(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
2021年11月18日(18.11.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/229630 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号： PCT/JP2020/018777

(22) 国際出願日： 2020年5月11日(11.05.2020)

(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: 松尾 達(**MATSUO, Tatsuru**); 〒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: 真田 有，外(**SANADA, Tamotsu** et al.); 〒1800004 東京都武蔵野市吉祥寺本町１丁目１０番３１号 ＮＭＦ吉祥寺本町ビル８階 Tokyo (JP).

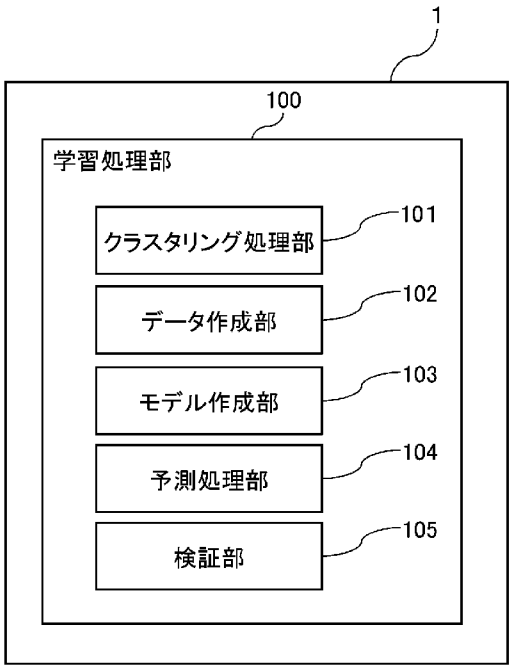
(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, 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, 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,

WO 2021/229630 A1

(54) **Title:** MACHINE LEARNING PROGRAM, MACHINE LEARNING METHOD, AND MACHINE LEARNING DEVICE

(54) 発明の名称： 機械学習プログラム，機械学習方法および機械学習装置



100 Learning processing unit
101 Clustering processing unit
102 Data creation unit
103 Model creation unit
104 Prediction processing unit
105 Verification unit

(57) **Abstract:** This invention makes it possible to avoid overtraining by causing a computer to execute processing for: clustering a plurality of data pieces; generating a model based on machine learning using data pieces sorted into a first group by the clustering; and using data pieces sorted into a second group by the clustering to verify the output accuracy of the generated model.

［続葉有］

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)）

(57) 要約：複数のデータをクラスタリングし、クラスタリングで第１のグループに分類されたデータを用いた機械学習によってモデルを生成し、クラスタリングで第２のグループに分類されたデータを用いて、生成されたモデルの出力精度を検証する、処理をコンピュータに実行させることで、過学習を回避できるようにする。

明　細　書

発明の名称 ：

機械学習プログラム，機械学習方法および機械学習装置

技術分野

[0001] 本発明は、機械学習技術に関する。

背景技術

[0002] 機械学習の手法の一つとして、入力データと出力データとを備える教師データ（正解付きデータ）を用いて入出力関係を学習させる教師あり学習が知られている。

[0003] また、一般的に、教師あり機械学習においては、複数の教師データのうちの一部をモデル作成用データ（訓練データ）として使用することでモデル（機械学習モデル）を作成し、複数の教師データのうち残りの一部をモデル検証用データ（評価データ）として使用することで過学習の判断を行なうことが知られている。

[0004] ここで、過学習（overfitting）とは、モデルが訓練データだけに最適化されてしまい汎用性がない状態に陥ることをいい、モデル作成用データについては高精度に予測できるが、それ以外のデータについては予測が低精度となる。

[0005] 上述した教師データの一部をモデル検証用データとして用いた過学習の判断手法においては、作成したモデルを用いてモデル作成用データを予測した時の予測精度とモデル検証用データを予測した時の予測精度が大幅に異なる場合に過学習の状態と判断される。

先行技術文献

特許文献

[0006] 特許文献1：特開２０１９－６６９９３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、教師データ取得時に偏りがある場合、入力データ空間全体で見たときに、例外的なクラスタが存在する可能性がある。

[0008] なお、ここでいう「偏り」とは、入力に関するものであり、偶然では起こりえないほど入力が類似したデータ群（クラスタ）が教師データに含まれている状況を指す。取得可能な教師データに制限がある等の事情により、このような偏りが生じ得る。

[0009] このような例外的なクラスタのデータまで正しく予測するモデルを作成すると過学習になりやすいが、上述の通りモデル作成／検証用データのいずれも高精度に予測できてしまうため、過学習となったことが検出されない場合がある。

[0010] １つの側面では、過学習を抑制することを目的とする。

課題を解決するための手段

[0011] このため、この機械学習プログラムは、複数のデータをクラスタリングし、前記クラスタリングで第１のグループに分類されたデータを用いた機械学習によってモデルを生成し、前記クラスタリングで第２のグループに分類されたデータを用いて、生成された前記モデルの出力精度を検証する、処理をコンピュータに実行させる。

発明の効果

[0012] 一実施形態によれば、過学習を抑制することができる。

図面の簡単な説明

[0013] [図1]実施形態の一例としての計算機システムのハードウェア構成を例示する図である。

[図2]実施形態の一例としての計算機システムの機能構成を例示する図である。

[図3]ニューラルネットワークの概要を示す図である。

[図4]実施形態の一例としての計算機システムのクラスタリング処理部によるクラスタリング手法を説明するための図である。

[図5]実施形態の一例としての計算機システムのデータ作成部による処理を説

明するための図である。

[図6]実施形態の一例としての計算機システムのモデル作成部による処理を説明するための図である。

[図7]実施形態の一例としての計算機システムの予測処理部による処理を説明するための図である。

[図8]実施形態の一例としての計算機システムの検証部による処理を説明するための図である。

[図9]実施形態の一例としての計算機システムにおける処理を説明するためのフローチャートである。

[図10]実施形態の一例としての計算機システムの学習処理部により行なわれる二値分類を説明するための図である。

[図11]図１０におけるモデル作成用クラスタを抽出して示す図である。

[図12]図１０におけるモデル検証用クラスタを抽出して示す図である。

[図13]機械学習方法における過学習について説明するための図である。

発明を実施するための形態

[0014] 図１３は機械学習方法における過学習について説明するための図であり、入力データ空間に配置された教師データを例示する。この図１３においては、多数の微小点が配置された入力データ空間を例示している。微小点のそれぞれは教師データを表しており、それぞれ入力データに応じた位置にプロットされている。

[0015] また、この入力データ空間においては、複数の教師データ（微小点）が局所的に集まることで複数の小規模のクラスタ（データ群）が形成されている。図１３中においては、教師データの集合によるクラスタに符号ａまたは符号ｂが付されている。

[0016] これらの符号ａまたは符号ｂは教師データの出力を表しており、符号ａが付されたクラスタを構成する教師データの出力はそれぞれａであり、符号ｂが付されたクラスタを構成する教師データの出力はそれぞれｂである。すなわち、図１３に示す例においては、ａまたはｂを予測する二値分類を表して

いる。

[0017] この図１３に示す例において、太破線は、モデル作成用データを用いて全て正解できる高精度なモデルを作成した場合の予測の境界を示す。当該モデルにおいては、この太破線の左側に位置する教師データの出力をｂと予測し、この太破線の右側に位置する教師データの出力をａと予測する。

[0018] ここで、この図１３に例示する入力データ空間上の教師データには、モデル作成用のデータとモデル検証用のデータとが混在しており、各クラスタを構成する教師データにも、モデル作成用のデータとモデル検証用のデータとが混在している。その場合、図１３における太破線を境界として予測するモデルはモデル作成用データおよびモデル検証用データのいずれに対しても高精度に予測できる。

[0019] しかしながら、教師データ取得時に偏りがある場合、入力データ空間全体で見たときに、例外的なクラスタが存在する可能性がある。

[0020] このような例外的なクラスタのデータまで正しく予測するモデルを作成すると過学習になりやすいが、上述の通りモデル作成／検証用データのいずれも高精度に予測できてしまうため、過学習となったことが検出されない場合がある。例えば、図１３中において、符号Ｐ１を付して示す四角点線で囲んだクラスタが例外的なクラスタであった場合、図１３に示す入力空間においては、太破線を境界として予測するモデルは過学習の状態にあり、同図中に一点鎖線を境界として予測するモデルの方が望ましい。機械学習においては、このような過学習を抑制することが望まれている。

[0021] 以下、図面を参照して本機械学習プログラム，機械学習方法および機械学習装置にかかる実施の形態を説明する。ただし、以下に示す実施形態はあくまでも例示に過ぎず、実施形態で明示しない種々の変形例や技術の適用を排除する意図はない。すなわち、本実施形態を、その趣旨を逸脱しない範囲で種々変形（実施形態および各変形例を組み合わせる等）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。

[0022] 図１は実施形態の一例としての計算機システム１のハードウェア構成を例示する図である。計算機システム１は、機械学習装置であって、例えば、ニューラルネットワークを実現する。計算機システム１は、図１に示すように、ＣＰＵ（Central Processing Unit）１０，メモリ１１およびアクセラレータ１２を備える。これらのＣＰＵ１０，メモリ１１およびアクセラレータ１２は、通信バス１３を介して相互に通信可能に接続されている。通信バス１３は、本計算機システム１内のデータ通信を行なう。

[0023] メモリ１１は、ＲＯＭ（Read Only Memory）およびＲＡＭ（Random Access Memory）を含む記憶メモリである。メモリ１１のＲＯＭには、後述するＣＰＵ１０によって実行されるプログラムやこのプログラム用のデータ類が書き込まれている。メモリ１１上のソフトウェアプログラムは、ＣＰＵ１０に適宜読み込まれて実行される。また、メモリ１１のＲＡＭは、一次記憶メモリあるいはワーキングメモリとして利用される。メモリ１１のＲＡＭには、教師データ（モデル作成用データ，モデル検証用データ）やモデルを構成する情報およびモデルを用いた予測結果等も格納される。アクセラレータ１２は、例えば、行列演算などのニューラルネットワークの計算に必要な演算処理を実行する。

[0024] ＣＰＵ１０は、種々の制御や演算を行なう処理装置（プロセッサ）であり、実装されたプログラムに基づき、計算機システム１全体を制御する。そして、このＣＰＵ１０がメモリ１１等に格納された機械学習プログラム（図示省略）を実行することで、後述する学習処理部１００（図２参照）としての機能を実現する。計算機システム１は、機械学習プログラムを実行することにより機械学習装置として機能する。

[0025] なお、学習処理部１００としての機能を実現するためのプログラム（機械学習プログラム）は、例えばフレキシブルディスク，ＣＤ（ＣＤ－ＲＯＭ，ＣＤ－Ｒ，ＣＤ－ＲＷ等），ＤＶＤ（ＤＶＤ－ＲＯＭ，ＤＶＤ－ＲＡＭ，ＤＶＤ－Ｒ，ＤＶＤ＋Ｒ，ＤＶＤ－ＲＷ，ＤＶＤ＋ＲＷ，ＨＤ　ＤＶＤ等），ブルーレイディスク，磁気ディスク，光ディスク，光磁気ディスク等の、コ

ンピュータ読取可能な記録媒体に記録された形態で提供される。そして、コンピュータ（計算機システム１）はその記録媒体からプログラムを読み取って内部記憶装置または外部記憶装置に転送し格納して用いる。また、そのプログラムを、例えば磁気ディスク，光ディスク，光磁気ディスク等の記憶装置（記録媒体）に記録しておき、その記憶装置から通信経路を介してコンピュータに提供するようにしてもよい。

[0026] 学習処理部１００としての機能を実現する際には、内部記憶装置（本実施形態ではメモリ１１のＲＡＭやＲＯＭ）に格納されたプログラムがコンピュータのマイクロプロセッサ（本実施形態ではＣＰＵ１０）によって実行される。このとき、記録媒体に記録されたプログラムをコンピュータが読み取って実行するようにしてもよい。

[0027] 図２は実施形態の一例としての計算機システム１の機能構成を例示する図である。計算機システム１は、図２に示すように、学習処理部１００としての機能を備える。学習処理部１００は、例えば、ニューラルネットワークにおける深層学習を実施する。

[0028] ニューラルネットワークは、ハードウェア回路であってもよいし、ＣＰＵ１０等によりコンピュータプログラム上で仮想的に構築される階層間を接続するソフトウェアによる仮想的なネットワークであってもよい。

[0029] 図３にニューラルネットワークの概要を示す。図３に示すニューラルネットワークは、入力層と出力層との間に複数の隠れ層を含むディープニューラルネットワークである。隠れ層は、例えば、畳み込み層、プーリング層または全結合層等である。図３中において、各層に示す丸印は、所定の計算をそれぞれ実行するノードを示す。

[0030] ニューラルネットワークは、入力データを入力層に入力し、畳み込み層やプーリング層などで構成される隠れ層にて所定の計算を順次実行することで、演算により得られる情報を入力側から出力側に順次伝えるフォーワッド方向の処理（順伝播処理）を実行する。フォーワッド方向の処理の実行後、出力層から出力される出力データと正解データとから得られる誤差関数の値を

小さくするために、フォーワッド方向の処理で使用するパラメータを決定するバックワード方向の処理（逆伝播処理）を実行する。そして、逆伝播処理の結果に基づいて重み等の変数を更新する更新処理が実行される。

[0031] 図２に示すように、学習処理部１００は、クラスタリング処理部１０１，データ作成部１０２，モデル作成部１０３，予測処理部１０４および検証部１０５を備える。

[0032] クラスタリング処理部１０１は、複数の教師データに対して、偏りが認識できるようにクラスタリングを行なうことで、複数のクラスタ（データ群）を作成する。教師データは、予め図示しない記憶装置に格納されてもよく、本計算機システム１の外部から入力されてもよい。クラスタリング処理部１０１は、複数の教師データに対して階層型クラスタリングを行なう。

[0033] 図４は実施形態の一例としての計算機システム１のクラスタリング処理部１０１によるクラスタリング手法を説明するための図である。この図４においては、階層型クラスタリングにおけるデンドログラム（樹形図）を例示している。

[0034] 階層型クラスタリングにおいては、複数の入力データに対して、データ間の距離に応じて結合（グルーピング，マージ）することを繰り返し行なうことでクラスタリングを実現する。

[0035] 本計算機システム１において、クラスタリング処理部１０１は、最遠隣法により、クラスタリングを実現する。なお、最遠隣法におけるデータ間の距離は、例えば、ユーグリッド距離を用いてもよく、適宜変更して実施することができる。

[0036] また、階層型クラスタリングにおいては、例えば、システム管理者等が、同一クラスタとするためのデータ間の距離を閾値として設定することができる。この閾値を設定することで、クラスタリング処理部１０１は、データ間の距離が閾値未満となるデータどうしを同一のクラスタとなるようにクラスタリングする。閾値は、クラスタのマージ停止条件に相当し、例えば、システム管理者等が任意に設定してもよい。図４においては、符号Ｄ０～Ｄ９で

表されるデータに対して階層型クラスタリングを行なう例を示しており、閾値=５が設定されている。

[0037] 隣接する入力データ間の距離が近いものから順に結合（グルーピング，マージ）することで、例えば、データＤ３，Ｄ４が一つのクラスタＣ１を形成している。同様に、データＤ８，Ｄ５，Ｄ７がクラスタＣ２を、データＤ２，Ｄ１，Ｄ６がクラスタＣ５をそれぞれ形成している。データＤ０，Ｄ９はいずれも他のデータからの距離が遠いものであるので、それぞれ単独で独立したクラスタＣ３，Ｃ４を形成する。

[0038] これらのクラスタＣ１～Ｃ５は、各クラスタ内におけるデータ間の距離が閾値（図４に示す例では５）未満であることが保証されており、データ空間内におけるデータの偏りを実現する。

クラスタリング処理部１０１は、このような階層型クラスタリング手法を用いて教師データに偏りが認識されるクラスタリングを実現する。

[0039] また、クラスタのマージ停止条件（閾値）は、教師データ取得時の偏りによるとみなせる入力データ間の距離とすることが望ましい。この閾値は、例えば、対象データに対するドメイン知識を持つ人がデータの素性に基づいて任意に設定してもよい。

[0040] データ作成部１０２は、モデル作成用データ（教師データ）およびモデル検証用データを作成する。モデル作成用データは、後述するモデル作成部１０３が機械学習のモデルを作成するために用いる教師データである。モデル検証用データは、後述する検証部１０５が作成されたモデルの検証を行なうために用いる教師データである。

[0041] 以下、モデル学習用データを用いてモデル作成を行なう過程を学習フェーズ（第１フェーズ）という場合があり、モデル作成用データを用いてモデルの検証を行なう過程を検証フェーズ（第２フェーズ）という場合がある。図５は実施形態の一例としての計算機システム１のデータ作成部１０２による処理を説明するための図である。

[0042] データ作成部１０２は、クラスタリング処理部１０１により作成された複

数のクラスタを、モデル作成用クラスタとモデル検証用クラスタとに分類する。なお、モデル作成用クラスタおよびモデル検証用クラスタの各数は、適宜変更して実施することができる。例えば、複数のクラスタをモデル作成用クラスタまたはモデル検証用クラスタへランダムに振り分けることで分類してもよく、適宜変更して分類を実施することができる。なお、複数のクラスタのモデル作成用クラスタまたはモデル検証用クラスタへの分類は、クラスタリング処理部１０１が行なってもよく、適宜変更して実施することができる。

[0043] 本計算機システム１においては、異なるクラスタのデータを使用して機械学習と検証とを実行する。すなわち、複数のクラスタのうち第１のクラスタ（第１のグループ）のデータを用いて機械学習のモデルを作成し、第２のクラスタ（第２のグループ）のデータを用いてモデルの出力精度の検証を行なう。

モデル作成用クラスタは、機械学習によってモデルを生成するために用いられるデータによる第１のグループであってもよい。また、モデル検証用クラスタは、生成されたモデルの出力精度を検証するために用いられるデータによる第２のグループであってもよい。

[0044] データ作成部１０２は、複数のモデル作成用クラスタからデータを均等にサンプリング（抽出）して、モデル作成用データを作成する。複数のモデル作成用クラスタからデータを均等にサンプリングする理由は、複数のモデル作成用クラスタ間においてデータ数の偏りがある恐れがあるからである。データ作成部１０２は、複数のモデル作成用クラスタから異なるサンプリングを行なうことで、複数のモデル作成用データを作成する。

[0045] 同様に、データ作成部１０２は、複数のモデル検証用クラスタからデータを均等にサンプリング（抽出）して、モデル検証用データを作成する。複数のモデル検証用クラスタからデータを均等にサンプリングする理由は、複数のモデル検証用クラスタ間においてもデータ数の偏りがあるおそれがあるからである。データ作成部１０２は、複数のモデル検証用クラスタから異なる

サンプリングを行なうことで、複数のモデル検証用データを作成する。

[0046] 複数のモデル作成用クラスタ，複数のモデル検証用クラスタ，複数のモデル作成用データおよび複数のモデル検証用データは、それぞれメモリ１１の所定の記憶領域に格納してもよく、また、図示しない記憶装置に格納してもよい。

[0047] モデル作成部１０３は、モデル作成用データ（教師データ）を用いた機械学習によってモデル（学習モデル）を作成する。モデルは、入力値を受け取り、何かしらの評価・判定をして出力値を出力する。モデルの出力を予測結果といってもよい。なお、モデルの作成は既知の手法を用いて実現することができ、モデル作成部１０３によるモデル作成手法の説明は省略する。また、モデル作成部１０３は、機械学習に複数のモデル検証用データを用いることで、これらのモデル作成用データに応じた複数のモデルを作成する。モデル検証用データは、クラスタリングで第３のグループに分類されたデータに相当する。図６は実施形態の一例としての計算機システム１のモデル作成部１０３による処理を説明するための図である。

[0048] 図６に示す例においては、２つのモデル作成用データ＃１，＃２が示されている。モデル作成部１０３は、モデル作成用データ＃１を用いて教師あり学習（機械学習）を行なうことでモデル＃１を作成し、モデル作成用データ＃２を用いて教師あり学習（機械学習）を行なうことでモデル＃２を作成する。作成されたモデル＃１，＃２には、モデル作成用データやモデル検証用データが入力される。モデル作成用データ＃１は、第１のグループに分類されたデータのうちの第１のデータに相当する。モデル作成用データ＃２は、第１のグループに分類されたデータのうちの第２のデータに相当する。

[0049] 予測処理部１０４は、モデル作成部１０３が作成した複数のモデルを用いて、予測対象データをこれらのモデルに入力した場合の出力の予測を行なう。予測処理部１０４は、予測対象データをモデル作成部１０３が作成した複数のモデルのそれぞれに入力し、各モデルの出力（予測結果）をアンサンブル（統合，集計）する。予測処理部１０４は、このアンサンブル結果を最終

的な出力（予測結果）とする。予測処理部１０４は、複数のモデルの各出力を統合（アンサンブル）して一の出力を生成するアンサンブル処理部に相当する。

予測対象データには、第１フェーズにおいてはモデル作成用データが用いられ、第２フェーズにおいてはモデル検証用データが用いられる。すなわち、予測処理部１０４は、第１フェーズにおいては、モデル作成用データを複数のモデルのそれぞれに入力し、各モデルの出力をアンサンブルした結果を最終的な出力（予測結果）とする。

[0050] また、予測処理部１０４は、第２フェーズにおいては、モデル検証用データを複数のモデルのそれぞれに入力し、各モデルの出力をアンサンブルした結果を最終的な出力（予測結果）とする。

図７は実施形態の一例としての計算機システム１の予測処理部１０４による処理を説明するための図である。この図７に示す例においては、２つのモデル＃１，＃２に予測対象データ、すなわち、モデル作成用データもしくはモデル検証用データが入力されている。各モデル＃１，＃２からそれぞれ出力される予測結果がアンサンブルされ、予測結果（予測対象データの予測結果）が出力される。

[0051] この図７に示す例において、予測対象データが第２のグループに分類されたデータに含まれる第３のデータに相当する。予測処理部１０４は、予測対象データ（第３のデータ）のモデル＃１への入力に応じて当該モデル＃１が出力した第１の結果と、予測対象データ（第３のデータ）のモデル＃２への入力に応じて当該モデル＃２が出力した第２の結果とに基づいて、第１の出力精度を算出する。

[0052] なお、複数のモデル出力のアンサンブルは平均値の演算等の既知の手法を用いて実現することができ、予測処理部１０４によるモデル出力のアンサンブル手法の説明は省略する。

[0053] 検証部１０５は、データ作成部１０２によって作成されたモデル検証用データを用いて、モデル作成部１０３が作成したモデルの検証を行なう。図８

は実施形態の一例としての計算機システム１の検証部１０５による処理を説明するための図である。検証部１０５は、データ作成部１０２によって作成されたモデル検証用データを用いて、モデル作成部１０３が作成したモデルの検証を行なう。

[0054] 検証部１０５は、データ作成部１０２によって作成された複数のモデル検証用データを、モデル作成部１０３によって作成された複数のモデルのそれぞれに入力させる。検証部１０５は、例えば、予測処理部１０４の機能を用いて、モデル検証用データ（予測対象データ）をモデル作成部１０３が作成した複数のモデルのそれぞれに入力し、各モデルの出力（予測結果）をアンサンブル（集計）する。予測処理部１０４は、このアンサンブル結果を最終的な出力（予測結果）とする。

[0055] 図８に示す例においては、モデル検証用データ＃１がモデル＃１，＃２にそれぞれ入力され、各モデル＃１，＃２からそれぞれ出力される予測結果がアンサンブルされ、予測結果（モデル検証用データの予測結果）＃１が出力されている。また、モデル検証用データ＃２がモデル＃１，＃２にそれぞれ入力され、各モデル＃１，＃２からそれぞれ出力される予測結果がアンサンブルされ、予測結果（モデル検証用データの予測結果）＃２が出力されている。

[0056] 検証部１０５は、予測結果＃１を、モデル検証用データ＃１の出力データと比較することで正答率（精度）を算出する。また、検証部１０５は、予測結果＃２を、モデル検証用データ＃２の出力データと比較することで正答率（精度）を算出する。検証部１０５は、これらの精度（正答率）の平均を算出することで、モデル検証用クラスタの精度を決定する。

[0057] すなわち、検証部１０５は、各モデル検証用データに対する予測精度の平均を算出して、モデル検証用クラスタについての最終的（全体的）な予測精度を取得する。

[0058] 例えば、検証部１０５は、モデル検証用データに基づいて出力された予測結果の精度と、モデル作成用データに基づいて出力された予測結果の精度と

の差が許容閾値内であるかを判断してもよい。すなわち、検証部１０５は、モデル検証用データに基づいて出力された予測結果の精度と、モデル作成用データに基づいて出力された予測結果の精度とが同レベルの精度であるかを判断してもよい。また、検証部１０５は、モデル検証用データに基づいて出力された予測結果の精度が所定の閾値以上であるかを判断してもよい。

[0059] 図８に示す例において、モデル検証用データ＃１は、第２のグループに分類されたデータに含まれる第３のデータに相当する。モデル検証用データ＃２は、第２のグループに分類されたデータに含まれる第４のデータに相当する。

[0060] 検証部１０５は、モデル検証用データ（第３のデータ）＃１のモデル＃１への入力に応じて当該モデル＃１が出力した第１の結果と、モデル検証用データ（第３のデータ）＃１のモデル＃２への入力に応じて当該モデル＃２が出力した第２の結果とに基づいて予測結果＃１（第１の出力精度）を算出する。

[0061] また、検証部１０５は、モデル検証用データ（第４のデータ）＃２のモデル＃１への入力に応じて当該モデル＃１が出力した第３の結果と、モデル検証用データ（第４のデータ）＃２のモデル＃２への入力に応じて当該モデル＃２が出力した第４の結果とに基づいて予測結果＃２（第２の出力精度）を算出する。検証部１０５は、これらの予測結果＃１（第１の出力精度）と予測結果＃２（第２の出力精度）とに基づいて予測精度の検証を行なう。

[0062] 上述の如く構成された実施形態の一例としての計算機システム１における処理を、図９に示すフローチャート（ステップＳ１～Ｓ４）に従って説明する。

[0063] ステップＳ１において、クラスタリング処理部１０１が、予め用意された教師データに対して階層型クラスタリングを行なうことで、偏りが認識できる複数のクラスタを作成する。データ作成部１０２は、クラスタリング処理部１０１が作成した複数のクラスタを、モデル作成用クラスタとモデル検証用クラスタとに分ける。

[0064] そして、データ作成部１０２は、複数のモデル作成用クラスタからデータを均等にサンプリングしてモデル作成用データを作成する。この際、データ作成部１０２は、複数のモデル作成用クラスタから異なるサンプリングを複数行なうことで、複数のモデル作成用データを作成する。

[0065] また、データ作成部１０２は、複数のモデル検証用クラスタからデータを均等にサンプリングしてモデル検証用データを作成する。この際、データ作成部１０２は、複数のモデル検証用クラスタから異なるサンプリングを複数行なうことで、複数のモデル検証用データを作成する。

ステップＳ２において、モデル作成部１０３は、機械学習にモデル作成用データ（教師データ）を用いてモデルを作成する。

[0066] ステップＳ３において、予測処理部１０４は、モデル作成部１０３が作成した複数のモデルを用いて、予測対象データをこれらのモデルに入力した場合の出力の予測を行なう。

[0067] ステップＳ４において、検証部１０５は、データ作成部１０２によって作成されたモデル検証用データを用いて、モデル作成部１０３が作成したモデルの検証を行なう。

[0068] このように、実施形態の一例としての計算機システム１によれば、クラスタリング処理部１０１が作成した一のクラスタを、データ作成部１０２がモデル作成用データもしくはモデル検証用データのいずれかに割り当てる。これにより、入力データ空間全体で見たときに例外的なクラスタが存在していても、同一クラスタ内のデータはモデル作成データとモデル検証用データとのいずれか一方にしか含まれない。そのため、モデル作成データの予測精度とモデル検証用データの予測精度とが同時に高くなることはない。このように、同一クラスタ内のデータがモデル作成データとモデル検証用データとに分かれることがないため、過学習を回避することができる。

[0069] 図１０は実施形態の一例としての計算機システム１の学習処理部１００により行なわれる二値分類を説明するための図であり、入力データ空間に配置された教師データを例示する。この図１０においては、多数の微小点が配置

された入力データ空間を例示している。微小点のそれぞれは教師データを表しており、それぞれ入力データに応じた位置にプロットされている。

[0070] また、この入力データ空間においては、破線の丸で囲まれた教師データの集合はモデル作成用クラスタを示し、実線の丸で囲まれた教師データの集合はモデル検証用クラスタを示す。

[0071] また、この図１０においては、各クラスタに符号ａまたは符号ｂが付されている。これらの符号ａまたは符号ｂは教師データの出力を表しており、符号ａが付されたクラスタを構成する教師データの出力はそれぞれａであり、符号ｂが付されたクラスタを構成する教師データの出力はそれぞれｂである。すなわち、図１０に示す例においては、ａまたはｂを予測する二値分類を表している。

[0072] この図１０に示す例において、モデル作成用クラスタからサンプリングしたデータで高精度なモデルを作成すると、符号αを付した太破線を境界として予測するモデルとなる。

[0073] 図１１は図１０におけるモデル作成用クラスタを抽出して示す図である。この図１１に示すように、モデル作成用クラスタに関しては、符号αを付した太破線の左側には全ての出力ｂが配置され、その右側には全ての出力ａが配置されている。すなわち、モデル作成用クラスタからサンプリングしたデータに対する予測精度が高いことがわかる。

[0074] 図１２は図１０におけるモデル検証用クラスタを抽出して示す図である。この図１２に示すように、モデル検証用クラスタに関しては、符号αを付した太破線の左側には出力ｂとともに出力ａも配置されており、図１１に示したモデル作成用クラスタからサンプリングしたデータに比べて予測精度が低いことがわかる。すなわち、過学習していると判断できる。図１０に示す例においては、符号βを付した一点鎖線を境界として予測するモデルが過学習のない好適なモデルとなる。

[0075] クラスタリング処理部１０１が階層型クラスタリングを行なうことで、複数の教師データに対して偏りが認識できるようにクラスタリングを行なうこ

とができる。

[0076] モデル作成部１０３が、機械学習に、モデル作成用クラスタに備えられる複数のクラスタデータ群のそれぞれから抽出（サンプリング）して生成したデータ（モデル作成用データ）を用いる。複数のクラスタから均等にサンプリングすることで取得したモデル作成用データを用いることで、モデルの出力精度を向上させることができる。

[0077] 検証部１０５が、複数のモデル検証用データをそれぞれモデルに適用することで、複数のクラスタの各データを検証に反映させることができ、検出精度を向上させることができる。

[0078] 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および各処理は、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。

[0079] 例えば、上述した実施形態においては、第１フェーズにおいて、データ作成部１０２が複数のモデル作成用データを作成し、モデル作成部１０３がこれらの複数のモデル作成用データを用いて複数のモデルを作成しているが、これに限定されるものではない。モデル作成部１０３は、全てのモデル作成用クラスタのデータを用いて一つのモデルを作成してもよい。

[0080] なおこの場合、第２フェーズにおいては、上述した実施形態と同様に、複数のモデル検証用データを作成し、これらの複数のモデル検証用データをモデルにそれぞれ適用することが望ましい。そして、予測処理部１０４は、これらの複数の入力データに基づいて出力された複数の予測結果を用いて精度を求めることが望ましい。

[0081] 検証を行なう際に、複数のクラスタのデータを一つにまとめた場合には、データ数が多いクラスタの精度が優先されてしまい検出精度が低下するおそれがある。そこで、複数のモデル検証用データをそれぞれモデルに適用することで、複数のクラスタの各データを検証に反映させることができ、検出精度を向上させることができる。

[0082] 上述した実施形態においては、機械学習をニューラルネットワークに適用した例を示しているが、これに限定されるものではなく、種々変形して実施することができる。また、上述した開示により本実施形態を当業者によって実施・製造することが可能である。

符号の説明

[0083] １　計算機システム

１０　ＣＰＵ

１１　メモリ

１２　アクセラレータ

１３　通信バス

１００　学習処理部

１０１　クラスタリング処理部

１０２　データ作成部

１０３　モデル作成部

１０４　予測処理部

１０５　検証部

請求の範囲

[請求項1] 複数のデータをクラスタリングし、

前記クラスタリングで第１のグループに分類されたデータを用いた機械学習によってモデルを生成し、

前記クラスタリングで第２のグループに分類されたデータを用いて、生成された前記モデルの出力精度を検証する、

処理をコンピュータに実行させることを特徴とする機械学習プログラム。

[請求項2] 前記クラスタリングは、階層型クラスタリングである、

ことを特徴とする請求項１に記載の機械学習プログラム。

[請求項3] 前記モデルを生成する処理は、前記クラスタリングで第３のグループに分類されたデータを用いた機械学習を含み、

前記検証する処理は、前記クラスタリングで第４のグループに分類されたデータを用いて実行される、

ことを特徴とする請求項１または２に記載の機械学習プログラム。

[請求項4] 前記モデルは、前記第１のグループに分類されたデータのうち第１のデータを用いた機械学習により生成され、

前記第１のグループに分類されたデータのうち第２のデータを用いた機械学習により他のモデルを生成する、

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乃至３のいずれか１項に記載の機械学習プログラム。

[請求項5] 前記検証する処理は、前記第２のグループに分類されたデータに含まれる第３のデータの前記モデルへの入力に応じて前記モデルが出力した第１の結果と、前記第３のデータの前記他のモデルへの入力に応じて前記他のモデルが出力した第２の結果とに基づいて、第１の出力精度を算出する処理を含む、

ことを特徴とする請求項４に記載の機械学習プログラム。

[請求項6] 前記検証する処理は、前記第２のグループに分類されたデータに含

まれる第４のデータの前記モデルへの入力に応じて前記モデルが出力した第３の結果と、前記第４のデータの前記他のモデルへの入力に応じて前記他のモデルが出力した第４の結果とに基づいて算出された第２の出力精度と、前記第１の出力精度とに基づいて実行される、

ことを特徴とする請求項５項に記載の機械学習プログラム。

[請求項7] 複数のデータをクラスタリングする処理と、

前記クラスタリングで第１のグループに分類されたデータを用いた機械学習によってモデルを生成する処理と、

前記クラスタリングで第２のグループに分類されたデータを用いて、生成された前記モデルの出力精度を検証する処理と、

を行なうことを特徴とする機械学習方法。

[請求項8] 前記クラスタリングは、階層型クラスタリングである、

ことを特徴とする、請求項７に記載の機械学習方法。

[請求項9] 前記モデルを生成する処理は、前記クラスタリングで第３のグループに分類されたデータを用いた機械学習を含み、

前記検証する処理は、前記クラスタリングで第４のグループに分類されたデータを用いて実行される、

ことを特徴とする、請求項７または８に記載の機械学習方法。

[請求項10] 前記モデルは、前記第１のグループに分類されたデータのうち第１のデータを用いた機械学習により生成され、

前記第１のグループに分類されたデータのうち第２のデータを用いた機械学習により他のモデルを生成する、

ことを特徴とする、請求項７乃至９のいずれか１項に記載の機械学習方法。

[請求項11] 前記検証する処理は、前記第２のグループに分類されたデータに含まれる第３のデータの前記モデルへの入力に応じて前記モデルが出力した第１の結果と、前記第３のデータの前記他のモデルへの入力に応じて前記他のモデルが出力した第２の結果とに基づいて、第１の出力

精度を算出する処理を含む、

ことを特徴とする、請求項１０に記載の機械学習方法。

[請求項12] 前記検証する処理は、前記第２のグループに分類されたデータに含まれる第４のデータの前記モデルへの入力に応じて前記モデルが出力した第３の結果と、前記第４のデータの前記他のモデルへの入力に応じて前記他のモデルが出力した第４の結果とに基づいて算出された第２の出力精度と、前記第１の出力精度とに基づいて実行される、

ことを特徴とする、請求項１１に記載の機械学習方法。

[請求項13] 複数のデータをクラスタリングし、

前記クラスタリングで第１のグループに分類されたデータを用いた機械学習によってモデルを生成し、

前記クラスタリングで第２のグループに分類されたデータを用いて、生成された前記モデルの出力精度を検証する、

処理部を有することを特徴とする機械学習装置。

[請求項14] 前記クラスタリングは、階層型クラスタリングである、

ことを特徴とする請求項１３に記載の機械学習装置。

[請求項15] 前記モデルを生成する処理は、前記クラスタリングで第３のグループに分類されたデータを用いた機械学習を含み、

前記検証する処理は、前記クラスタリングで第４のグループに分類されたデータを用いて実行される、

ことを特徴とする請求項１３または１４に記載の機械学習装置。

[請求項16] 前記モデルは、前記第１のグループに分類されたデータのうち第１のデータを用いた機械学習により生成され、

前記第１のグループに分類されたデータのうち第２のデータを用いた機械学習により他のモデルを生成する、

ことを特徴とする請求項１３乃至１５のいずれか１項に記載の機械学習装置。

[請求項17] 前記検証する処理は、前記第２のグループに分類されたデータに含

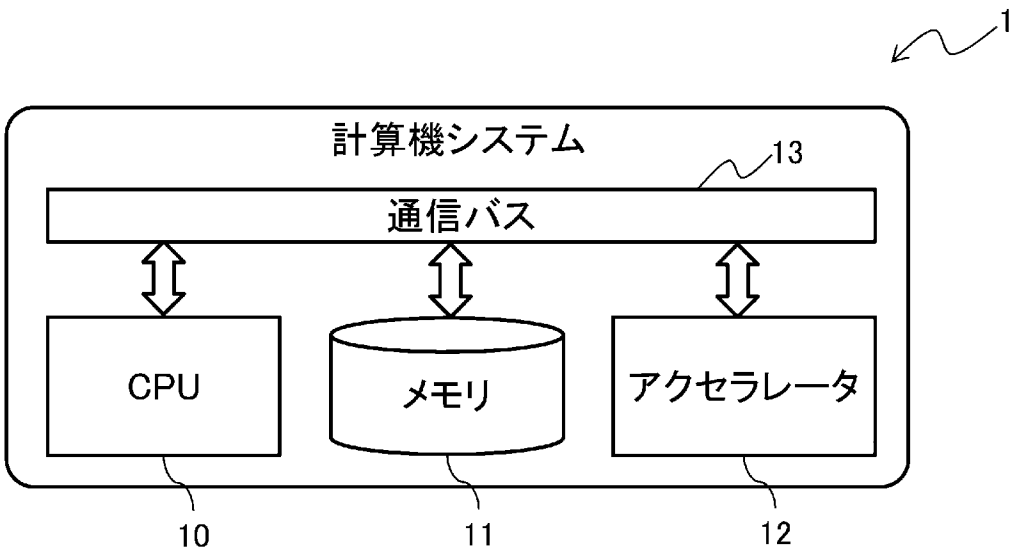
まれる第３のデータの前記モデルへの入力に応じて前記モデルが出力した第１の結果と、前記第３のデータの前記他のモデルへの入力に応じて前記他のモデルが出力した第２の結果とに基づいて、第１の出力精度を算出する処理を含む、

ことを特徴とする請求項１６に記載の機械学習装置。

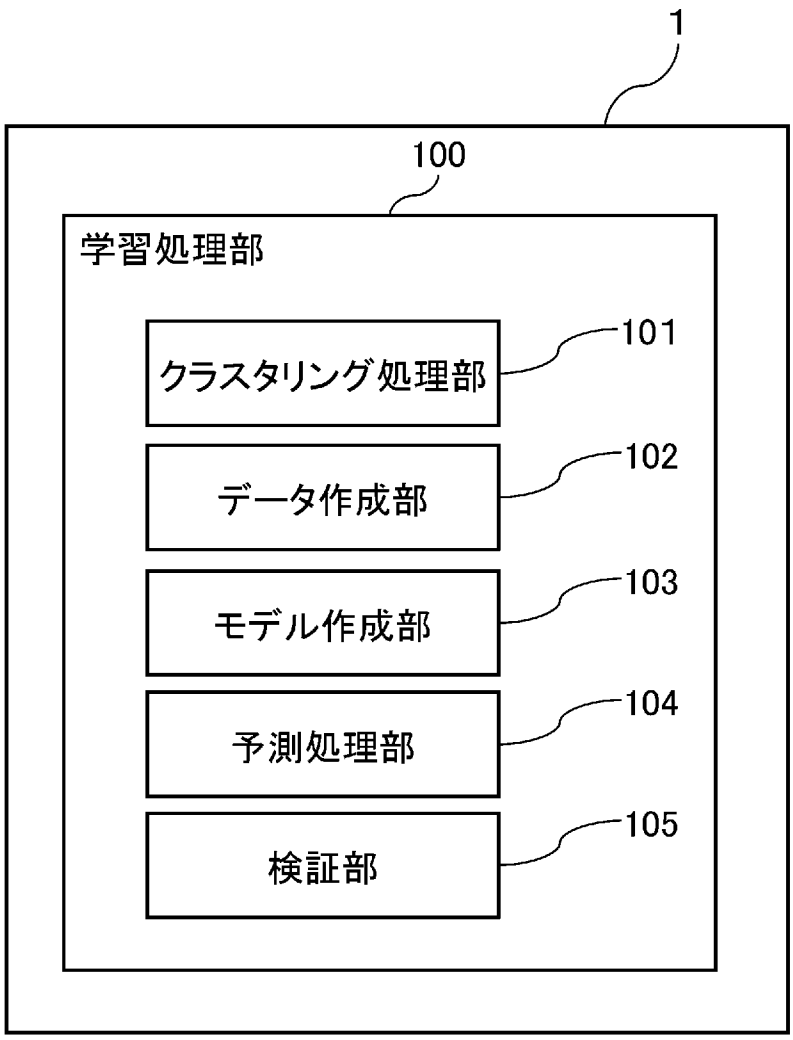
[請求項18] 前記検証する処理は、前記第２のグループに分類されたデータに含まれる第４のデータの前記モデルへの入力に応じて前記モデルが出力した第３の結果と、前記第４のデータの前記他のモデルへの入力に応じて前記他のモデルが出力した第４の結果とに基づいて算出された第２の出力精度と、前記第１の出力精度とに基づいて実行される、

ことを特徴とする請求項１７項に記載の機械学習装置。

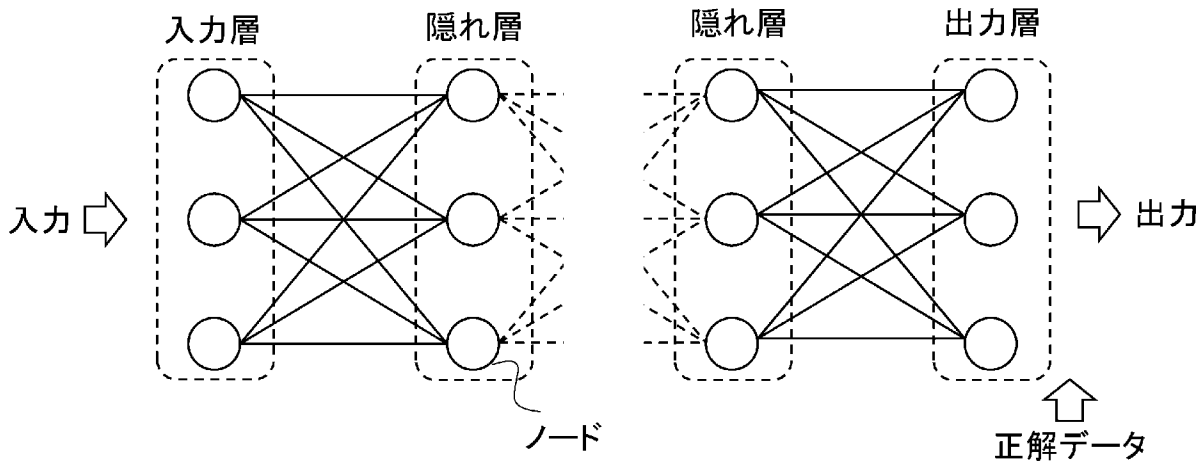
[図1]


1
計算機システム
13
通信バス
CPU
メモリ
アクセラレータ
10
11
12

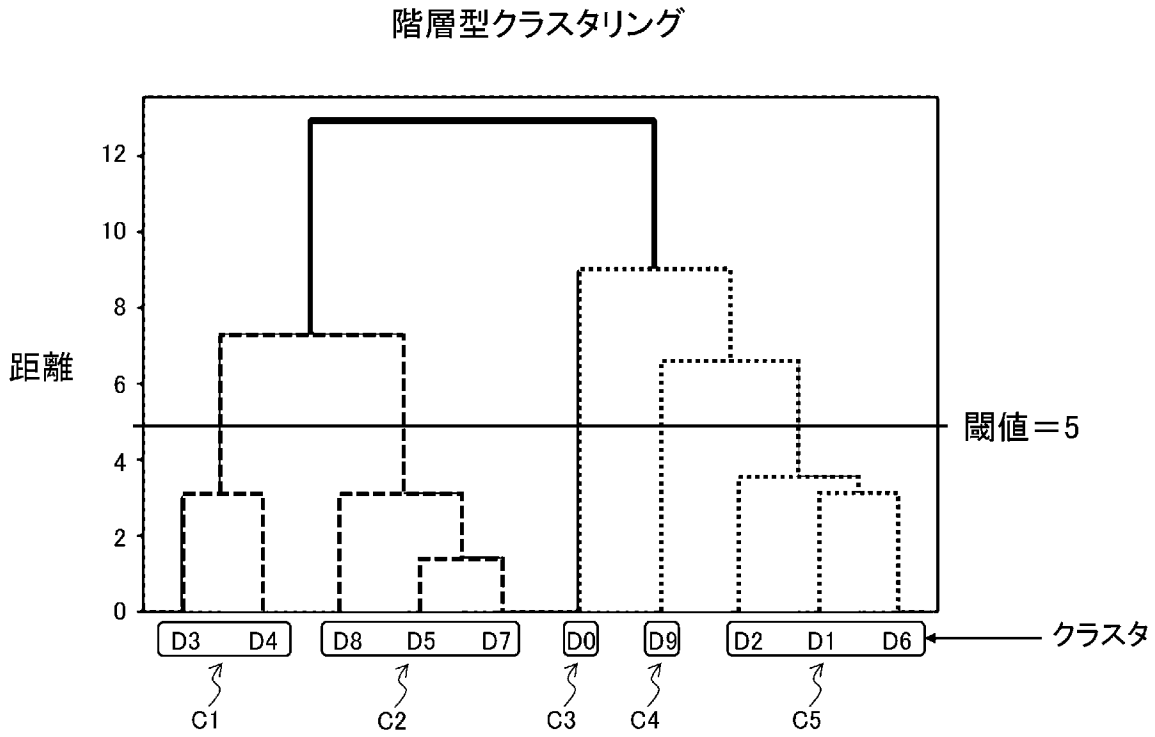
[図2]


1
100
学習処理部
クラスタリング処理部
101
データ作成部
102
モデル作成部
103
予測処理部
104
検証部
105

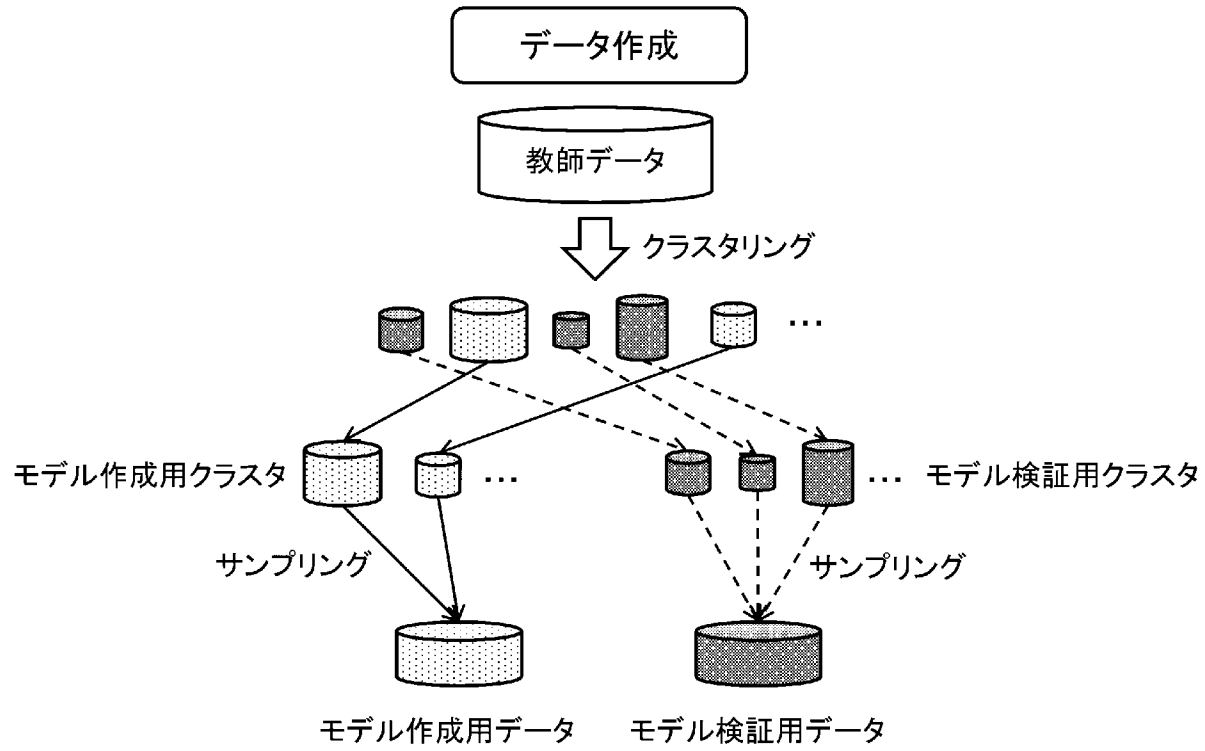
[図3]


入力層
隠れ層
隠れ層
出力層
入力
出力
ノード
正解データ

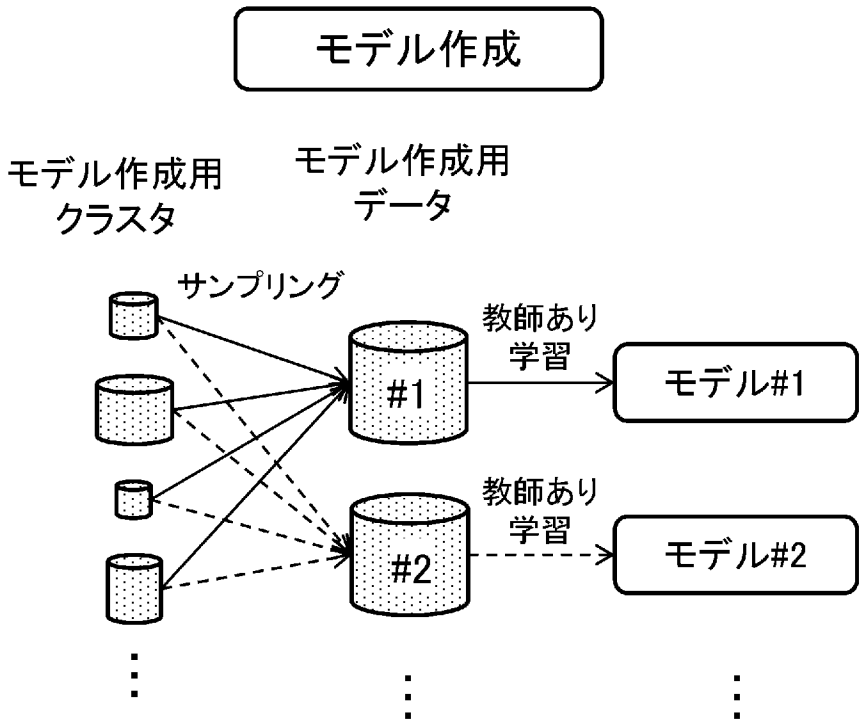
[図4]


階層型クラスタリング
距離
12
10
8
6
4
2
0
閾値＝5
D3
D4
D8
D5
D7
D0
D9
D2
D1
D6
クラスタ
C1
C2
C3
C4
C5

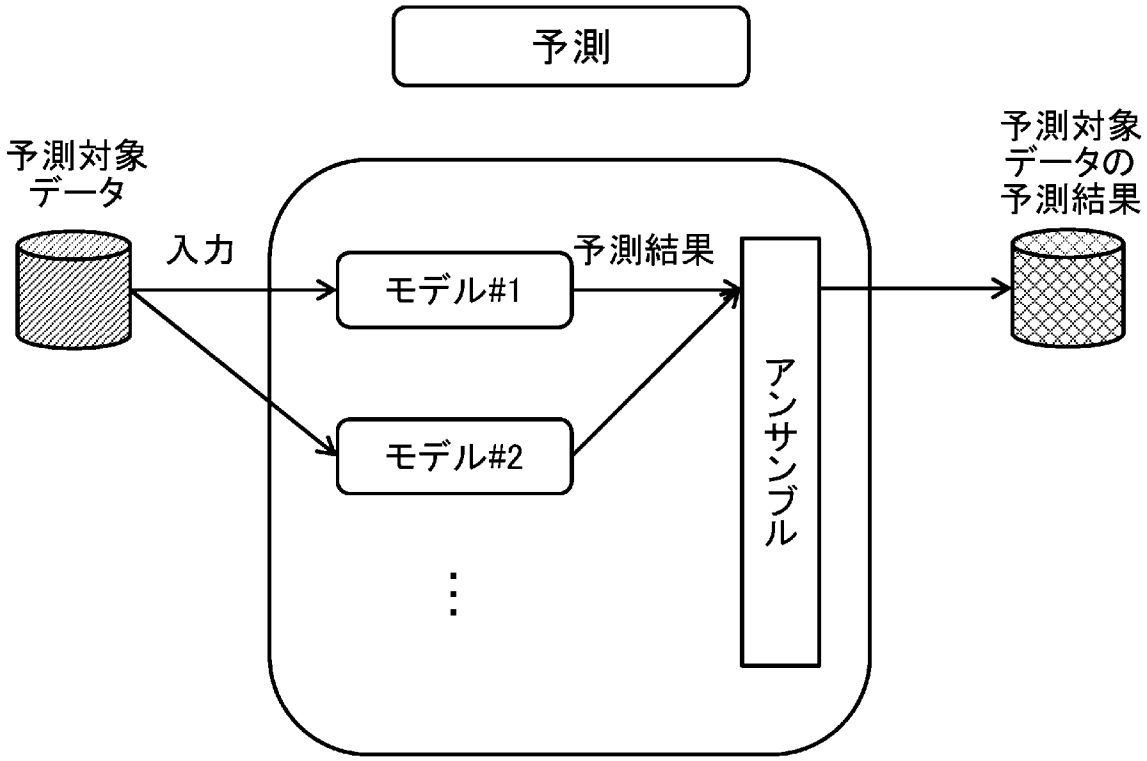
[図5]


データ作成
教師データ
クラスタリング
モデル作成用クラスタ
モデル検証用クラスタ
サンプリング
サンプリング
モデル作成用データ
モデル検証用データ

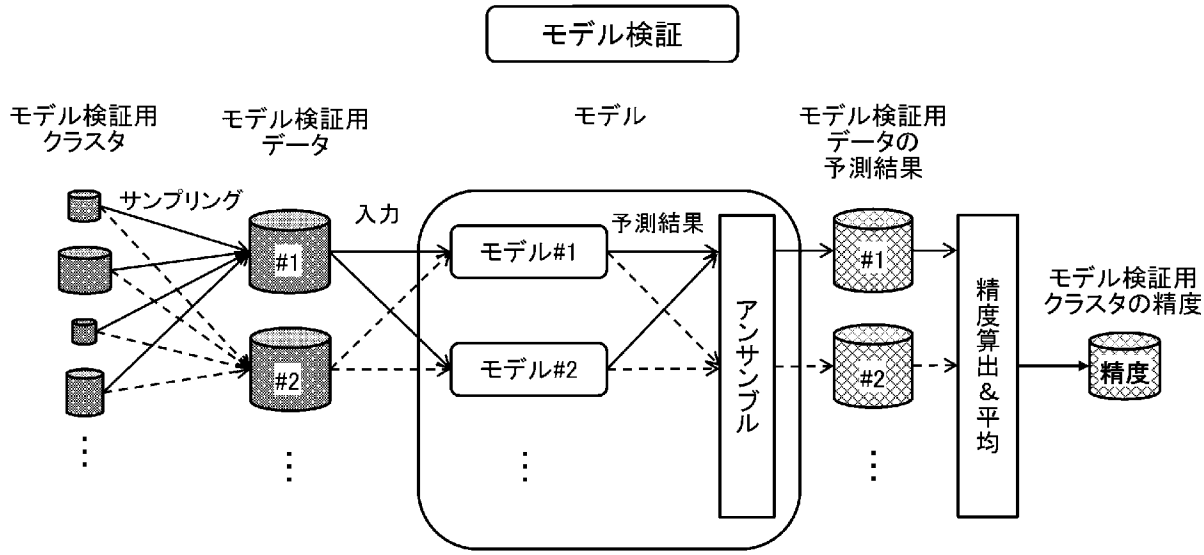
[図6]


モデル作成
モデル作成用
クラスタ
モデル作成用
データ
サンプリング
#1
#2
教師あり
学習
教師あり
学習
モデル#1
モデル#2

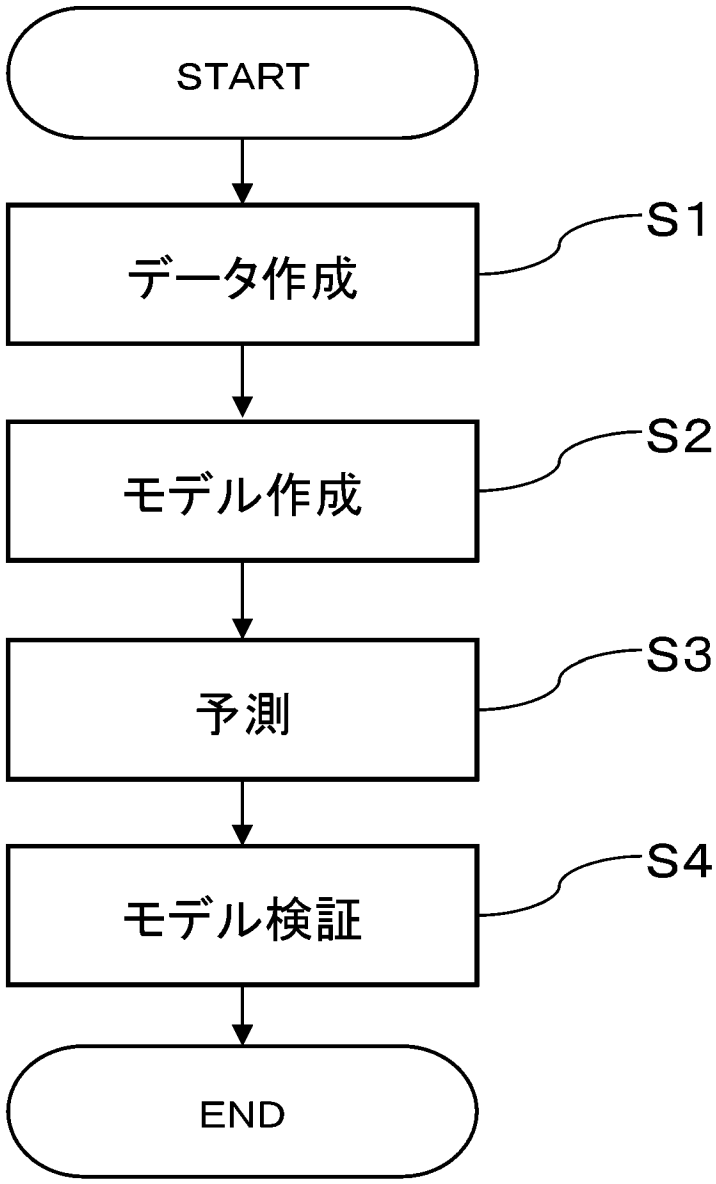
[図7]


予測
予測対象
データ
入力
モデル#1
予測結果
モデル#2
アンサンブル
予測対象
データの
予測結果

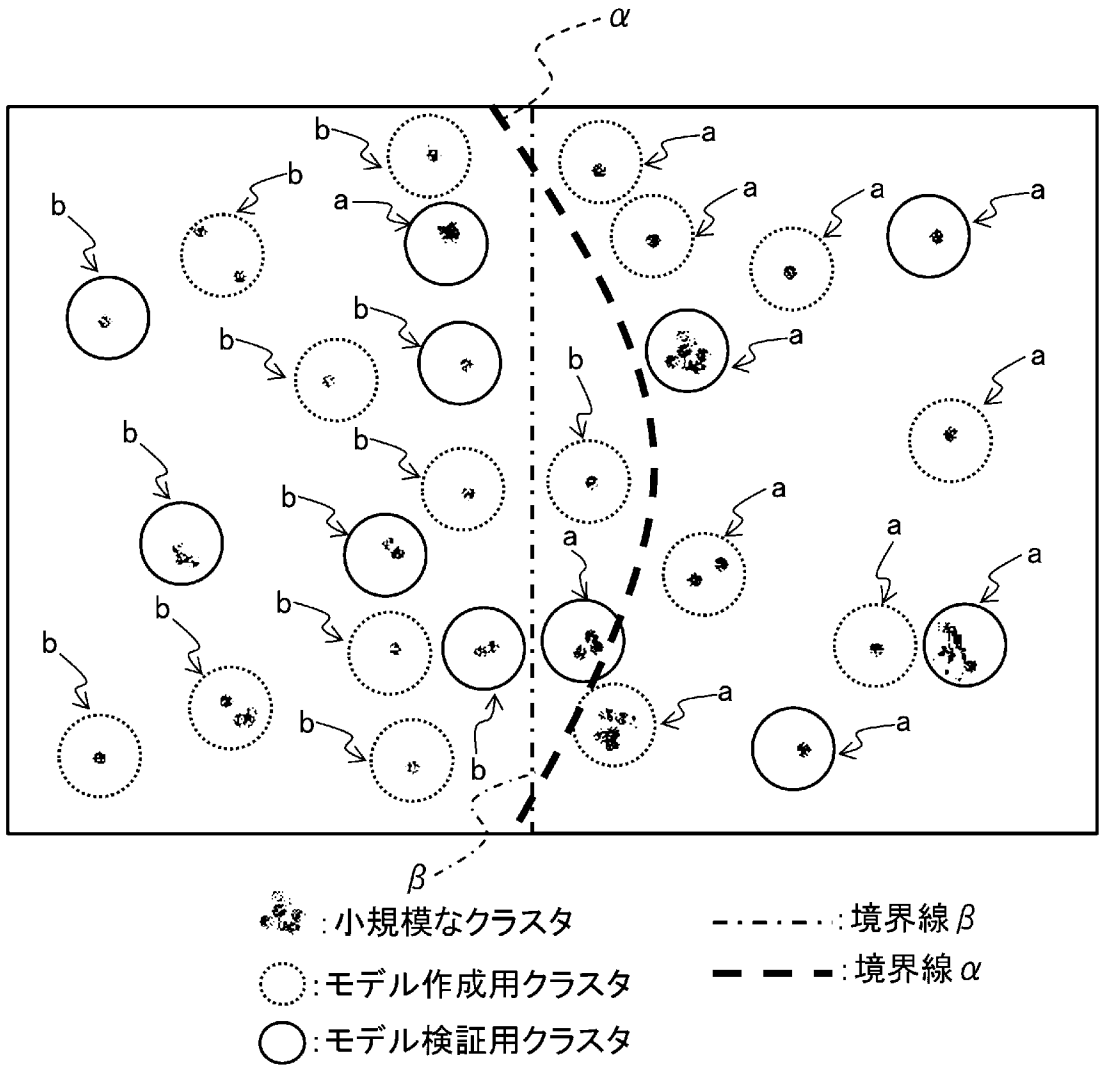
[図8]


モデル検証
モデル検証用
クラスタ
モデル検証用
データ
モデル
モデル検証用
データの
予測結果
サンプリング
入力
予測結果
#1
#2
モデル#1
モデル#2
アンサンブル
#1
#2
精度算出＆平均
モデル検証用
クラスタの精度
精度

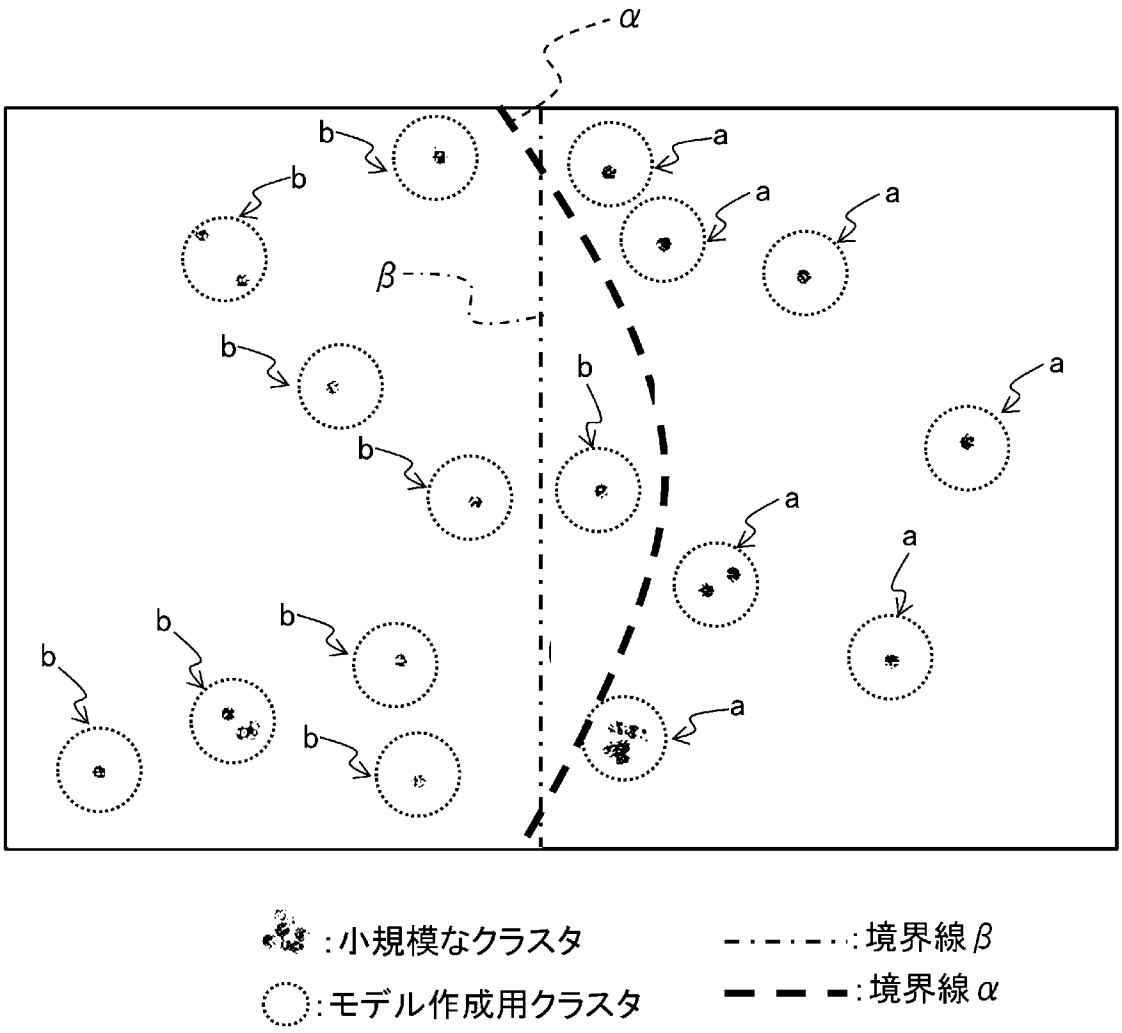
[図9]



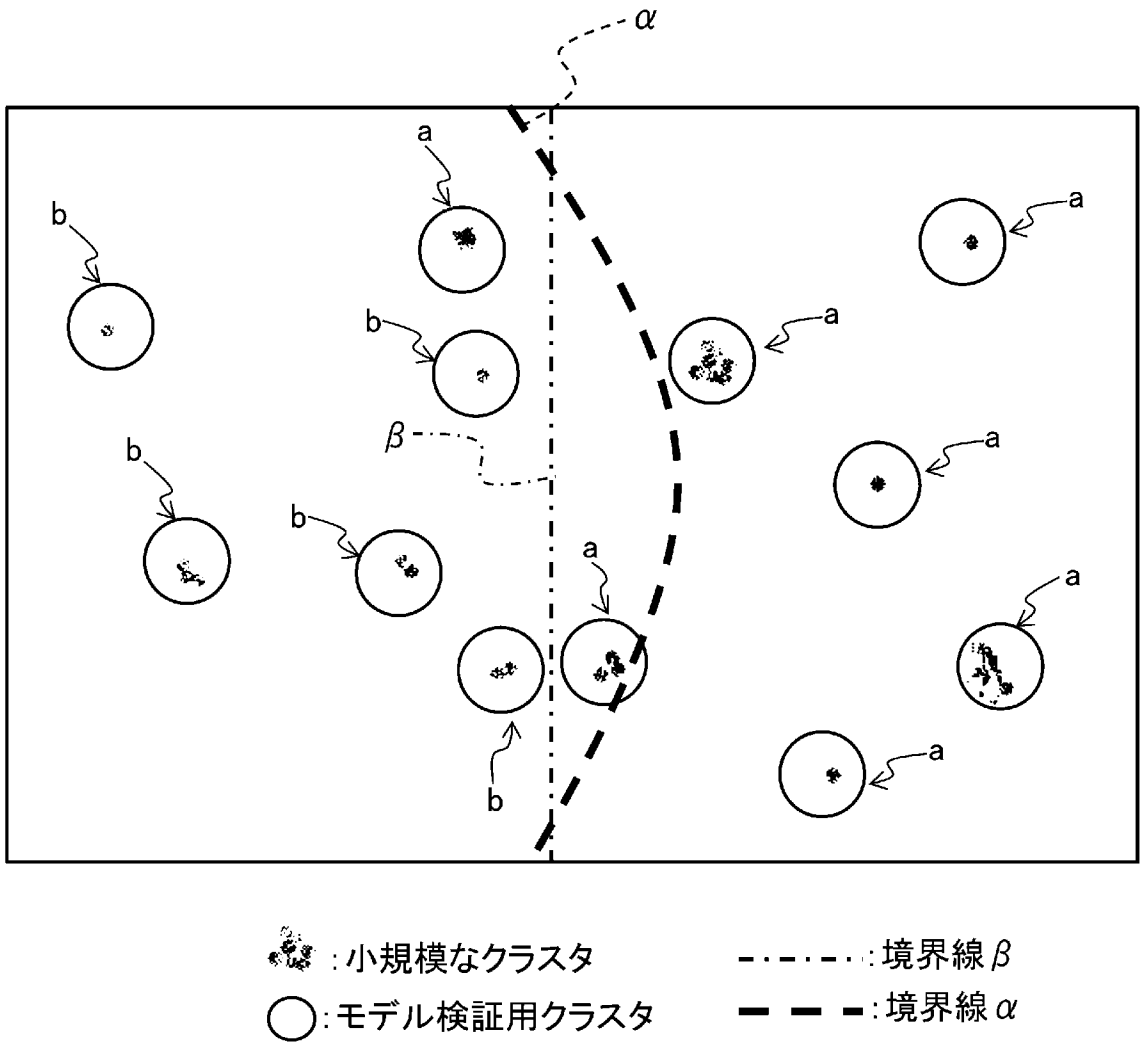
[図10]



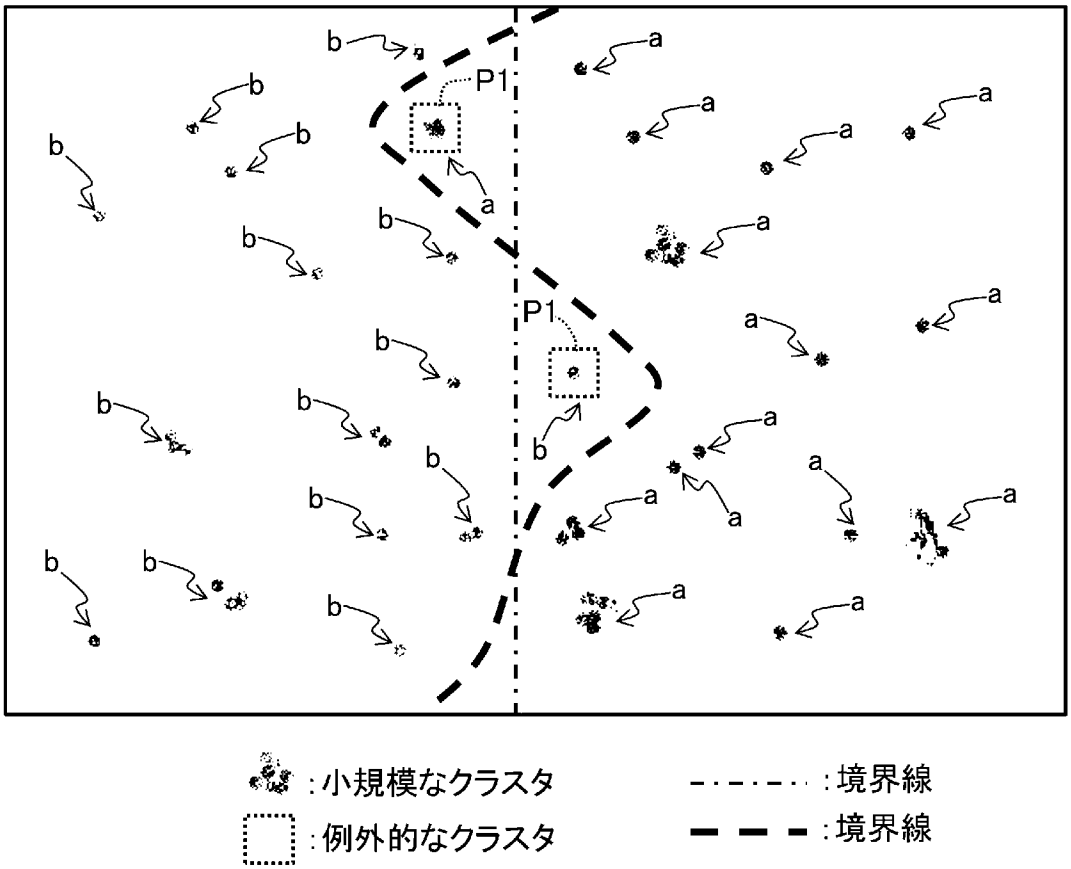
[図11]



[図12]



[図13]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/018777

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G06N20/00(2019.01)i
FI: 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)
Int. Cl. 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 1971-2020
Registered utility model specifications of Japan 1996-2020
Published registered utility model applications of Japan 1994-2020

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.
X	JP 2019-159538 A (OMRON CORP.) 19 September 2019, paragraphs [0006], [0019], [0020], [0033], [0034]	1-3, 7-9, 13-15
A		4-6, 10-12, 16-18
A	JP 2016-133895 A (CANON INC.) 25 July 2016, paragraphs [0034]-[0049], fig. 5	4-6, 10-12, 16-18
A	JP 2018-190125 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 29 November 2018, paragraphs [0026]-[0032]	4-6, 10-12, 16-18
A	JP 2019-45929 A (CANON INC.) 22 March 2019, paragraphs [0003]-[0007]	1-18

☐ 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
13.07.2020	21.07.2020
Name and mailing address of the ISA/ Japan Patent Office 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/JP2020/018777

Patent Documents referred to in the Report	Publication Date	Patent Family	Publication Date
JP 2019-159538 A	19.09.2019	(Family: none)	
JP 2016-133895 A	25.07.2016	US 2016/0210535 A1 paragraphs [0048]-[0061], fig. 5	
JP 2018-190125 A	29.11.2018	(Family: none)	
JP 2019-45929 A	22.03.2019	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2020/018777

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

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	１９７１－２０２０年
日本国実用新案登録公報	１９９６－２０２０年
日本国登録実用新案公報	１９９４－２０２０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
X	JP 2019-159538 A (オムロン株式会社) 19.09.2019 (2019-09-19) [0006],[0019]-[0020],[0033]-[0034]	1-3,7-9,13-15
A		4-6,10-12,16-18
A	JP 2016-133895 A (キヤノン株式会社) 25.07.2016 (2016-07-25) [0034]-[0049],[図5]	4-6,10-12,16-18
A	JP 2018-190125 A (日本電信電話株式会社) 29.11.2018 (2018-11-29) [0026]-[0032]	4-6,10-12,16-18
A	JP 2019-45929 A (キヤノン株式会社) 22.03.2019 (2019-03-22) [0003]-[0007]	1-18

☐ 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
13.07.2020	21.07.2020

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）
渡部　博樹 5B 3868
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/JP2020/018777

引用文献	公表日	パテントファミリー文献	公表日
JP 2019-159538 A	19.09.2019	（ファミリーなし）	
JP 2016-133895 A	25.07.2016	US 2016/0210535 A1 [0048]-[0061],Fig.5	
JP 2018-190125 A	29.11.2018	（ファミリーなし）	
JP 2019-45929 A	22.03.2019	（ファミリーなし）	