビジネスが存在する。このようなＡＩ製品としては、例えば、顧客から預かったデータを利用して機械学習し、事象を予測するモデルがある。
【０００３】
　図１は、ＡＩ製品の機械学習を例示する図である。
【０００４】
　図１のＡ１に示すＩｄと職業と性別と年収とが対応付けられたデータに対して、年収を目的変数として機械学習を実行することにより、符号Ａ２に示すように、職業及び性別から、年収を予測するモデルが生成される。
【０００５】
　利用するデータが個人情報である場合には、匿名化処理が実行されることがある。
【先行技術文献】
【特許文献】
【０００６】
【文献】特開２０１７－１８２５０８号公報
再表２０１３－１１４４４５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、匿名化処理された匿名化データを学習で利用すると、匿名化データはオリジナルのデータよりも情報量が低下するため、生成されるモデルの精度が低下するおそれがある。
【０００８】
１つの側面では、機械学習によって生成する学習モデルの精度を向上させることを目的とする。
【課題を解決するための手段】
【０００９】
１つの側面では、生成方法は、複数の項目値を含む複数の個人情報を受け付け、前記複数の項目値それぞれに対応付けられた第１のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成し、前記複数の項目値それぞれに対応付けられた前記第１のパラメータとは異なる第２のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成し、前記第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する、処理をコンピュータが実行する。
【発明の効果】
【００１０】
１つの側面では、機械学習によって生成する学習モデルの精度を向上できる。
【図面の簡単な説明】
【００１１】
【図１】ＡＩ製品の機械学習を例示する図である。
【図２】実施形態の一例におけるｋ－匿名化を例示するテーブルである。
【図３】実施形態の一例における情報処理装置のハードウェア構成例を模式的に示すブロック図である。
【図４】図３に示した情報処理装置のソフトウェア構成例を模式的に示すブロック図である。
【図５】関連例における抑制ｋ－匿名化を例示するテーブルである。
【図６】関連例における一般化匿名化を例示するテーブルである。
【図７】図６に示したテーブルに対応する一般化木を例示する図である。
【図８】実施形態の一例におけるｋ－匿名化データの結合例を示す図である。
【図９】実施形態の一例における匿名化データの生成処理を説明するフローチャートである。
【図１０】実施形態の一例における抑制ｋ－匿名化処理を利用した匿名化処理の生成処理を説明するフローチャートである。
【図１１】図１０に示したオリジナルデータからの匿名化データの結合例を示す図である。
【図１２】実施形態の一例における一般化匿名化処理を利用した匿名化処理の生成処理を説明するフローチャートである。
【図１３】図１２に示したオリジナルデータからの匿名化データの結合例を示す図である。
【図１４】関連例における学習モデルの生成処理を模式的に示す図である。
【図１５】実施形態の一例における学習モデルの生成処理を模式的に示すブロック図である。
【図１６】関連例における学習モデルと実施形態の一例における学習モデルとの精度を比較するテーブルである。
【図１７】実施形態の一例におけるデータの削除箇所を例示する図である。

【図１８】実施形態の一例におけるデータの削除箇所の決定アルゴリズムの第１の例を説明するテーブルである。
【図１９】学習モデルの生成処理の違いによる学習モデルの精度を比較するテーブルである。
【発明を実施するための形態】
【００１２】
〔Ａ〕実施形態
以下、図面を参照して一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とができる。
【００１３】
以下、図中において、同一の各符号は同様の部分を示しているので、その説明は省略する。
【００１４】
〔Ａ－１〕概要
図２は、実施形態の一例におけるｋ－匿名化を例示するテーブルである。
【００１５】
実施形態の一例においては、分類モデルの精度を下げないｋ－匿名化データが構築される。ｋ－匿名化データとは、同一レコードが少なくともｋ個ある個人を特定できないデータである。ただし、全て欠損しているレコードは無視される。ｋ－匿名化が施されているか否かは容易にチェックできる。
【００１６】
図２においては、符号Ｂ１に示すオリジナルデータを入力として、２－匿名化により、符号Ｂ２に示す匿名化データ（別言すれば、匿名情報）が構築される。
【００１７】
ｋ－匿名化においては、「どの属性に関する情報を残したいか」といった属性に関する設定が、入力として求められる場合がある。図示する例では、年収及び性別ができるだけ残されるように、匿名化が実行されている。
【００１８】
図３は、実施形態の一例における情報処理装置１のハードウェア構成例を模式的に示すブロック図である。
【００１９】
図３に示すように、情報処理装置１は、Central Processing Unit（ＣＰＵ）１１，メモリ部１２，表示制御部１３，記憶装置１４，入力Interface（ＩＦ）１５，外部記録媒体処理部１６及び通信ＩＦ１７を備える。
【００２０】
メモリ部１２は、記憶部の一例であり、例示的に、Read Only Memory（ＲＯＭ）及びRandom Access Memory（ＲＡＭ）などである。メモリ部１２のＲＯＭには、Basic Input/Output System（ＢＩＯＳ）等のプログラムが書き込まれてよい。メモリ部１２のソフトウェアプログラムは、ＣＰＵ１１に適宜に読み込まれて実行されてよい。また、メモリ部１２のＲＡＭは、一時記録メモリあるいはワーキングメモリとして利用されてよい。
【００２１】
表示制御部１３は、表示装置１３０と接続され、表示装置１３０を制御する。表示装置１３０は、液晶ディスプレイやOrganic Light-Emitting Diode（ＯＬＥＤ）ディスプレイ，Cathode Ray Tube（ＣＲＴ），電子ペーパーディスプレイ等であり、オペレータ等に対する各種情報を表示する。表示装置１３０は、入力装置と組み合わされたものでもよく、例えば、タッチパネルでもよい。

【００２２】
記憶装置１４は、高ＩＯ性能の記憶装置であり、例えば、Dynamic Random Access Memory（ＤＲＡＭ）やSolid State Drive（ＳＳＤ），Storage Class Memory（ＳＣＭ），Hard Disk Drive（ＨＤＤ）が用いられてよい。
【００２３】
入力ＩＦ１５は、マウス１５１やキーボード１５２等の入力装置と接続され、マウス１５１やキーボード１５２等の入力装置を制御してよい。マウス１５１やキーボード１５２は、入力装置の一例であり、これらの入力装置を介して、オペレータが各種の入力操作を行なう。
【００２４】
外部記録媒体処理部１６は、記録媒体１６０が装着可能に構成される。外部記録媒体処理部１６は、記録媒体１６０が装着された状態において、記録媒体１６０に記録されている情報を読み取り可能に構成される。本例では、記録媒体１６０は、可搬性を有する。例えば、記録媒体１６０は、フレキシブルディスク、光ディスク、磁気ディスク、光磁気ディスク、又は、半導体メモリ等である。
【００２５】
通信ＩＦ１７は、外部装置との通信を可能にするためのインタフェースである。
【００２６】
ＣＰＵ１１は、種々の制御や演算を行なう処理装置であり、メモリ部１２に格納されたOperating System（ＯＳ）やプログラムを実行することにより、種々の機能を実現する。
【００２７】
情報処理装置１全体の動作を制御するための装置は、ＣＰＵ１１に限定されず、例えば、ＭＰＵやＤＳＰ，ＡＳＩＣ，ＰＬＤ，ＦＰＧＡのいずれか１つであってもよい。また、情報処理装置１全体の動作を制御するための装置は、ＣＰＵ，ＭＰＵ，ＤＳＰ，ＡＳＩＣ，ＰＬＤ及びＦＰＧＡのうちの２種類以上の組み合わせであってもよい。なお、ＭＰＵはMicro Processing Unitの略称であり、ＤＳＰはDigital Signal Processorの略称であり、ＡＳＩＣはApplication Specific Integrated Circuitの略称である。また、ＰＬＤはProgrammable Logic Deviceの略称であり、ＦＰＧＡはField Programmable Gate Arrayの略称である。
【００２８】
図４は、図３に示した情報処理装置１のソフトウェア構成例を模式的に示す図である。
【００２９】
情報処理装置１は、受付処理部１１１，匿名情報生成部１１２及び結合情報生成部１１３として機能する。
【００３０】
受付処理部１１１は、複数の項目値を含む複数の個人情報を受け付ける。
【００３１】
匿名情報生成部１１２は、複数の項目値それぞれに対応付けられたパラメータに応じて複数の項目値それぞれの秘匿化されやすさが決まる秘匿化アルゴリズムを、受け付けた複数の個人情報に適用して、複数の項目値の少なくとも何れかの項目値が匿名化された匿名情報を生成する。具体的には、匿名情報生成部１１２は、複数の項目値それぞれに対応付けられた第１のパラメータに応じて複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた複数の個人情報に適用して、複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成する第１匿名情報生成部の一例として機能する。また、匿名情報生成部１１２は、複数の項目値それぞれに対応付けられた第１のパラメータとは異なる第２のパラメータに応じて複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた複数の個人情報に適用して、複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成する第２匿名情報生成部の一例として機能する。なお、詳細は後述されるが、開示の技術におけるパラメータまたは秘匿化アルゴリズムは２つに限定されるわけではなく、３つ以上で

あってもよい。
【００３２】
結合情報生成部１１３は、第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する。
【００３３】
〔Ａ－２〕匿名化処理
ｋ－匿名化においては、データ品質を表す量を定義し、その量が最大化される。データの品質を表す量としては、例えば、加工（削除等）されなかったセルの数やエントロピーがある。
【００３４】
しかしながら、データ品質を表す量が大きくても、うまく学習モデルを構成できる保証はないため、出力を学習データとして用いると精度劣化が大きくなるおそれがある。モデル学習に最適な匿名化データを見つけることは計算量的に容易でなく、現実的ではない。
【００３５】
図５は、関連例における抑制ｋ－匿名化を例示するテーブルである。
【００３６】
抑制ｋ－匿名化においては、セルを削除することにより、ｋ－匿名化を実現することができる。符号Ｃ１に示すような表形式のオリジナルデータと、ｋ（正の整数）と、属性の順番付け（属性優先順序と称されてもよい。）とが入力されると、符号Ｃ２及びＣ３に示すような表形式のｋ－匿名化データが出力される。
【００３７】
符号Ｃ２に示すｋ－匿名化データでは、属性優先順序が「年収＞性別＞職業」に設定されている。また、符号Ｃ３に示すｋ－匿名化データでは、属性優先順序が「年収＞職業＞性別」に設定されている。
【００３８】
抑制ｋ－匿名化においては、匿名化の際に、属性優先順序が「属性に関する設定」として使用される。これにより、属性優先順序が高い属性ほど、データが残りやすいように匿名化が実行される。
【００３９】
図６は、関連例における一般化匿名化を例示するテーブルである。図７は、図６に示したテーブルに対応する一般化木を例示する図である。
【００４０】
一般化ｋ－匿名化においては、抑制ｋ－匿名化におけるセルの削除に加えて、セルの置き換えも実施される。
【００４１】
図６の符号Ｄ１におけるオリジナルデータは、符号Ｄ２及びＤ３に示すように、一般化されたｋ－匿名化データとして出力される。符号Ｄ２に示すｋ－匿名化データでは、職業の一般化が行なわれている。また、符号Ｄ３に示すｋ―匿名化データでは、住所の一般化が行なわれている。
【００４２】
図７の符号Ｅ１に示す一般化木においては、「杉並区」，「世田谷区」及び「目黒区」の一般化概念として「東京都」が定義されており、「横浜市」及び「川崎市」の一般化概念として「神奈川県」が定義されている。
【００４３】
また、図７の符号Ｅ２に示す一般化木においては、「飲食店」及び「美容師」の一般化概念として「自営業」が定義されており、「開発」及び「営業」の一般化概念として「会社員」が定義されており、「教授」及び「教諭」の一般化概念として「教育職」が定義されている。
【００４４】
なお、図７の符号Ｅ３に示す一般化木において、年収の「５００万円以上」及び「５０

０万円未満」については、一般化概念が定義されていない。

【００４５】

匿名化は、図７に示した一般化木に加えて、各属性の重みも入力とし、次式のNCPを最小化するように実行されてよい。

【００４６】

【数１】

$$NCP = \sum_A (A\text{の重み}) \sum_T \frac{T\text{の}A\text{における値を先祖とするリーフの数}}{A\text{に対応する一般化木のリーフの数}}$$

なお、Aは属性であり、Tはレコードである。重みが大きい（別言すれば、重要な）属性は、匿名化されづらくなる。

【００４７】

ここで、重みを(住所,職業,年収)= (a,b,c)とすると、図６の符号Ｄ２に示した匿名化データのNCPは5a+(5/3)b+4cとなり、図６の符号Ｄ３に示した匿名化データのNCPは2a+5b+3cとなる。

【００４８】

(a,b,c) = (0.1,0.6,0.3)とすると、図６の符号Ｄ２に示した匿名化データのNCPは5*0.1+(5/3)*0.6+4*0.3 = 2.7となり、図６の符号Ｄ３に示した匿名化データのNCPは2*0.1+5*0.6+3*0.3 = 4.1となる。すなわち、図６の符号Ｄ２に示した匿名化データの方がNCPが低くなる。

【００４９】

一方、(a,b,c) = (0.3,0.1,0.6)とすると、図６の符号Ｄ２に示した匿名化データのNCPは5*0.3+(5/3)*0.1+4*0.6 = 4.0666…となり、図６の符号Ｄ３に示した匿名化データのNCPは2*0.3+5*0.1+3*0.6 = 2.9となる。すなわち、図６の符号Ｄ３に示した匿名化データの方がNCPが低くなる。

【００５０】

〔Ａ－３〕結合処理

図８は、実施形態の一例におけるｋ－匿名化データの結合例を示す図である。

【００５１】

実施形態の一例において、オリジナルデータを機械学習するにあたって、レコード数は保たれなくてもよい。そこで、複数の異なるｋ－匿名化データが結合されることで、機械学習に適した匿名化データが生成される。

【００５２】

符号Ｆ１に示すオリジナルデータを入力として、符号Ｆ２に示すように２つのｋ－匿名化データが出力される。そして、符号Ｆ３に示すように、２つのｋ－匿名化データが結合されることにより、学習モデルが生成される。

【００５３】

図９は、実施形態の一例における匿名化データの生成処理を説明するフローチャートである。図９に示すフローチャート（ステップＳ１～Ｓ３）に従って、匿名化データの生成処理を説明する。

【００５４】

必要に応じて優先順序等の補助入力を受け付け、属性に関するパラメータとして、匿名化アルゴリズム設定＃１～＃ｎが生成される（ステップＳ１）。

【００５５】

表形式のオリジナルデータの入力を受け付け、各匿名化アルゴリズム設定＃１～＃ｎに応じた匿名化アルゴリズムで匿名化処理が実行され、匿名化データ＃１～＃ｎが出力される（ステップＳ２）。

【００５６】

匿名化データ＃１～＃ｎについて互いに結合処理が実行され、結合された匿名化データが出力される（ステップＳ３）。そして、匿名化データの生成処理は終了する。

【００５７】

図１０は、実施形態の一例における抑制ｋ－匿名化処理を利用した匿名化処理の生成処理を説明するフローチャートである。図１１は、図１０に示したオリジナルデータからの匿名化データの結合例を示す図である。図１０に示すフローチャート（ステップＳ１１～Ｓ１３）に従って、抑制ｋ－匿名化処理を利用した匿名化処理の生成処理を説明する。

【００５８】

表形式のオリジナルデータが入力として受け付けられる（ステップＳ１１）。

【００５９】

属性に関するパラメータ（別言すれば、属性優先順序）が、匿名化データの生成個数と順序を固定したい属性とを追加入力とした上で、ランダムに生成される（ステップＳ１２）。

【００６０】

生成した各属性優先順序の設定の元でそれぞれの匿名化データが出力され、得られた匿名化データが結合される（ステップＳ１３）。そして、抑制ｋ－匿名化処理を利用した匿名化処理の生成処理は終了する。

【００６１】

図１１に示す例では、生成個数が「２」に設定され、「年収」の属性優先順序が一番として固定され、残った２つの属性について順序をランダムにして匿名化データが生成される。符号Ｇ１に示す例では、「年収〉職業〉性別」及び「年収〉性別〉職業」を属性優先順序とする匿名化データが生成される。そして、符号Ｇ２に示すように、２つの匿名化データが結合されて学習モデルが出力される。

【００６２】

図１２は、実施形態の一例における一般化匿名化処理を利用した匿名化処理の生成処理を説明するフローチャートである。図１３は、図１２に示したオリジナルデータからの匿名化データの結合例を示す図である。図１２に示すフローチャート（ステップＳ２１～Ｓ２３）に従って、一般化匿名化処理を利用した匿名化処理の生成処理を説明する。

【００６３】

表形式のオリジナルデータが入力として受け付けられる（ステップＳ２１）。

【００６４】

属性に関するパラメータ（別言すれば、属性に対する重み）がランダムに生成される（ステップＳ２２）。

【００６５】

生成した各属性優先順序の設定の元でそれぞれの匿名化データが出力され、得られた匿名化データが結合される（ステップＳ２３）。そして、一般化匿名化処理を利用した匿名化処理の生成処理は終了する。

【００６６】

図１３に示す例では、属性に対する重みとして、(住所,職業,年収) = (0.1,0.6,0.3), (0.3,0.1,0.6)が生成されると、符号Ｈ１に示す匿名化データが生成される。そして、符号Ｈ２に示すように、２つの匿名化データが結合されて学習モデルが出力される。

【００６７】

〔Ａ－４〕関連例との比較

図１４は、関連例における学習モデルの生成処理を模式的に示す図である。

【００６８】

関連例においては、図１４の符号Ｉ１に示すオリジナルデータを入力として、符号Ｉ２に示す１つの匿名化データが生成される。そして、１つの匿名化データに対して機械学習が実施されることにより、符号Ｉ３に示すように、学習モデルが生成される。

【００６９】

図１５は、実施形態の一例における学習モデルの生成処理を模式的に示すブロック図で

ある。

【００７０】

一方、実施形態の一例においては、図１５の符号Ｊ１に示すオリジナルデータを入力として、符号Ｊ２に示す複数の匿名化データ＃１～＃ｎが生成される。複数の匿名化データ＃１～＃ｎが結合されて、符号Ｊ３に示すように、結合匿名化データが生成される。そして、結合匿名化データに対して機械学習が実施されることにより、符号Ｊ４に示すように、学習モデルが生成される。

【００７１】

図１６は、関連例における学習モデルと実施形態の一例における学習モデルとの精度を比較するテーブルである。

【００７２】

図１６に示す例では、単一の匿名化データで学習したモデルと、複数の匿名化データで学習したモデルとの精度が比較されている。

【００７３】

実験方法として、２から５０までのｋに対して、８個のｋ－匿名化データを生成する。次に、２から５０までのｋに対し、８個のｋ－匿名化データそれぞれで学習した８個の学習モデルによる精度のうち最高のものと、８個のｋ－匿名化データを結合したデータで学習した学習モデルの精度とを記録する。そして、それぞれの精度について、ｋに関する平均値、最小値をとった。

【００７４】

このような実験の結果、図１６に示すように、複数の匿名化データから複数のモデルを作るよりも、複数の匿名化データを結合して一つのモデルを作った方が、平均値及び最小値の両方の精度が高くなる。

【００７５】

〔Ａ－５〕匿名化データの生成処理の詳細

匿名化データの生成処理の具体例としては、データオーギュメンテーションとランダム生成とが想定される。

【００７６】

データオーギュメンテーションにおいては、安定性は高くなるものの、精度が低くなるおそれがある。例えば、よく似た匿名化データが結合されることにより、多様性が失われてしまい、学習モデルに汎用性がなくなる。特に、表形式のオリジナルデータの場合には、画像データとは異なり、似たようなデータが入力されても、機械学習の効果が薄くなる。また、匿名化データが少しずつ変更されるため、結合匿名化データが、匿名化アルゴリズムのハイパーパラメータの初期値に強く依存するおそれがある。

【００７７】

一方、ランダム生成においては、安定性が低くなるおそれがある。匿名化データがランダムに生成されてしまうので、学習に適した匿名化データが生成されないことがある。また、大量に生成して多様性を担保することは、学習コストの面で非効率になる。

【００７８】

図１７は、実施形態の一例におけるデータの削除箇所を例示する図である。

【００７９】

符号Ｋ１に示す様に、オリジナルデータから複数の匿名化データ＃１～＃ｍ間で似たような部分を削除するのは学習に適さないと想定される一方、符号Ｋ２に示す様に、オリジナルデータから複数の匿名化データ＃１～＃ｍ間で大きく異なる部分を削除するのが学習に適すると想定される。

【００８０】

すなわち、“大いに異なる”匿名化データを構成することができれば、データを補完し合えるので、学習に適したデータを生成できると想定される。

【００８１】

匿名化データの生成処理の第１の具体例として、匿名化データの水増し件数ｍ及び順序

同士の距離関数を入力とする。順序同士の距離関数は、Kendallのτ距離やCayley距離等の任意の関数でよい。また、属性数をｎとしたとき、ｍ≦ｎ！を満たすものとする。

【００８２】

与えられた距離関数によって、順序の間の距離の総和といった多様性を表す指標が最大となるｍ個の異なる属性順序が、全ての組み合わせを調べることによって決定される。

【００８３】

決定されたｍ個の異なる属性順序を使って、属性に関する設定がｍ個作成され、ｋ－匿名化データが生成・結合される。

【００８４】

図１８は、実施形態の一例におけるデータの削除箇所の決定アルゴリズムの第１の例を説明するテーブルである。

【００８５】

ここで、匿名化データの生成処理の第１の具体例において、水増し件数をm=2、属性数はn=3、距離関数をKendallのτ距離関数とする。なお、Kendallのτ距離は、二つの1からnの整数からなる配列a,bが与えられたとき、i〈jを満たす1からnの整数の組 (i,j) であって、aのi番目の数とaのj番目の数の間の大小関係が、bのそれと食い違っているものを数え上げた数である。

【００８６】

順序間の距離は、図１８に示すようになる。

【００８７】

多様性を表す指標として、距離が最大になる異なる２つの属性優先順序を、全ての組み合わせを調べ決定する。この場合の異なる２つの属性優先順序は、{(123),(132)}, {(123),(213)}, {(123),(231)}, {(123),(312)}, {(123),(321)}, {(132),(213)}, {(132),(231)}, {(132),(312)}, {(132),(321)}, {(213),(231)}, {(213),(312)}, {(213),(321)}, {(231),(312)}, {(231),(321)}, {(312),(321)}となる。

【００８８】

２つの属性優先順序の間の距離を、図１８を用いて計算すると、それぞれ1,1,2,2,3,2,1,3,2,3,1,2,2,1,1となる。

【００８９】

そして、最大となるものが選択される。複数ある場合はランダムにとることで、一つ選択される。本例では、{(123),(321)}, {(132),(312)},{(213),(231)}から一つが選択される。

【００９０】

匿名化データの生成処理の第２の具体例として、匿名化データの水増し件数mを入力とする。また、属性数をnとしたとき、m ≦ 2n を満たすものとする。

【００９１】

以下、匿名化データの生成処理の第２の具体例におけるアルゴリズムを説明する。

【００９２】

まず、初期値である属性優先順序に対して、順序を固定する属性が選ばれる。その後、動かす順序が決定され、その並びを [1,2,…,n] とおく。

【００９３】

次に、数列(0,1/2,1/3,2/3,1/4,2/4,3/4,…)にnを掛け、整数に切り下げ、前から見て重複して現れた番号は除くことでできる数列を (a1,…,an)とする。

【００９４】

次に、1からnの各iに対して、属性優先順序biを[ai+1,ai+2,…,n,1,…,ai]と置く。

【００９５】

次に、属性優先順序の列B=(b1,r(b1),b2,r(b2),…,bn,r(bn))が生成される。ここで、優先順序bに対して、r(b)はそれをひっくり返したものを表す。例えば、b=[3,4,1,2]ならばr(b) = [2,1,4,3]である。

【００９６】

そして、Bの先頭からm項をとり、それらの属性優先順序を用いて、k-匿名化データが生成・結合される。

【００９７】

匿名化データの生成処理の第２の具体例におけるアルゴリズムでは、計算量は非常に少なくなる。またアルゴリズムの2～3が作用して、データ間のKendallのτ距離の総和は、mが偶数なら最大で、奇数の場合もランダムに比べ大きくなると期待できる。

【００９８】

ここで、匿名化データの生成処理の第２の具体例において、水増し件数をm=3とする。属性は{年収、職業、住所、性別、学歴}の5種類で、さらに年収は必ず属性優先順序の最初に置くと決める。すなわち、動かす属性は{職業、住所、性別、学歴}の4つであるため、n=4の場合に対応する。

【００９９】

初期値である属性優先順序を固定し、その並びを[1,2,3,4]とおく。なお、実際には[職業、学歴、性別、住所]などと並ぶが、便宜上数値とする。

【０１００】

数列(0,1/2,1/3,2/3,1/4,2/4,3/4,…)に4を掛け、整数に切り下げ、前から見て重複して現れた番号は除くことでできる数列は(0,2,1,3) となる。

【０１０１】

1から4の各iに対して、属性優先順序biは、b1=[1,2,3,4], b2=[3,4,1,2], b3=[2,3,4,1], b4=[4,1,2,3]となる。

【０１０２】

属性優先順序の列B=(b1,r(b1),b2,r(b2),…,bn,r(bn))は、([1,2,3,4],[4,3,2,1],[3,4,1,2],[2,1,4,3],[2,3,4,1],[1,4,3,2],[4,1,2,3],[3,2,1,4])となる。

【０１０３】

Bの先頭からm=3項をとると、[1,2,3,4], [4,3,2,1], [3,4,1,2]が生成され、それらの属性優先順序を用いて、k-匿名化データが生成・結合される。

【０１０４】

本例では、結果として、[職業、学歴、性別、住所]、[住所、性別、学歴、職業]、[性別、住所、職業、学歴]の３つが生成される。

【０１０５】

図１９は、学習モデルの生成処理の違いによる学習モデルの精度を比較するテーブルである。

【０１０６】

図１９においては、データオーギュメンテーション及びランダム生成による匿名化データで学習した学習モデルと、匿名化データの生成処理の第２の具体例による匿名化データで学習した学習モデルとが比較されている。

【０１０７】

実験方法として、データオーギュメンテーション、ランダム生成Ａ、ランダム生成Ｂ、匿名化データの生成処理の第２の具体例により、2から15までのｋに対して、ｋ－匿名化データを８個作成する。なお、ランダム生成A,Bは、異なるシードによるランダム生成を意味する。また、目的変数は必ず優先順序を1位とした。初期値となる属性優先順序は、学習器の特徴量重要度を用いた。

【０１０８】

それぞれの場合で、データを結合し、学習したモデルの精度を各ｋについて比較し、ｋに関する精度の平均、最小値、最大値を記録すると、図１９に示すテーブルが得られた。

【０１０９】

図１９に示す実験結果において、匿名化データの生成処理の第２の具体例は、データオーギュメンテーションよりも精度が高くなる。また、ランダム生成は精度が良い場合もあるが、シードによって値がばらつくため安定しないことが確認される。匿名化データの生成処理の第２の具体例はシードに寄らず、平均値は最大となる。

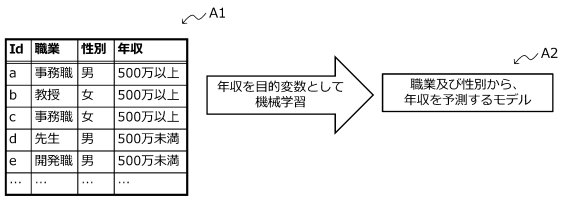
【０１１０】
〔Ａ－６〕効果
上述した実施形態の一例における生成方法，情報処理装置１及び生成プログラムによれば、例えば、以下の作用効果を奏することができる。
【０１１１】
受付処理部１１１は、複数の項目値を含む複数の個人情報を受け付ける。匿名情報生成部１１２は、複数の項目値それぞれに対応付けられた第１のパラメータに応じて複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた複数の個人情報に適用して、複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成する。また、匿名情報生成部１１２は、複数の項目値それぞれに対応付けられた第１のパラメータとは異なる第２のパラメータに応じて複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた複数の個人情報に適用して、複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成する。そして、結合情報生成部１１３は、第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する。
【０１１２】
これにより、個人情報の匿名性を確保しつつ、機械学習によって生成する学習モデルの精度を向上できる。
【０１１３】
第１及び第２のパラメータについての多様性に関する指標が最大化されるように、当該第１及び第２のパラメータを決定する。これにより、匿名情報の多様性を向上させることができる。
【０１１４】
第１及び第２のパラメータは、例えば、前記複数の項目値間の優先順序である。第１及び第２のパラメータは、例えば、前記複数の項目値それぞれの重み付け値であってもよい。これにより、多様性を有する匿名情報を容易に生成することができる。
【０１１５】
〔Ｂ〕その他
開示の技術は上述した実施形態に限定されるものではなく、本実施形態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。本実施形態の各構成及び各処理は、必要に応じて取捨選択することができ、あるいは適宜組み合わせてもよい。
【０１１６】
〔Ｃ〕付記
以上の実施形態に関し、更に以下の付記を開示する。
【０１１７】
（付記１）
複数の項目値を含む複数の個人情報を受け付け、
前記複数の項目値それぞれに対応付けられた第１のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成し、
前記複数の項目値それぞれに対応付けられた前記第１のパラメータとは異なる第２のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成し、
前記第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する、
処理をコンピュータが実行することを特徴とする生成方法。
【０１１８】
（付記２）
前記第１及び第２のパラメータについての多様性に関する指標が最大化されるように、

当該第１及び第２のパラメータを決定する、
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に記載の生成方法。
【０１１９】
（付記３）
前記第１及び第２のパラメータは、前記複数の項目値間の優先順序である、
ことを特徴とする、付記１又は２に記載の生成方法。
【０１２０】
（付記４）
前記第１及び第２のパラメータは、前記複数の項目値それぞれの重み付け値である、
ことを特徴とする、付記１又は２に記載の生成方法。
【０１２１】
（付記５）
複数の項目値を含む複数の個人情報を受け付ける受付処理部と、
前記複数の項目値それぞれに対応付けられた第１のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成する第１匿名情報生成部と、
前記複数の項目値それぞれに対応付けられた前記第１のパラメータとは異なる第２のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成する第２匿名情報生成部と、
前記第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する結合情報生成部と、
を備えることを特徴とする情報処理装置。
【０１２２】
（付記６）
前記第１及び第２のパラメータについての多様性に関する指標が最大化されるように、
当該第１及び第２のパラメータを決定する、
ことを特徴とする、付記５に記載の情報処理装置。
【０１２３】
（付記７）
前記第１及び第２のパラメータは、前記複数の項目値間の優先順序である、
ことを特徴とする、付記５又は６に記載の情報処理装置。
【０１２４】
（付記８）
前記第１及び第２のパラメータは、前記複数の項目値それぞれの重み付け値である、
ことを特徴とする、付記５又は６に記載の情報処理装置。
【０１２５】
（付記９）
複数の項目値を含む複数の個人情報を受け付け、
前記複数の項目値それぞれに対応付けられた第１のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第１の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第１の匿名情報を生成し、
前記複数の項目値それぞれに対応付けられた前記第１のパラメータとは異なる第２のパラメータに応じて前記複数の項目値それぞれの秘匿化されやすさが決まる第２の秘匿化アルゴリズムを、受け付けた前記複数の個人情報に適用して、前記複数の項目値の少なくとも何れかの項目値が匿名化された第２の匿名情報を生成し、
前記第１の匿名情報及び前記第２の匿名情報を含む匿名情報を生成する、
処理をコンピュータに実行させることを特徴とする生成プログラム。

【符号の説明】
【０１２６】
１　　：情報処理装置
１１　　：ＣＰＵ
１１１　　：受付処理部
１１２　　：匿名情報生成部
１１３　　：結合情報生成部
１２　　：メモリ部
１３　　：表示制御部
１４　　：記憶装置
１５　　：入力ＩＦ
１５１　　：マウス
１５２　　：キーボード
１６　　：外部記録媒体処理部
１６０　　：記録媒体
１３０　　：表示装置
１７　　：通信ＩＦ

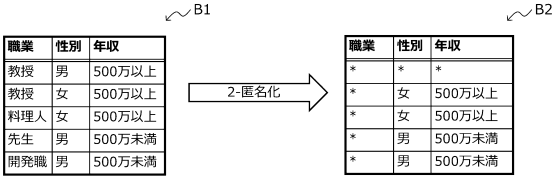
【図面】

【図１】




| Id | 職業 | 性別 | 年収 |
|---|---|---|---|
| a | 事務職 | 男 | 500万以上 |
| b | 教授 | 女 | 500万以上 |
| c | 事務職 | 女 | 500万以上 |
| d | 先生 | 男 | 500万未満 |
| e | 開発職 | 男 | 500万未満 |
| … | … | … | … |

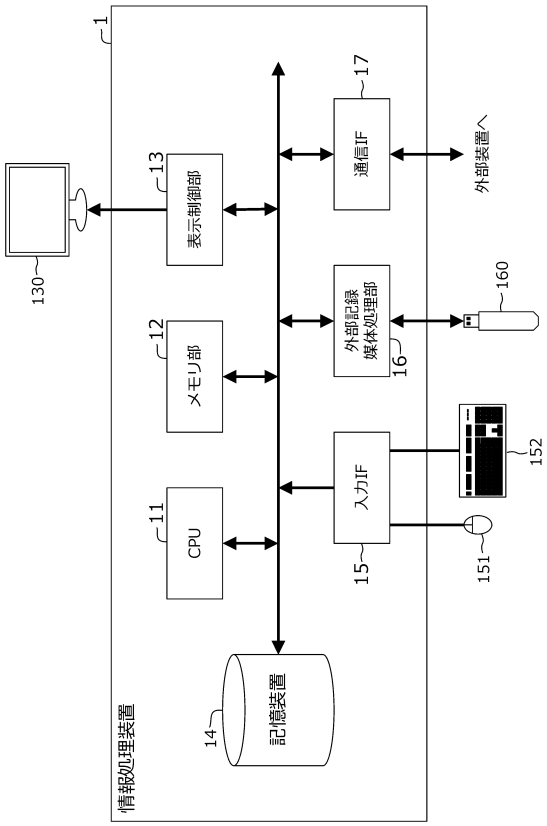
【図２】



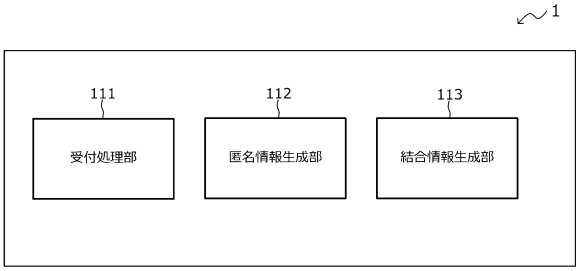

| 職業 | 性別 | 年収 |
|---|---|---|
| 教授 | 男 | 500万以上 |
| 教授 | 女 | 500万以上 |
| 料理人 | 女 | 500万以上 |
| 先生 | 男 | 500万未満 |
| 開発職 | 男 | 500万未満 |

| 職業 | 性別 | 年収 |
|---|---|---|
| * | * | * |
| * | 女 | 500万以上 |
| * | 女 | 500万以上 |
| * | 男 | 500万未満 |
| * | 男 | 500万未満 |

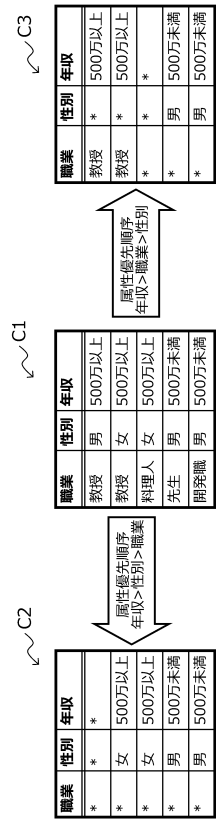
【図３】




【図４】



【図５】




C1

| 職業 | 性別 | 年収 |
|---|---|---|
| 教授 | 男 | 500万以上 |
| 教授 | 女 | 500万以上 |
| 料理人 | 女 | 500万以上 |
| 先生 | 男 | 500万未満 |
| 開発職 | 男 | 500万未満 |

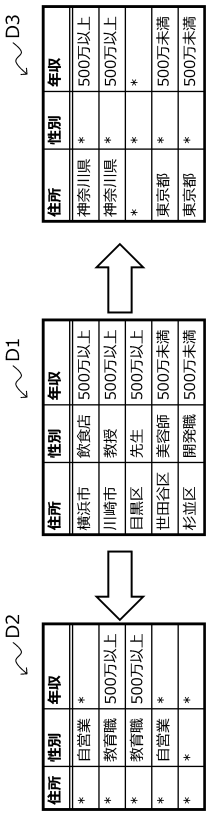
C2

| 職業 | 性別 | 年収 |
|---|---|---|
| * | * | * |
| * | 女 | 500万以上 |
| * | 女 | 500万以上 |
| * | 男 | 500万未満 |
| * | 男 | 500万未満 |

C3

| 職業 | 性別 | 年収 |
|---|---|---|
| 教授 | * | 500万以上 |
| 教授 | * | 500万以上 |
| * | * | * |
| * | 男 | 500万未満 |
| * | 男 | 500万未満 |

【図６】



D1

| 住所 | 性別 | 年収 |
|---|---|---|
| 横浜市 | 飲食店 | 500万以上 |
| 川崎市 | 教授 | 500万以上 |
| 目黒区 | 先生 | 500万以上 |
| 世田谷区 | 美容師 | 500万未満 |
| 杉並区 | 開発職 | 500万未満 |

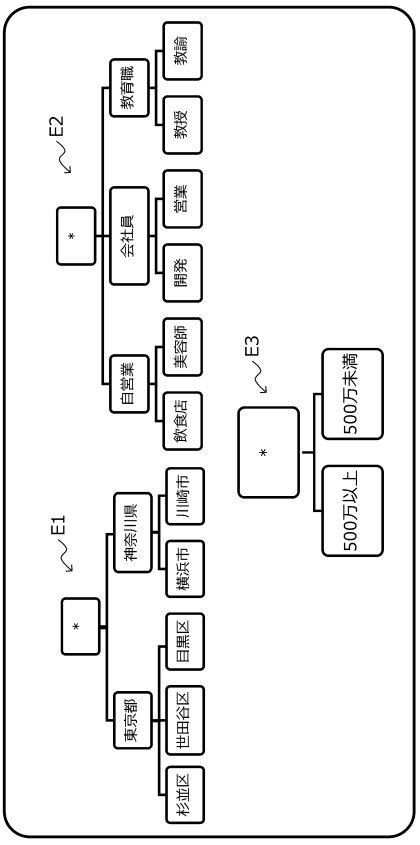
D2

| 住所 | 性別 | 年収 |
|---|---|---|
| * | 自営業 | * |
| * | 教育職 | 500万以上 |
| * | 教育職 | 500万以上 |
| * | 自営業 | * |
| * | * | * |

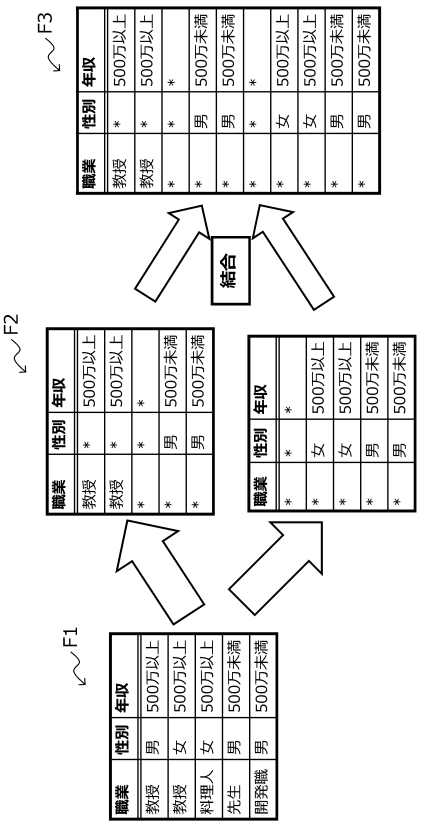
D3

| 住所 | 性別 | 年収 |
|---|---|---|
| 神奈川県 | * | 500万以上 |
| 神奈川県 | * | 500万以上 |
| * | * | * |
| 東京都 | * | 500万未満 |
| 東京都 | * | 500万未満 |

【図７】




【図８】




F1

| 職業 | 性別 | 年収 |
|---|---|---|
| 教授 | 男 | 500万以上 |
| 教授 | 女 | 500万以上 |
| 料理人 | 女 | 500万以上 |
| 先生 | 男 | 500万未満 |
| 開発職 | 男 | 500万未満 |

F2

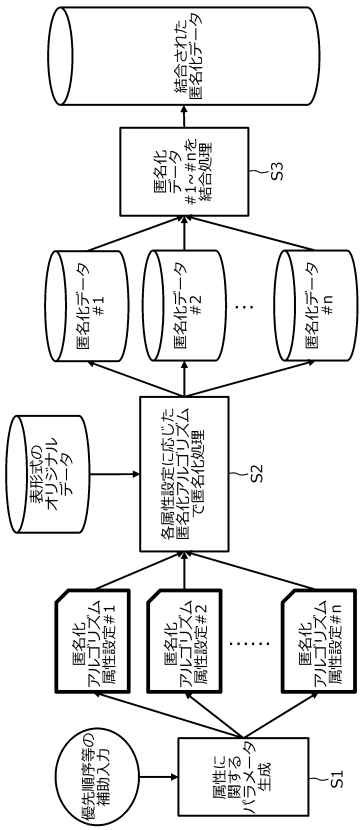
| 職業 | 性別 | 年収 |
|---|---|---|
| 教授 | * | 500万以上 |
| 教授 | * | 500万以上 |
| * | * | * |
| * | 男 | 500万未満 |
| * | 男 | 500万未満 |

| 職業 | 性別 | 年収 |
|---|---|---|
| * | * | * |
| * | 女 | 500万以上 |
| * | 女 | 500万以上 |
| * | 男 | 500万未満 |
| * | 男 | 500万未満 |

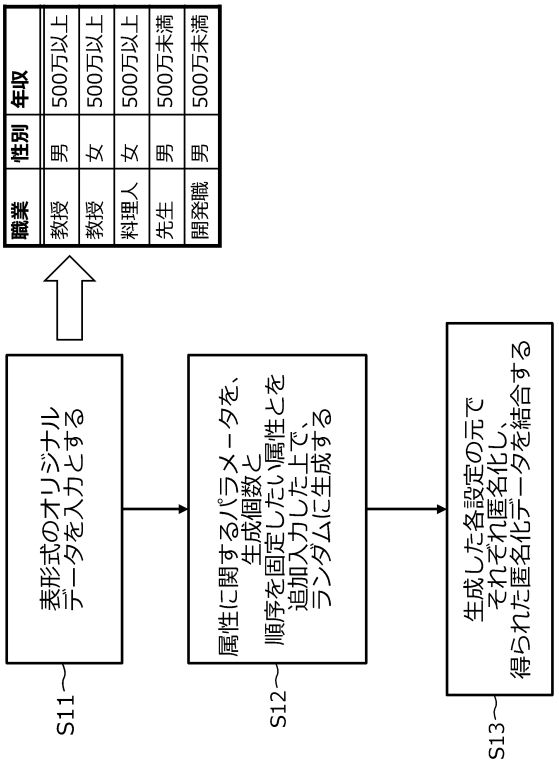
F3

| 職業 | 性別 | 年収 |
|---|---|---|
| 教授 | * | 500万以上 |
| 教授 | * | 500万以上 |
| * | * | * |
| * | 男 | 500万未満 |
| * | 男 | 500万未満 |
| * | * | * |
| * | 女 | 500万以上 |
| * | 女 | 500万以上 |
| * | 男 | 500万未満 |
| * | 男 | 500万未満 |

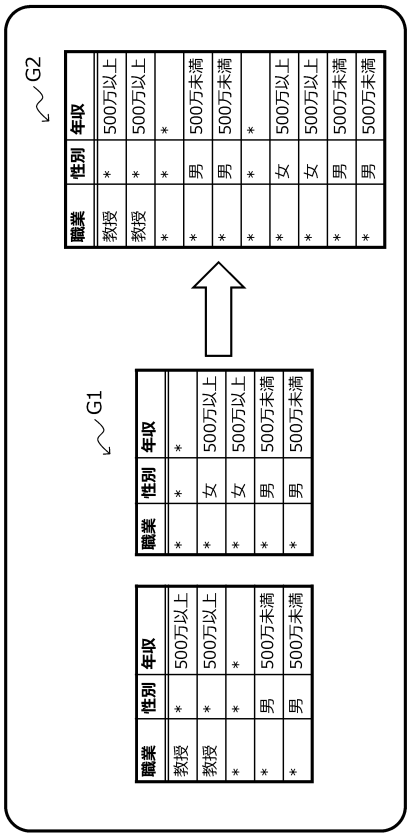
【図９】



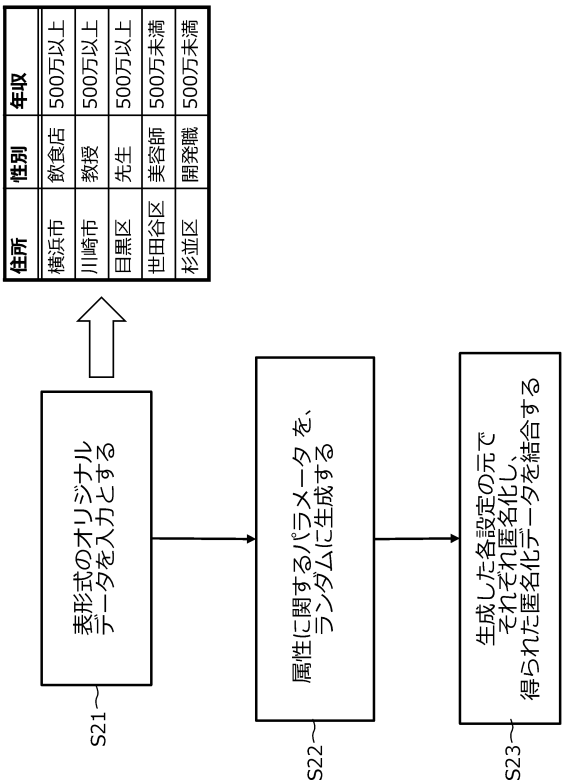

【図１０】



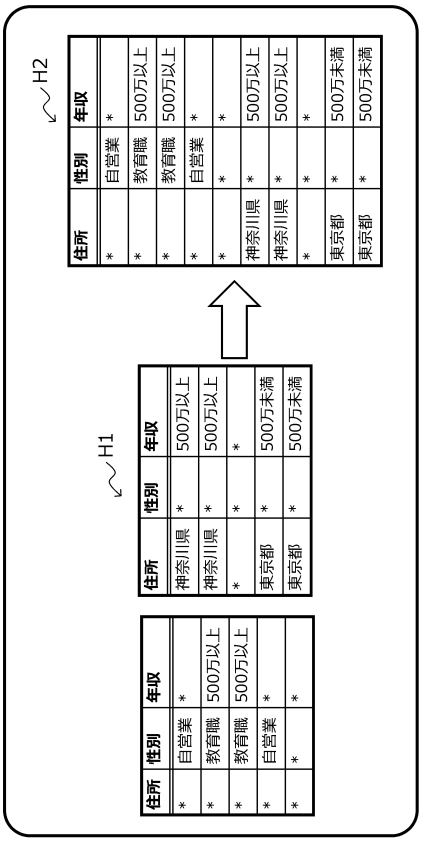

【図１１】



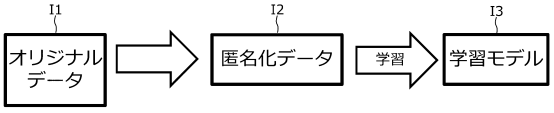

【図１２】



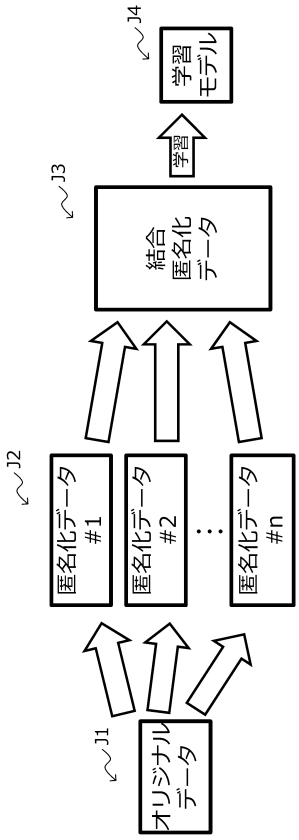
【図１３】




【図１４】



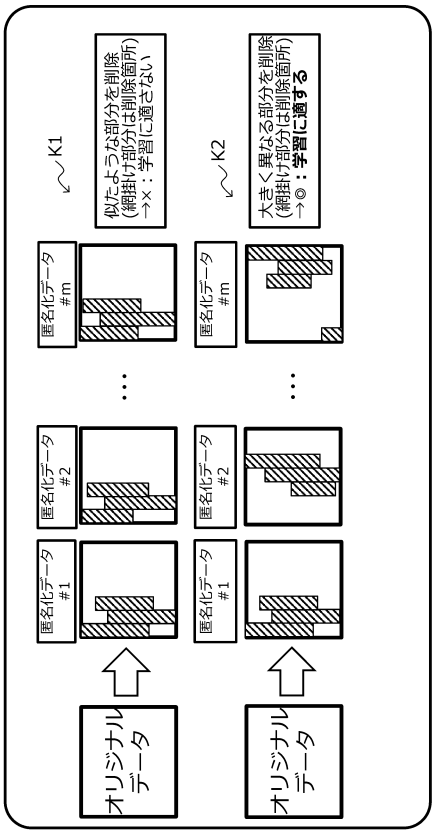

【図１５】




【図１６】

| 精度のkに関する要約統計量 | 平均 | 最小値 |
|---|---|---|
| それぞれ学習した場合の最大精度 | 0.8270 | 0.8236 |
| 結合したデータで学習した場合の精度 | 0.8275 | 0.8248 |

【図１７】




【図１８】

| 順序間の距離 | (123) | (132) | (213) | (231) | (312) | (321) |
|---|---|---|---|---|---|---|
| (123) | 0 | 1 | 1 | 2 | 2 | 3 |
| (132) | 1 | 0 | 2 | 1 | 3 | 2 |
| (213) | 1 | 2 | 0 | 3 | 1 | 2 |
| (231) | 2 | 1 | 3 | 0 | 2 | 1 |
| (312) | 2 | 3 | 1 | 2 | 0 | 1 |
| (321) | 3 | 2 | 2 | 1 | 1 | 0 |

【図１９】

| 精度のkに関する要約統計量 | 平均 | 最小値 | 最大値 |
|---|---|---|---|
| データオーギュメンテーション | 0. 8271 | 0. 8256 | 0. 8278 |
| ランダム生成A | 0. 8274 | 0. 8262 | 0. 8288 |
| ランダム生成B | 0. 8277 | 0. 8262 | 0. 8293 |
| 赤：補助手段その2の実施例2 | 0. 8281 | 0. 8265 | 0. 8293 |

フロントページの続き

(56)参考文献　特表２０１９－５２６８５１（ＪＰ，Ａ）
特開２０１４－２２９０３９（ＪＰ，Ａ）
国際公開第２０１２／０６７２１３（WO，Ａ１）
特開２０１１－２０９８００（ＪＰ，Ａ）
米国特許出願公開第２０１８／０００４９７８（ＵＳ，Ａ１）
米国特許出願公開第２０１６／０１３２６９７（ＵＳ，Ａ１）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　２１／６２