

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7367872号
(P7367872)

(45)発行日 令和5年10月24日(2023.10.24)

(24)登録日 令和5年10月16日(2023.10.16)

(51)国際特許分類

F I

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

請求項の数 11 (全17頁)

(21)出願番号	特願2022-532227(P2022-532227)	(73)特許権者	000005223
(86)(22)出願日	令和2年6月26日(2020.6.26)		富士通株式会社
(86)国際出願番号	PCT/JP2020/025360		神奈川県川崎市中原区上小田中4丁目1
(87)国際公開番号	WO2021/260945		番1号
(87)国際公開日	令和3年12月30日(2021.12.30)	(74)代理人	110001519
審査請求日	令和4年9月29日(2022.9.29)		弁理士法人太陽国際特許事務所
		(72)発明者	中尾 悠里
			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	真木 健彦

最終頁に続く

(54)【発明の名称】 訓練データ生成プログラム、装置、及び方法

(57)【特許請求の範囲】

【請求項1】

複数のデータに含まれる属性毎のデータ数に基づいて算出された値に対して、評価値を受け付け、

受け付けた前記評価値と前記属性毎のデータ数とに基づいて、属性毎の基準値を決定し、前記属性毎の基準値に応じて、前記複数のデータの少なくとも一部のデータの属性を変更することによって機械学習用の訓練データを生成する、
処理をコンピュータに実行させることを特徴とする訓練データ生成プログラム。

【請求項2】

複数の評価者の各々から前記評価値を受け付け、
受け付けた前記評価値の散らばり度合いが所定値以下の場合に、前記複数の評価者の各々から受け付けた前記評価値を統合し、前記複数の評価者による合意済みの評価値として受け付ける、
ことを特徴とする請求項1に記載の訓練データ生成プログラム。

【請求項3】

前記散らばり度合いが前記所定値を超える場合、前記複数の評価者の各々から受け付けた前記評価値を、それぞれの前記散らばり度合いが前記所定値以下となるまでクラスタリングし、各クラスタに含まれる評価値を統合し、前記合意済みの評価値の各々として受け付ける、
ことを特徴とする請求項2に記載の訓練データ生成プログラム。

【請求項 4】

前記合意済みの評価値が複数存在する属性については、前記合意済みの評価値の各々を選択肢として提示することにより、最終的な評価値を受け付ける、
ことを特徴とする請求項 3 に記載の訓練データ生成プログラム。

【請求項 5】

前記属性毎のデータ数に基づいて算出された値を、前記評価値の大きさに応じた割合で引き下げた値を、前記属性毎の基準値として決定する、
ことを特徴とする請求項 1～請求項 4 のいずれか 1 項に記載の訓練データ生成プログラム。

【請求項 6】

前記属性毎の基準値と前記値との差が所定値以下となるように、前記複数のデータの少なくとも一部のデータの属性を変更する、
ことを特徴とする請求項 1～請求項 5 のいずれか 1 項に記載の訓練データ生成プログラム。

【請求項 7】

前記属性は、判定に用いられる属性と、判定結果を表す属性とを含み、前記判定に用いられる属性の前記判定結果に対する寄与度が前記値として算出されている、
ことを特徴とする請求項 1～請求項 6 のいずれか 1 項に記載の訓練データ生成プログラム。

【請求項 8】

前記寄与度が前記属性毎の基準値以下となるように、前記複数のデータの少なくとも一部のデータについて、前記判定結果を表す属性の属性値を、異なる判定結果を示す属性値に変更する、
ことを特徴とする請求項 7 に記載の訓練データ生成プログラム。

【請求項 9】

前記評価値として、前記判定に用いられる属性が前記判定結果に差別的に寄与している度合いを示す差別度を受け付ける、
ことを特徴とする請求項 7 又は請求項 8 に記載の訓練データ生成プログラム。

【請求項 10】

複数のデータに含まれる属性毎のデータ数に基づいて算出された値に対して、評価値を受け付け、
受け付けた前記評価値と前記属性毎のデータ数とに基づいて、属性毎の基準値を決定し、前記属性毎の基準値に応じて、前記複数のデータの少なくとも一部のデータの属性を変更することによって機械学習用の訓練データを生成する、
処理を実行する制御部を含むことを特徴とする訓練データ生成装置。

【請求項 11】

複数のデータに含まれる属性毎のデータ数に基づいて算出された値に対して、評価値を受け付け、
受け付けた前記評価値と前記属性毎のデータ数とに基づいて、属性毎の基準値を決定し、前記属性毎の基準値に応じて、前記複数のデータの少なくとも一部のデータの属性を変更することによって機械学習用の訓練データを生成する、
処理をコンピュータが実行することを特徴とする訓練データ生成方法。

【発明の詳細な説明】

【技術分野】

【0001】

開示の技術は、訓練データ生成プログラム、訓練データ生成装置、及び訓練データ生成方法に関する。

【背景技術】

【0002】

例えば、ローン申請の可否、採用試験の可否等の判定を、機械学習されたモデルを用い

10

20

30

40

50

で行う場合がある。このような判定に用いるモデルは、公平性配慮型機械学習により生成するなどして、性別や人種といったセンシティブ属性によって判定結果に差別が生じないようにする必要がある。また、性別や人種等、判定の際に法的に差別的な取り扱いをしないように規定されている属性（以下、直接差別属性という）以外にも、住所や職業等、一見差別的には見えない属性が、直接差別属性の持つ傾向と一致し、差別的な結果を助長する場合がある。以下では、このような、一見差別的に見えない属性でありながら、差別的な判定結果を引き起こす属性を間接差別属性という。また、直接差別属性及び間接差別属性を総称して「差別属性」という。

【0003】

従来、データに含まれる属性と判定結果との相関を表す指標として、属性毎のデータ数を用いて算出したExtended lift (e l i f t) という値を用いて、差別属性を検出する技術が提案されている。この技術では、e l i f t が所定の閾値以上となる属性値を持つ属性を、その属性が結果に対して有効に作用していると判定し、差別属性として検出する。

【0004】

また、差別属性の属性値によって、差別的な判定結果が生じないように、モデルを機械学習するための訓練データのラベルを修正する技術が存在する。

【先行技術文献】

【非特許文献】

【0005】

【文献】Dino Pedreschi, Salvatore Ruggieri, Franco Turini, "Discrimination-aware Data Mining", KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, August 2008, Pages 560-568

【発明の概要】

【発明が解決しようとする課題】

【0006】

公平性配慮型機械学習により生成されたモデルを用いた判定を、公平性を要するビジネス等に適用する場合、何が差別属性とみなされるかは、文化、法規制、ビジネス領域、商慣習等の文脈に依存する。しかしながら、従来技術のように、データに含まれる各属性に対して一律の閾値が設定される場合には、差別属性の誤検出又は検出漏れが生じる可能性がある、という問題がある。この場合、差別的な判定へのバイアスを適切に低減できるように訓練データのラベルを修正することができない。そのため、そのような訓練データを用いて機械学習されたモデルは、公平性を配慮した適切な判定結果を出力することができない、という問題がある。

【0007】

一つの側面として、開示の技術は、文脈を考慮した差別属性による差別的な判定へのバイアスが適切に低減されるモデルを機械学習するための訓練データを生成することを目的とする。

【課題を解決するための手段】

【0008】

一つの態様として、開示の技術は、複数のデータに含まれる属性毎のデータ数に基づいて算出された値に対して、評価値を受け付け、受け付けた前記評価値と前記属性毎のデータ数とに基づいて、属性毎の基準値を決定する。また、開示の技術は、前記属性毎の基準値に応じて、前記複数のデータの少なくとも一部のデータの属性を変更することによって機械学習用の訓練データを生成する。

【発明の効果】

【0009】

一つの側面として、文脈を考慮した差別属性による差別的な判定へのバイアスが適切に低減されるモデルを機械学習するための訓練データを生成することができる、という効果を有する。

10

20

30

40

50

【図面の簡単な説明】

【0010】

【図1】訓練データ生成装置の機能ブロック図である。

【図2】一定の基準値を設定する場合を説明するための図である。

【図3】属性毎に寄与度の基準値を決定する場合を説明するための図である。

【図4】受付画面の一例を示す図である。

【図5】合意済み差別度の設定を説明するための図である。

【図6】合意済み差別度の設定を説明するための図である。

【図7】多様な差別度モードを説明するための図である。

【図8】各属性について設定された合意済み差別度の一例を示す図である。

10

【図9】選択画面の一例を示す図である。

【図10】追加受付画面の一例を示す図である。

【図11】第2の評価者に選択された差別度の集計結果の一例を示す図である。

【図12】最終的な差別度に応じた、属性毎の寄与度の基準値の決定を説明するための図である。

【図13】各属性の寄与度の基準値の一例を示す図である。

【図14】寄与度の基準値に基づくデータの変更を説明するための図である。

【図15】訓練データ生成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。

20

【図16】訓練データ生成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】

【0011】

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。

【0012】

図1に示すように、本実施形態に係る訓練データ生成装置10には、データ集合が入力される。データ集合に含まれる各データは、公平性配慮型機械学習によりモデルを生成するための訓練データの元となるデータであり、複数の属性の各々についての属性値を含む。例えば、生成されるモデルが、ローン審査の承認又は却下を判定するためのモデルの場合、属性は、ユーザ毎の性別、人種、収入、職業、ローンの目的（以下、単に「目的」ともいう）、返済期間、住所、年齢、判定結果等である。属性値は、属性の種類に応じた値であり、例えば、属性「性別」については、「男性」又は「女性」を示す値であり、属性「判定結果」については、「承認」又は「却下」を示す値である。各データは、各属性の属性値を連結したベクトル値とすることができる。

30

【0013】

属性「判定結果」は、開示の技術の「判定結果を表す属性」の一例である。以下では、属性「判定結果」を「ラベル」ともいう。また、属性「判定結果」以外の属性は、開示の技術の「判定に用いられる属性」の一例である。訓練データ生成装置10は、データに含まれる属性のうち、直接差別属性及び属性「判定結果」以外の属性から、間接差別属性を特定する。例えば、上記のローン審査の例で、性別及び人種が直接差別属性として規定されている場合、訓練データ生成装置10は、収入、職業、ローンの目的、返済期間、住所、及び年齢から間接差別属性を特定する。

40

【0014】

ここで、図2に示すように、コンピュータが、各属性について、その属性の判定結果に対する寄与度を算出し、算出した寄与度が所定の基準値を超える属性を間接差別属性として特定することが考えられる。寄与度は、従来技術の`elift`のような、属性毎のデータ数に基づいて算出される、その属性と判定結果との相関を示す値である。

【0015】

しかし、上述したように、公平性配慮型機械学習により生成されたモデルを用いた判定を、公平性を要するビジネス等に適用する場合、何が差別属性とみなされるかは文脈に依存する。コンピュータが、文脈依存性を考慮して間接差別属性を特定するためには、属性

50

毎の「寄与効果」を考慮する必要がある。寄与効果とは、人間にとって、間接差別属性の判断の根拠となる寄与度の基準値は属性毎に異なるように観測されるという効果である。例えば、属性の中には、判定結果に少しでも寄与すれば差別的であるとみなされる属性もあれば、判定結果に大きく寄与した場合にのみ差別的であるとみなされる属性もある。図2に示すように、コンピュータが、いずれの属性に対しても一律の基準値を決定し、決定した基準値を寄与度が超えたか否かによって間接差別属性を特定する場合には、この属性毎の寄与効果を考慮した特定を行うことができない。

【0016】

本実施形態では、図3に示すように、訓練データ生成装置10は、属性毎に寄与度の基準値を決定する。具体的には、訓練データ生成装置10は、各属性の寄与度に対して、その属性がどの程度差別的であるかに応じて、属性毎の寄与度の基準値を決定する。以下、本実施形態に係る訓練データ生成装置10の機能構成について詳述する。

【0017】

図1に示すように、訓練データ生成装置10は、機能的には、算出部12と、受付部14と、決定部16と、生成部18とを含む。

【0018】

算出部12は、訓練データ生成装置10に入力されたデータ集合から、属性毎の寄与度を算出する。上述したように、寄与度は、属性毎のデータ数に基づいて算出される値であり、その属性と判定結果との相関を示す値である。例えば、算出部12は、ある属性についてXという属性値を持つデータのうち、他の属性についてYという属性値も持つデータの割合で表される信頼度 $\text{conf}(X \rightarrow Y)$ や、従来技術の elift 等を用いて寄与度を算出してよい。なお、寄与度は、開示の技術の「複数のデータに含まれる属性毎のデータ数に基づいて算出された値」の一例である。

【0019】

一例として、属性「目的」の属性値として、「中古車の購入」及び「休暇中の予算」があるとする。この場合において、算出部12は、属性「判定結果」の属性値が「承認」であるデータ数に対する、属性「判定結果」の属性値が「承認」、かつ属性「目的」の属性値が「中古車の購入」であるデータ数の割合を、「中古車の購入」についての信頼度として算出する。また、算出部12は、同様に、「休暇中の予算」についての信頼度を算出する。そして、算出部12は、「中古車の購入」の信頼度と、「休暇中の予算」の信頼度との差又は比を、属性「目的」についての判定結果に対する寄与度として算出する。これは、ある属性について、属性値毎の信頼度の差又は比が大きい場合には、その属性の属性値による判定結果への偏りが大きいことを表している。

【0020】

また、 elift は、“ $\text{elift} = \text{conf}(X, Z \rightarrow Y) / \text{conf}(X \rightarrow Y)$ ”の式で表される。一例として、算出部12は、Xを属性「住所」の属性値「ニューヨーク」、Zを属性「人種」の属性値「アフリカ系アメリカ人」、Yを属性「判定結果」の属性値「却下」とし、該当のデータ数を用いて上式から elift を算出する。この elift は、「ニューヨークに住む人でローン審査が却下された人の割合」に対して、「ニューヨークに住むアフリカ系アメリカ人でローン審査が却下された人の割合」の比を表す。また、同様に、算出部12は、Xを属性「住所」の属性値「ロスアンゼルス」、Y及びZは上記と同様として、 elift を算出する。算出部12は、この2つの elift の差又は比を、属性「住所」の寄与度として算出する。この場合の寄与度は、「人種」という直接差別属性との相関により、属性「住所」が間接差別属性となるか否かを判定するための指標となる。

【0021】

なお、属性が取り得る属性値が3種類以上の場合には、各属性値について算出された elift や信頼度の最大値と最小値との差又は比を寄与度として算出してもよい。また、寄与度の算出方法は上記の例に限定されず、属性と判定結果との相関を、属性毎のデータ数に基づいて算出するものであればよい。

【0022】

受付部14は、算出部12により算出された属性毎の寄与度に対して、評価値を受け付ける。受付部14は、評価値として、判定に用いられる属性毎の寄与度に基づいて判定された、判定に用いられる属性が判定結果に差別的に寄与している度合いを示す差別度を受け付ける。言い換えると、差別度は、各属性が差別的でなくなるために、寄与度を小さくする割合とすることができる。

【0023】

具体的には、受付部14は、第1の評価者集団に含まれる複数の第1の評価者の各々が利用する情報処理装置の表示装置に、例えば図4に示すような受付画面30を表示し、複数の第1の評価者の各々から、属性毎の差別度を受け付ける。図4の例では、受付画面30には、属性毎に寄与度が対応付けて表示されると共に、属性毎の差別度を選択するためのスライドバーと、スライドバーにより選択された差別度とが表示される。図4の例では、差別度を、0、1、2、・・・、10の離散値としているが、差別度はこの例に限定されない。

【0024】

ここで、差別度の決定方法として容易に考えられる方法としては、全ての評価者の各々に対して、各属性をランダムな順番で表示し、全ての評価者の各々から受け付けた差別度を集計する方法が挙げられる。しかし、評価が難しい属性の差別度の評価の際には、評価者は熟考プロセスに入るが、この方法では、その「熟考プロセス」が考慮されない。熟考プロセスでは、評価者は、1つの属性に関するシナリオを考え付く限り考慮し、差別度を選択する。例えば、評価者は、属性「住所」を考慮することは、通常では差別的ではないが、特定の人種が特定の地域に住んでいる場合には差別的に扱われるかもしれない、等の複数のシナリオを検討する。熟考プロセスに入った場合、大きな認知コストがかかる。人間の認知特性として、評価者は、認知コストが大きい意思決定を続ける場合、より安直な評価を行う傾向がある。すなわち、評価者は、熟考プロセスに入った属性について、非差別的である、と安直に評価してしまう場合がある。したがって、この方法では、各属性に対して与えられた評価への信頼性が確保できない場合がある。そこで、本実施形態に係る訓練データ生成装置10は、評価者が差別度を評価する際の認知負荷を軽減し、評価が難しい属性に対しても、評価者が素早く、かつ適切に差別的又は非差別的の評価を行うことができるように支援する。

【0025】

具体的には、受付部14は、ある属性に対して第1の評価者集団による差別度の合意があるか否かを判定する。第1の評価者集団内で差別度についての合意がある属性は、各第1の評価者からの差別度の評価傾向が一致する。すなわち、差別度の評価が集中する。一方、第1の評価者集団内で合意がない属性は、各第1の評価者からの差別度の評価が散らばる。そこで、受付部14は、第1の評価者集団内での合意の有無を、その属性に対する差別度の散らばり度合いに基づいて判定する。より具体的には、受付部14は、上述の受付画面30を介して第1の評価者集団から受け付けた差別度の散らばり度合いが所定値以下の場合に、第1の評価者の各々から受け付けた差別度を統合し、第1の評価者集団による合意済み差別度として受け付ける。

【0026】

例えば、受付部14は、差別度の散らばり度合いの一例として、各属性について評価された差別度の分散を計算する。図5に、30人の第1の評価者集団から受け付けた属性「年齢」についての差別度の集計結果を示す。図5の例では、属性「年齢」の差別度の分散は0.45と計算される。例えば、所定値を1とした場合、受付部14は、属性「年齢」を、第1の評価者集団内で単一の合意された差別度が存在する属性（以下、「合意済み属性」という）であると判定し、例えば、差別度の平均6.57を合意済み差別度として設定する。なお、散らばり度合いは分散に限らず、標準偏差等であってもよい。また、合意済み差別度は平均に限らず、最頻値、中央値等であってもよい。

【0027】

一方、受付部 14 は、散らばり度合いが所定値を超える場合、第 1 の評価者の各々から受け付けた差別度を、それぞれの散らばり度合いが所定値以下となるまでクラスタリングし、各クラスタに含まれる差別度を統合し、合意済み差別度の各々として受け付ける。図 6 に、30 人の第 1 の評価者集団から受け付けた属性「目的」についての差別度の集計結果を示す。図 6 の例では、属性「目的」の差別度の分散は 1.45 と計算される。上記の例と同様に、所定値を 1 とした場合、受付部 14 は、属性「目的」は、第 1 の評価者集団内で単一の差別度の合意がない属性（以下、「非合意属性」という）であると判定する。

【0028】

この場合、受付部 14 は、非合意属性に対して、多様な差別度モードを適用する。多様な差別度モードを適用することは、その属性に対して、複数の合意済み差別度を設定することである。第 1 の評価者集団内で単一の差別度への合意がない場合であっても、属性によっては、合意が全くないのではなく、差別度の見解が一致しておらず、第 1 の評価者集団の中に多様な差別度への合意がある可能性がある。例えば、図 7 に示すように、属性「住所」に対して、差別度が高いと認識している評価者と、差別度が低いと認識している評価者が存在する場合、多様な差別度モードが適用されることになる。これは、多様な差別度モードが適用される属性は、熟考プロセスを引き起こし易い属性である可能性が高いため、1 つの属性に対して複数の合意済み差別度を設定することで、評価者集団内での多様な公平性基準を表現するものである。

【0029】

例えば、受付部 14 は、直前まで、合意済み差別度の有無を判定していたクラスタ数に 1 を追加したクラスタ数になるように、評価者集団により評価された差別度をクラスタリングする。図 6 の例では、直前のクラスタ数は 1 であるため、受付部 14 は、クラスタ数を 2 とし、例えば、k-means アルゴリズム等により差別度をクラスタリングする。そして、受付部 14 は、クラスタ毎に、クラスタ内の差別度の分散を計算する。図 6 の例では、クラスタ 1（図 6 中の破線部分）については、分散は 0.14 と計算され、クラスタ 2（図 6 中の一点鎖線部分）については、分散は 0.44 と計算される。上記と同様に所定値を 1 とすると、いずれのクラスタも分散が所定値以下となるため、受付部 14 は、各クラスタに含まれる差別度の平均の各々を、その属性についての合意済み差別度として設定する。図 6 の例では、受付部 14 は、属性「目的」について、合意済み差別度として、2.83 及び 5 の 2 つを設定する。

【0030】

そして、受付部 14 は、1 つの合意済み差別度が設定されている属性については、その合意済み差別度を最終的な差別度として受け付ける。また、受付部 14 は、合意済み差別度が複数設定されている属性については、第 2 の評価者集団に含まれる第 2 の評価者の各々に、合意済み差別度の各々を選択肢として提示する。そして、受付部 14 は、第 2 の評価者の各々に選択肢を選択させ、選択結果を集計して、最終的な評価値として受け付ける。これは、上述した多様な差別度モードにより表現された、第 1 の評価者集団内での多様な公平性基準に基づいて、後述する「最終的な差別度」の決定について、第 2 の評価者集団による早期の合意を促すものである。

【0031】

例えば、図 8 に示すように、各属性について合意済み差別度が設定されたとする。この場合、属性「収入」、「返済期間」、及び「年齢」については、1 つの合意済み差別度が設定されているため、受付部 14 は、その合意済み差別度をそのまま最終的な差別度として受け付ける。また、属性「職業」、「目的」、及び「住所」については、複数の合意済み差別度が設定されている。そこで、受付部 14 は、例えば、図 9 に示すような選択画面 32 を、第 2 の評価者の各々が利用する情報処理装置の表示装置に表示する。第 2 の評価者は、第 1 の評価者集団に含まれる第 1 の評価者とは異なる評価者でもよいし、少なくとも一部の第 2 の評価者を第 1 の評価者と共通としてもよい。

【0032】

図 9 の例では、選択画面 32 には、複数の合意済み差別度のいずれか、又は、いずれの

10

20

30

40

50

合意済み差別度にも合意しないこと（図9中の「なし」）を選択するためのラジオボタン等の表示部品が表示される。第2の評価者は、いずれかの合意済み差別度、又は「なし」を選択する。このように、複数ある属性のうち、合意済み差別度が複数設定されている属性、すなわち、第1の評価者集団内で差別度の評価が割れている属性が、その属性の合意済み差別度を選択肢として優先的に表示される。これにより、第2の評価者は、1つの属性に対して考えるべき多様なケースを自力で列挙する必要がなくなり、第2の評価者が評価困難な属性を評価する際の認知コストを軽減し、各属性の差別度の早期合意を促すことができる。

【0033】

また、受付部14は、第2の評価者が、いずれの合意済み差別度にも合意しなかった場合、すなわち、選択画面32において「なし」が選択された場合、例えば図10に示すような追加受付画面34を第2の評価者が利用する情報処理装置の表示装置に表示する。図10の例では、追加受付画面34には、「なし」が選択された属性について、第2の評価者に差別度を選択させるためのスライドバーが表示される。受付部14は、第2の評価者から、追加受付画面34を介して、差別度を受け付ける。受付部14は、第2の評価者から受け付けた差別度と、第1の評価者集団から受け付けた差別度と合わせて、合意済み差別度の設定からやり直す。

【0034】

受付部14は、選択画面32を介して第2の評価者に選択された差別度を集計し、例えば、第2の評価者により最も多く選択された合意済み差別度を、最終的な差別度として受け付ける。図11に、第2の評価者に選択された差別度の集計結果の一例を示す。図11の例では、第2の評価者は100人である。例えば、属性「職業」について、受付部14は、合意済み差別度「2.11」、「3.98」、及び「6.53」のうち、第2の評価者により最も多く選択された「6.53」を、最終的な差別度として受け付ける。図11の下段に、各属性の最終的な差別度の一例を示す。本実施形態では、最終的な差別度が0より大きい属性が間接属性として特定される。

【0035】

決定部16は、受付部14により受け付けられた最終的な差別度と、データ集合における属性毎のデータ数とに基づいて、属性毎の寄与度の基準値を決定する。具体的には、図12に示すように、決定部16は、算出部12により算出された寄与度を、最終的な差別度の大きさに応じた割合で引き下げた値を、属性毎の寄与度の基準値として決定する。これにより、図12に示すように、寄与効果を考慮した属性毎の基準値が決定される。

【0036】

例えば、決定部16は、算出部12により算出された属性毎の寄与度から、最終的な差別度の $1/10$ で表される割合を減じた値を、その属性の寄与度の基準値として決定する。例えば、属性「職業」について、寄与度0.65、最終的な差別度6.53であるため、決定部16は、 $0.65 - (0.65 \times 0.653) = 0.23$ を寄与度の基準値として決定する。図13に、上記の計算例で計算された各属性の寄与度の基準値を示す。

【0037】

生成部18は、決定部16により決定された属性毎の寄与度の基準値に応じて、データ集合に含まれる少なくとも一部のデータの属性を変更することによって、機械学習用の訓練データを生成する。具体的には、生成部18は、属性毎の寄与度の基準値と、算出部12により算出される寄与度との差が所定値以下となるように、データ集合に含まれる少なくとも一部のデータの属性「判定結果」の属性値、すなわちラベルを変更する。より具体的には、生成部18は、算出部12により算出される寄与度が基準値以下となるように、データ集合に含まれる少なくとも一部のデータについて、属性「判定結果」の属性値を、元のデータの属性値とは異なる判定結果を示す属性値に変更する。

【0038】

例えば、生成部18は、マッサージングという手法を用いて、一部のデータの属性を変更する。図14に、マッサージングにより、属性「目的」の寄与度を基準値以下となるよ

10

20

30

40

50

うにラベルを変更する場合を示す。この場合、生成部 18 は、データ集合に含まれる各データを、例えばランダムフォレスト等の一般的な分類器で分類し、各データが、どの程度の確実さで、そのデータのラベルが示すクラスに分類されたかを示す分類スコアをデータ毎に算出する。図 14 において、中に数字が表記された丸印が各データに相当し、その数字が各データの分類スコアを表している。また、斜線の丸印は、そのデータのラベルが「承認」であることを表し、白い丸印は、そのデータのラベルが「却下」であることを表している。また、図 14 では、各データを示す丸印と共に、属性「目的」の属性値が「中古車の購入」、かつラベルが「承認」のデータ数、及び属性「目的」の属性値が「休暇中の予算」、かつラベルが「承認」のデータ数を示す棒グラフを示している。ある属性についての属性値間における同一ラベルのデータ数の差（図 14 中の A）が、その属性の寄与度に相当する。

10

【0039】

生成部 18 は、属性値毎に、ラベルの境界に向かって分類スコアが低くなるようにデータをソートする。すなわち、境界に最も近い、ラベルが「承認」のデータは、そのデータが「承認」に分類される確率が最も低いデータであり、境界に最も近い、ラベルが「却下」のデータは、そのデータが「却下」に分類される確率が最も低いデータである。ここで、図 13 の例では、属性「目的」の寄与度の基準値は、算出部 12 により算出された寄与度 0.60 から差別度に応じて引き下げられた 0.43 となっている。したがって、生成部 18 は、境界に最も近いデータから順に、寄与度が小さくなるように、ラベルを他方のラベルに変更する。すなわち、生成部 18 は、属性値が「中古車の購入」のデータについては、境界に最も近い、ラベルが「承認」のデータのラベルを「却下」に変更する（図 14 中の白い二重丸で示すデータ）。また、生成部 18 は、属性値が「休暇中の予算」のデータについては、境界に最も近い、ラベルが「却下」のデータのラベルを「承認」に変更する（図 14 中の斜線の二重丸で示すデータ）。これにより、属性「目的」についての属性値間における、ラベルが「承認」のデータ数の差に相当する寄与度が小さくなり、決定部 16 により決定された基準値に近づく（図 14 中の B）。

20

【0040】

生成部 18 は、上記のようなラベルの変更を、寄与度が基準値以下となるまで繰り返し、寄与度が基準値以下となった場合におけるラベルが付与されたデータを、訓練データとする。生成部 18 は、生成した複数の訓練データを含む訓練データ集合を出力する。なお、ラベルの変更は、上記のマッサージングにより行う場合に限定されず、e l i f t や信頼度といった他の基準に基づいてデータを変更するようにしてもよい。

30

【0041】

出力された訓練データ集合は、図 1 に示すように、機械学習装置 20 において、ニューラルネットワーク等のモデルの機械学習に用いられる。具体的には、機械学習装置 20 は、訓練データに含まれる、判定に用いられる属性の属性値を入力した際のモデルの出力と、判定結果を示す属性の属性値、すなわちラベルとが対応するように、モデルのパラメータを機械学習する。

【0042】

訓練データ生成装置 10 は、例えば図 15 に示すコンピュータ 40 で実現することができる。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、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。CPU 41、メモリ 42、記憶部 43、入出力装置 44、R/W 部 45、及び通信 I/F 46 は、バス 47 を介して互いに接続される。

40

【0043】

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できる。記憶媒体としての記憶部 43 には、コンピュータ 40

50

を、訓練データ生成装置 10 として機能させるための訓練データ生成プログラム 50 が記憶される。訓練データ生成プログラム 50 は、算出プロセス 52 と、受付プロセス 54 と、決定プロセス 56 と、生成プロセス 58 とを有する。

【0044】

CPU 41 は、訓練データ生成プログラム 50 を記憶部 43 から読み出してメモリ 42 に展開し、訓練データ生成プログラム 50 が有するプロセスを順次実行する。CPU 41 は、算出プロセス 52 を実行することで、図 1 に示す算出部 12 として動作する。また、CPU 41 は、受付プロセス 54 を実行することで、図 1 に示す受付部 14 として動作する。また、CPU 41 は、決定プロセス 56 を実行することで、図 1 に示す決定部 16 として動作する。また、CPU 41 は、生成プロセス 58 を実行することで、図 1 に示す生成部 18 として動作する。これにより、訓練データ生成プログラム 50 を実行したコンピュータ 40 が、訓練データ生成装置 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 CPU 41 はハードウェアである。

10

【0045】

なお、訓練データ生成プログラム 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。

【0046】

次に、本実施形態に係る訓練データ生成装置 10 の作用について説明する。訓練データ生成装置 10 にデータ集合が入力されると、訓練データ生成装置 10 が、図 16 に示す訓練データ生成処理を実行する。なお、訓練データ生成処理は、開示の技術の訓練データ生成方法の一例である。

20

【0047】

ステップ S11 で、算出部 12 が、訓練データ生成装置 10 に入力されたデータ集合を取得し、属性毎の寄与度を算出する。次に、ステップ S12 で、受付部 14 が、第 1 の評価者集団に含まれる複数の第 1 の評価者の各々が利用する情報処理装置の表示装置に、例えば図 4 に示すような受付画面 30 を表示する。そして、受付部 14 は、複数の第 1 の評価者の各々から、属性毎の差別度を受け付ける。

【0048】

次に、ステップ S13 で、受付部 14 が、後述するステップ S14 で実行されるクラスタリング処理におけるクラスタ数を示す変数 k に 1 を設定する。次に、ステップ S14 で、受付部 14 が、属性毎に、上記ステップ S12 で受け付けられた差別度を k 個にクラスタリングする。そして、受付部 14 が、各クラスタについて差別度の分散を計算する。なお、 $k=1$ の場合、受付部 14 は、本ステップにおけるクラスタリング処理を省略し、上記ステップ S12 で受け付けられた差別度を 1 つのクラスタとして、差別度の分散を計算する。

30

【0049】

次に、ステップ S15 で、受付部 14 が、属性毎に、各クラスタに含まれる差別度の分散が、予め定めた所定値以下か否かを判定する。全てのクラスタについての分散が所定値以下の場合には、処理はステップ S17 へ移行する。一方、いずれかのクラスタの分散が所定値を超えている場合には、処理はステップ S16 へ移行し、受付部 14 が、 k を 1 インクリメントして、処理はステップ S14 へ戻る。ステップ S17 では、受付部 14 が、属性毎に、各クラスタに含まれる差別度の平均の各々を、その属性についての合意済み差別度として設定する。

40

【0050】

次に、ステップ S18 で、受付部 14 が、合意済み差別度が複数設定されている属性について、第 2 の評価者の各々が利用する情報処理装置の表示装置に、例えば図 9 に示すような選択画面 32 を表示する。これにより、受付部 14 は、第 2 の評価者の各々に、合意済み差別度が複数設定されている属性について、合意済み差別度の各々を選択肢として提示する。

50

【0051】

次に、ステップS19で、受付部14が、第2の評価者により、いずれかの合意済み差別度が選択されたか否かを判定する。いずれかの合意済み差別度が選択された場合には、処理はステップS20へ移行し、受付部14が、選択画面32を介して第2の評価者を選択された差別度を集計し、第2の評価者により最も多く選択された合意済み差別度を、最終的な差別度として受け付ける。一方、いずれの合意済み差別度も選択されなかった場合、すなわち、選択画面32において「なし」が選択された場合、処理はステップS21へ移行する。ステップS21では、受付部14が、例えば図10に示すような追加受付画面34を、「なし」を選択した第2の評価者が利用する情報処理装置の表示装置に表示する。そして、受付部14は、第2の評価者から、追加受付画面34を介して差別度を受け付け、第1の評価者集団から受け付けた差別度に、第2の評価者から受け付けた差別度を追加し、処理はステップS13へ戻る。

10

【0052】

次に、ステップS22で、決定部16が、上記ステップS11で算出された寄与度を、上記ステップS20で受け付けられた最終的な差別度の大きさに応じた割合で引き下げた値を、属性毎の寄与度の基準値として決定する。次に、ステップS23で、生成部18が、属性毎に、寄与度が基準値以下となるように、データ集合に含まれる少なくとも一部のデータについて、属性「判定結果」の属性値（ラベル）を、元のデータの属性値とは異なる判定結果を示す属性値に変更する。これにより、生成部18は、訓練データを生成し、生成した複数の訓練データを訓練データ集合として出力して、訓練データ生成処理は終了する。

20

【0053】

そして、機械学習装置20が、訓練データ生成装置10から出力された訓練データ集合を用いて、ニューラルネットワーク等のモデルを機械学習する。

【0054】

以上説明したように、本実施形態に係る訓練データ生成装置は、複数のデータに含まれる属性毎のデータ数に基づいて算出された値に対して、評価値を受け付け、受け付けた評価値と属性毎のデータ数とに基づいて、属性毎の基準値を決定する。そして、訓練データ生成装置は、属性毎の基準値に応じて、複数のデータの少なくとも一部のデータの属性を変更することによって機械学習用の訓練データを生成する。これにより、文脈を考慮した差別属性による差別的な判定へのバイアスが適切に低減されるモデルを機械学習するための訓練データを生成することができる。

30

【0055】

また、本実施形態に係る訓練データ生成装置は、最終的な差別度を受け付ける過程で、同一の属性に対して複数の合意済み差別度を設定することで、同一の属性に対して多様な差別度を認め、評価者間で評価が割れる属性に関しても、早期の合意を促すことができる。例えば、コンピュータが、第1の評価者集団により評価された差別度の平均を単一の合意済み差別度として第2の評価者集団に提示した場合、提示された合意済み差別度に納得できる第2の評価者が少なくなり、早期合意を阻害する可能性がある。一方、本実施形態に係る訓練データ生成装置は、複数の合意済み差別度を提示することで、より多くの第2の評価者が提示された差別度で納得できるようになり、早期合意を促進することができる。

40

【0056】

また、本実施形態に係る訓練データ生成装置は、複数の合意済み差別度が設定されている属性について、複数の合意済み差別度を選択肢として第2の評価者に提示する。すなわち、属性の数が膨大であったとしても、第2の評価者は、複数の属性の中の一部の属性について、選択肢の中から差別度を選択すればよい。このため、第2の評価者が熟考プロセスに入ることを防止できる、又は、熟考プロセスに入った第2の評価者が早期に熟考プロセスから抜け出すことができるようになる。結果として、全ての評価者の各々が全ての属性について差別度を評価する場合に比べ、各属性の差別度について、複数の評価者間で早期合意を図ることができる。

50

【 0 0 5 7 】

なお、上記実施形態では、訓練データ生成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】

【 0 0 5 8 】

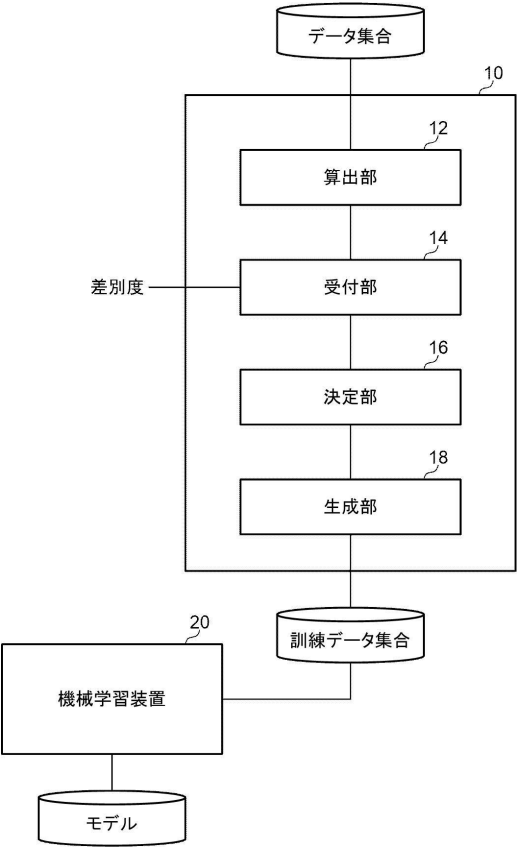
1 0	訓練データ生成装置	
1 2	算出部	
1 4	受付部	10
1 6	決定部	
1 8	生成部	
2 0	機械学習装置	
3 0	受付画面	
3 2	選択画面	
3 4	追加受付画面	
4 0	コンピュータ	
4 1	ＣＰＵ	
4 2	メモリ	
4 3	記憶部	20
4 9	記憶媒体	
5 0	訓練データ生成プログラム	

30

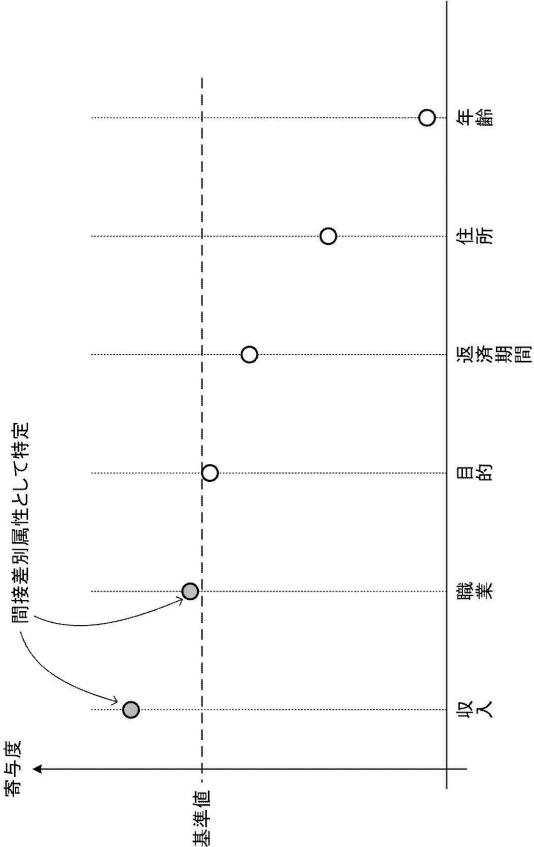
40

50

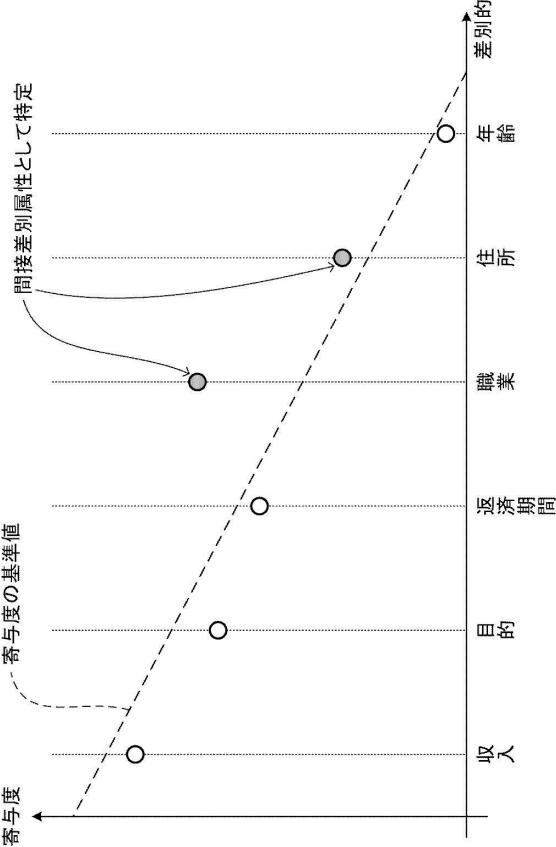
【図面】
【図 1】



【図 2】



【図 3】

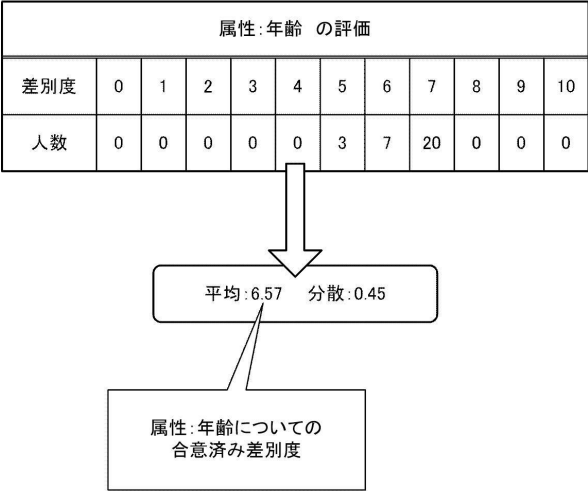


【図 4】

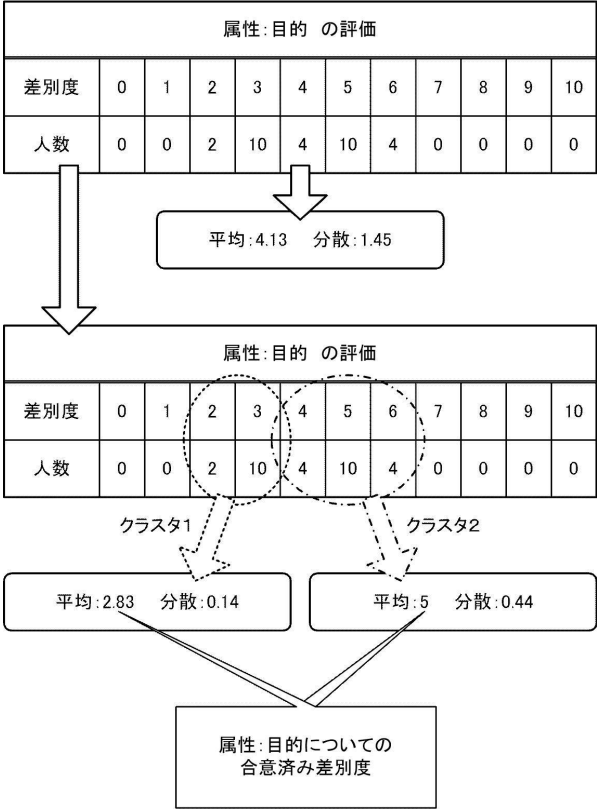
スライドバーで各属性の差別度を選択してください
差別度: 差別的でなくするために寄与度を小さくする割合

属性	寄与度	(非差別的) ← 差別度 → (差別的)	差別度
収入	0.80	0 1 2 3 4 5 6 7 8 9 10	0
職業	0.65	0 1 2 3 4 5 6 7 8 9 10	5
目的	0.60	0 1 2 3 4 5 6 7 8 9 10	3
返済期間	0.50	0 1 2 3 4 5 6 7 8 9 10	0
住所	0.30	0 1 2 3 4 5 6 7 8 9 10	3
年齢	0.05	0 1 2 3 4 5 6 7 8 9 10	7

【図 5】



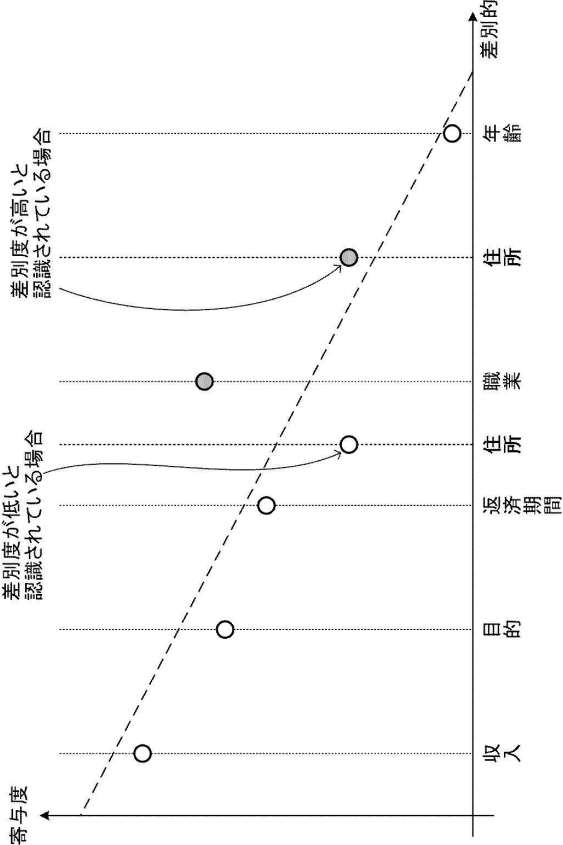
【図 6】



10

20

【図 7】



【図 8】

属性	寄与度	合意済み差別度		
収入	0.80	0.00		
職業	0.65	2.11	3.98	6.53
目的	0.60	2.83	5.00	
返済期間	0.50	0.00		
住所	0.30	3.21	6.34	
年齢	0.05	5.15		

30

40

【図 9】

32

妥当だと思う差別度を選択してください

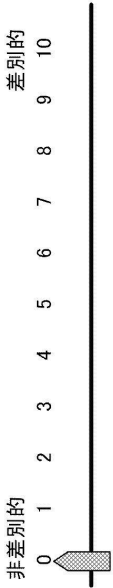
属性	寄与度	合意済み差別度		
職業	0.65	2.11	3.98	6.53
		☐	☐	☐
目的	0.60	2.83	5.00	なし
		☐	☐	☐
住所	0.30	3.21	6.34	なし
		☐	☐	☐

【図 10】

34

属性	寄与度	合意済み差別度		
住所	0.30	3.21	6.34	なし
		☐	☐	☐

どの程度の差別度が妥当だと思いますか？



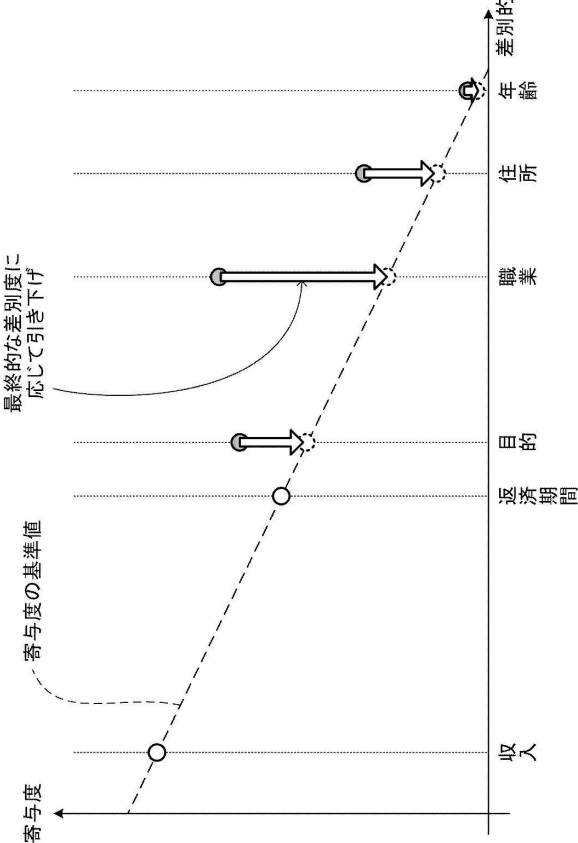
【図 11】

属性	寄与度	合意済み差別度		
収入	0.80	0.00		
職業	0.65	2.11	3.98	6.53
		選択した人数→ 10人	15人	75人
目的	0.60	2.83	5.00	
		選択した人数→ 75人	25人	
返済期間	0.50	0.00		
住所	0.30	3.21	6.34	
		選択した人数→ 40人	60人	
年齢	0.05	5.15		

↓

属性	寄与度	最終的な差別度
収入	0.80	0.00
職業	0.65	6.53
目的	0.60	2.83
返済期間	0.50	0.00
住所	0.30	6.34
年齢	0.05	5.15

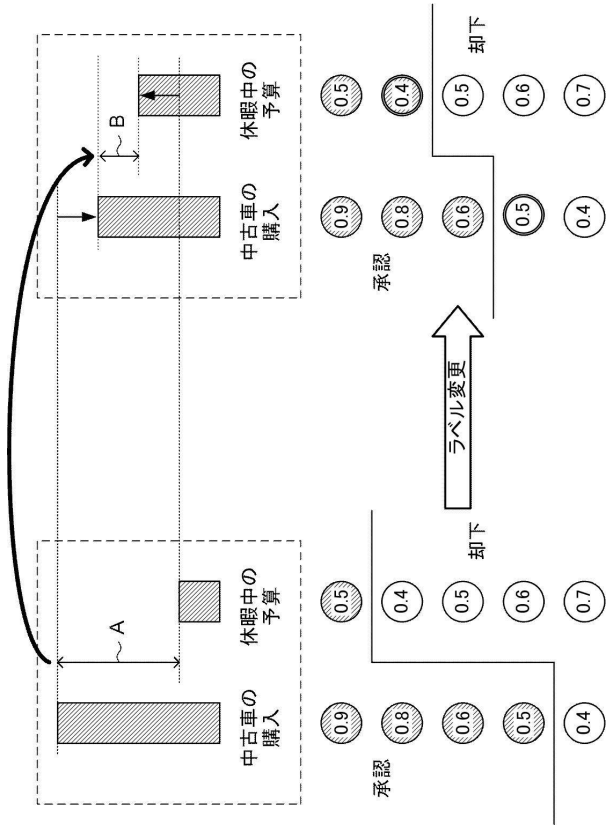
【図 12】



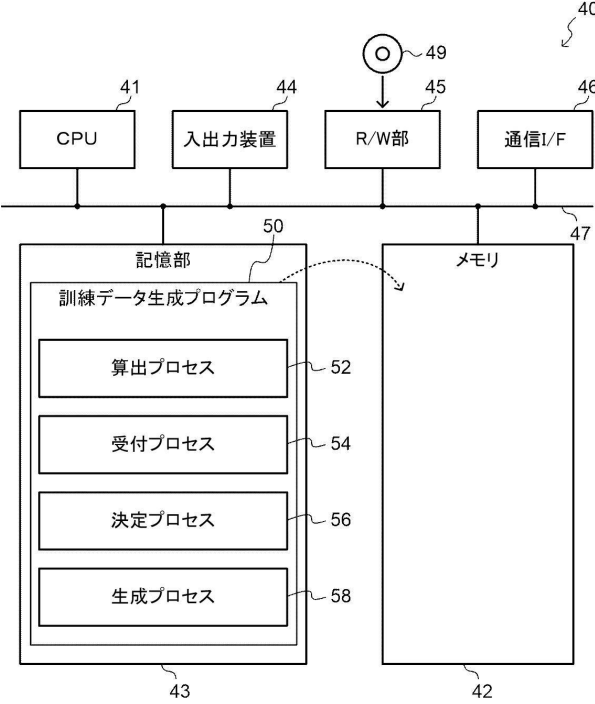
【図 1 3】

属性	寄与度	最終的な差別度	寄与度の基準値
収入	0.80	0.00	0.80
職業	0.65	6.53	0.23
目的	0.60	2.83	0.43
返済期間	0.50	0.00	0.50
住所	0.30	6.34	0.11
年齢	0.05	5.15	0.02

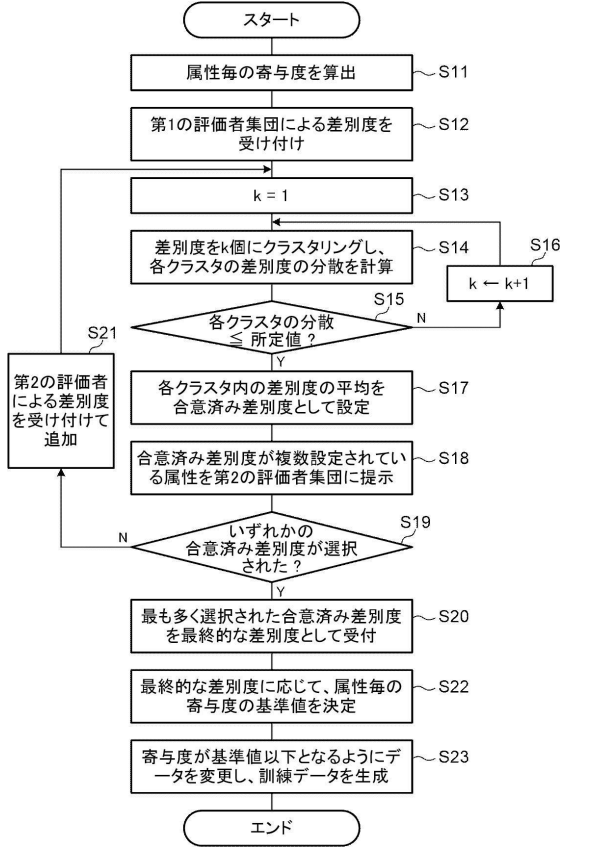
【図 1 4】



【図 1 5】



【図 1 6】



10

20

30

40

50

フロントページの続き

- (56)参考文献 米国特許出願公開第2019/0188605 (US, A1)
米国特許出願公開第2019/0147371 (US, A1)
特開2010-204966 (JP, A)
- (58)調査した分野 (Int.Cl., DB名)
- | | |
|------|-------|
| G06N | 20/00 |
| G06N | 3/00 |
| G06T | 7/00 |