(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
2020年12月17日(17.12.2020)

WIPO | PCT

(10) 国際公開番号
WO 2020/250597 A1

(51) 国際特許分類:
G06Q 50/06 (2012.01)

(21) 国際出願番号： PCT/JP2020/018400

(22) 国際出願日： 2020年5月1日(01.05.2020)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2019-109461 2019年6月12日(12.06.2019) JP

(71) 出願人: ソニー株式会社 (**SONY CORPORATION**) [JP/JP]; 〒1080075 東京都港区港南１丁目７番１号 Tokyo (JP).

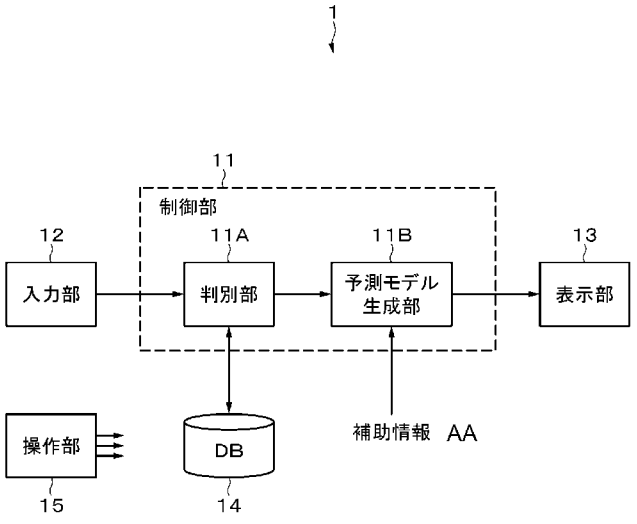
(72) 発明者: 堀口 裕士 (**HORIGUCHI, Yuji**); 〒1080075 東京都港区港南１丁目７番１号 ソニー株式会社内 Tokyo (JP). 高松 慎吾 (**TAKAMATSU, Shingo**); 〒1080075 東京都港区港南１丁目７番１号 ソニー株式会社内 Tokyo (JP). 飯田 紘士 (**IIDA, Hiroshi**); 〒1080075 東京都港区港南１丁目７番１号 ソニー株式会社内 Tokyo (JP). 中田 健人 (**NAKADA, Kento**); 〒1080075 東京都港区港南１丁目７番１号 ソニー株式会社内 Tokyo (JP). 宮原 正典(**MIYAHARA, Masanori**); 〒1080075 東京都港区港南１丁目７番１号 ソニー株式会社内 Tokyo (JP).

(74) 代理人: 杉浦 正知, 外(**SUGIURA, Masatomo** et al.); 〒1710022 東京都豊島区南池袋１－１－１１ カドラービル４０２ 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,

(54) **Title:** INFORMATION PROCESSING DEVICE, INFORMATION PROCESSING METHOD, AND PROGRAM

(54) 発明の名称： 情報処理装置、情報処理方法及びプログラム



11 Control unit
11A Determination unit
11B Prediction model generation unit
12 Input unit
13 Display unit
14 Database
15 Operation unit
AA Supplementary information

(57) **Abstract:** An information processing device comprises: an input unit through which a first dataset including a plurality of data is input; a determination unit that determines processing applied when generating a prediction model based on a second dataset which resembles the first dataset; and a prediction model generation unit that generates a prediction model based on the first dataset by applying, to the first dataset, the processing determined by the determination unit. Figure 1

WO 2020/250597 A1

［続葉有］

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, 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: 複数のデータを含む第１データセットが入力される入力部と、第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別する判別部と、第１データセットに対して、判別部により判別された処理を適用することにより第１データセットに基づく予測モデルを生成する予測モデル生成部とを有する情報処理装置である。 図１

明　細　書

発明の名称 ： 情報処理装置、情報処理方法及びプログラム

技術分野

[0001]　本開示は、情報処理装置、情報処理方法及びプログラムに関する。

背景技術

[0002]　従来から、過去のデータに基づいて各種の情報を予測する技術が提案されている。例えば、下記特許文献１には、取引対象の不動産の特徴量に応じて取引期間における当該不動産の成約確率を予測する装置が記載されている。

先行技術文献

特許文献

[0003] 特許文献1：特開２０１７－１６３２１号公報

発明の概要

発明が解決しようとする課題

[0004]　このような分野では、予測が効率的に行われることが望まれる。

[0005]　本開示は、上述した点に鑑みてなされたものであり、予測を効率的に行う情報処理装置、情報処理方法及びプログラムを提供することを目的の一つとする。

課題を解決するための手段

[0006]　本開示は、例えば、

複数のデータを含む第１データセットが入力される入力部と、

第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別する判別部と、

第１データセットに対して、判別部により判別された処理を適用することにより第１データセットに基づく予測モデルを生成する予測モデル生成部とを有する情報処理装置である。

[0007]　本開示は、例えば、

判別部が、入力部に入力された複数のデータを含む第１データセットと類

似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別し、

予測モデル生成部が、第１データセットに対して、判別部により判別された処理を適用することにより第１データセットに基づく予測モデルを生成する

を有する情報処理方法である。

[0008] 本開示は、例えば、

判別部が、入力部に入力された複数のデータを含む第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別し、

予測モデル生成部が、第１データセットに対して、判別部により判別された処理を適用することにより第１データセットに基づく予測モデルを生成する

を有する情報処理方法をコンピュータに実行させるプログラムである。

図面の簡単な説明

[0009] [図1]図１は、実施形態にかかる情報処理装置の構成例を示すブロック図である。

[図2]図２は、実施形態に係る表形式データの一例を示す図である。

[図3]図３は、実施形態に係るデータベースに記憶される情報の一例を示す図である。

[図4]図４は、所定のアルゴリズムに適用されるパラメータ及びその値の一例を示す図である。

[図5]図５は、予測モデルを作成するための新規プロジェクトを設定するための表示例を示す図である。

[図6]図６は、表形式データを選択して情報処理装置に読み込ませるための表示例を示す図である。

[図7]図７は、選択された表形式データの中で、予測モデルを生成する処理にどの特徴を用いるかを設定するための表示例を示す図である。

[図8]図８は、アルゴリズムのパラメータ等をチューニング中に表示される表示例を示す図である。

[図9]図９は、生成された予測モデルの表示例を説明するための図である。

[図10]図１０は、アルゴリズム毎の特性の一例を説明するための図である。

[図11]図１１は、優先すべき処理項目を設定できる画面の一例を説明するための図である。

[図12]図１２は、第１データセットと類似するデータセットに基づいてアルゴリズム等を検索した結果の一例を示す図である。

[図13]図１３は、補助情報をユーザに問いかける表示例を示す図である。

[図14]図１４は、補助情報をユーザに問いかける他の表示例を示す図である。

[図15]図１５は、補助情報をユーザに問いかける他の表示例を示す図である。

[図16]図１６は、補助情報をユーザに問いかける他の表示例を示す図である。

[図17]図１７は、特徴毎の有用度の表示例を示す図である。

発明を実施するための形態

[0010] 以下、本開示の一実施形態等について図面を参照した説明がなされる。なお、説明は以下の順序で行われる。

＜一実施形態において考慮すべき問題＞

＜一実施形態＞

＜変形例＞

以下に説明する実施形態等は本開示の好適な具体例であり、本開示の内容がこれらの実施形態等に限定されるものではない。

[0011] ＜一実施形態において考慮すべき問題＞

上述したように、各種の項目（売上、人口、渋滞等）を予測する予測分析技術が提案されている。予測分析技術が一般的に認知されるにつれ、統計や予測分析の専門家ではないが予測分析を適用したいデータを持つ人が多数存

在するようになった。予測分析においてより高い予測性能を出すためには種々の前処理と予測アルゴリズムとそれらに付随するハイパーパラメータを適切に選択する必要がある。アルゴリズムとハイパーパラメータを選択するには実際に予測モデルを生成して検証する必要があるが、このステップを多数行うためには大量の計算を要する。一方で、実際に予測分析を行いたいユーザとして、例えば売上予測をしたい営業の人などが挙げられるが、これらのユーザが計算資源を大量に保持しているケースは稀であり、何度も予測モデルの生成を試して予測性能の高いモデルを得ることは難しい。

[0012] 大量の計算資源はクラウドサービスを利用すれば取得可能であるが、クラウドサービスを用いた予測分析には専門的な知識を要する。また、外部のサーバにデータを持ち出す必要があり、プライバシーやセキュリティの面から適用できないケースではユーザの手元の環境で予測分析を行う必要が生じる。

[0013] 既存のパラメータチューニング手法としてベイズ最適化を用いたものが多数提案されているが、これらの手法は一般的にパラメーターつごとに数百回の探索を行って最適化を行う。デスクトップ型のパーソナルコンピュータのような環境でこれらの最適化手法に基づいて複数のパラメータのチューニングとアルゴリズムの選択を同時に行うには数千～数万回の探索が必要であり非常に長時間の計算を要する。従って、これらの計算を行うコンピュータ資源を持たないユーザは、不利である。

[0014] 予測モデルの生成を完全に自動化するには上述したように多数の探索を実行する必要がある。この分野の専門家は、経験則を用いて、探索するパラメータとアルゴリズムとの候補を絞り込むことにより短時間で予測モデルを生成する。しかしながら、専門家でない人は自身の持つデータセットに関する事前知識と予測モデルのパラメータがどのように対応しているかを知らないため、探索対象を絞り込むことが難しい。

[0015] これらの点に鑑み、以下の一実施形態では、専門的な知識や高度なコンピュータ資源を持たないユーザであっても、予測分析を効率的に行うことがで

きようにした技術に関する説明がなされる。

[0016] ＜一実施形態＞

［情報処理装置の構成例］

図１は、一実施形態にかかる情報処理装置（情報処理装置１）の構成例を示すブロック図である。情報処理装置１は、具体的には、パーソナルコンピュータ、タブレット型コンピュータ、スマートフォン、クラウド上のサーバ装置等として構成される。

[0017] 情報処理装置１は、例えば、制御部１１と、入力部１２と、表示部１３と、データベース（ＤＢ）１４と、操作部１５とを有している。制御部１１は、その機能ブロックとして、判別部１１Ａと、予測モデル生成部１１Ｂとを有している。

[0018] 制御部１１は、情報処理装置１を統括的に制御する。制御部１１は、ＣＰＵ（Central Processing Unit）等から構成されている。制御部１１は、プログラムが格納されるＲＯＭ(Read Only Memory)及びプログラムが実行される際のワークメモリとして使用されるＲＡＭ（Random Access Memory）等を有している（なお、これらの構成に関する図示は、省略されている。）。

[0019] 判別部１１Ａは、第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別する。かかる処理は、例えば、第２データセットに基づく予測モデルを生成する際に適用されたアルゴリズム及びアルゴリズムにおけるパラメータの値（以下、アルゴリズム等と適宜、称される場合がある）である。予測モデル生成部１１Ｂは、第１データセットに対して、判別部１１Ａにより判別された処理を適用することにより第１データセットに基づく予測モデルを生成する。予測モデル生成部１１Ｂには、補助情報が入力される。なお、判別部１１Ａの動作、予測モデル生成部１１Ｂの動作及び補助情報の詳細については後述される。

[0020] 入力部１２は、複数のデータを含む第１データセットが入力されるインターフェースである。入力部１２には、第２データセットも入力される。第１データセットは、現在の操作に基づいて入力部１２に入力されるデータセッ

トである。また、第２データセットは、過去に入力部１２に入力されたデータセットである。入力部１２に入力されたデータセットは、判別部１１Ａに供給される。

[0021] 表示部１３は、予測モデル生成部１１Ｂにより生成された予測モデルを表示するディスプレイ（ディスプレイを駆動するドライバを含む）である。表示部１３としては、ＬＣＤ(Liquid Crystal Display)、ＯＬＥＤ（Organic Light Emitting Diode）等を適用することができる。表示部１３は、プロジェクターによって情報を表示するものであっても良い。

[0022] データベース１４は、各種のデータを記憶するものである。データベース１４としては、ＨＤＤ（Hard Disk Drive）等の磁気記憶デバイス、半導体記憶デバイス、光記憶デバイス、光磁気記憶デバイス等が挙げられる。データベース１４は、情報処理装置１に対して着脱自在のものであっても良い。

[0023] 操作部１５は、ユーザの操作入力を受け付ける構成を総称したものである。操作部１５としては、マウス、タッチパネル、ボタン等の物理キー等を挙げることができる。操作部１５になされた操作入力に応じて操作信号が生成され、操作信号に応じた処理が行われる。

[0024] ［各種のデータについて］

（表形式データ、第１データセット及び第２データセットについて）

次に、本実施形態にかかる処理で用いられる各種のデータについての説明がなされる。始めに、表形式データに関する説明がなされる。

[0025] 図２は、表形式データの一例を示す図である。表形式データの内容は任意の内容とすることができる。図２に示す例は、商品の販売の履歴に関する内容の表形式データである。表形式データを構成する各種のデータには、その特徴としてデータの内容を示す項目（図２の第１行に規定される内容）が設定されている。表形式データは、例えば、ユーザにより指定される。表形式データは、情報処理装置１に記憶されているデータでも良いし、外部機器から情報処理装置１に取り込まれたデータであっても良い。

[0026] 第１データセットは、上述した表形式データにおける特徴の全て又は一部

が指定されたデータである。即ち、本実施形態における第１データセットは、所定のデータの一例である表形式データに対するユーザ入力に応じて、その内容が設定されるデータセットである。指定された特徴に対応する第１データセットが、予測モデル生成部１１Ｂが予測モデルを生成する際に用いられる。即ち、表形式データの全てが第１データセットとなる場合もあれば、表形式データの一部が第１データセットとなる場合もある。

[0027] 第２データセットは、予測モデル生成部１１Ｂが過去に予測モデルを生成した際に使用されたデータセットの中で、第１データセットと類似するデータセットである。詳細は後述するが、第１データセット及び第２データセットには、それぞれを特徴付ける指標が割り当てられる。かかる指標を対比することにより、第１データセットと類似する第２データセットを判別することができる。

[0028] （データベースに記憶される情報）

図３は、データベース１４に記憶される情報（以下、データベース情報と適宜、称される）の一例を示す図である。データベース情報の項目には、例えば、モデル名、表形式データファイル名、データセット情報、データセットを構成する各特徴の情報、予測モデル生成時間、予測モデル使用メモリ、アルゴリズムで使用された各パラメータの実験結果、予測モデル生成条件が設定されている。

[0029] モデル名は、予測モデルを生成したときに設定された名称である。モデル名は、予測モデルの内容に応じて適宜、設定することができる。図３では、ある予測モデルのモデル名として「Ａローンの貸し倒れ予測モデル」と、他の予測モデルのモデル名として「Ｂ店の廃棄量予測モデル」が設定された例が示されている。

[0030] 表形式データファイル名は、予測モデルを生成した際に使用された第２データセットの基になる表形式データ及びそのファイル名である。

[0031] データセットの情報とは、過去に生成された予測モデル対応する第２データセットに関する各種の情報である。データセットの情報は、例えば、デー

タセットに含まれるデータ数、特徴数、欠損したデータの割合、ファイルサイズ、ドメイン（気象データ、売上データ等、何についてのデータなのかを示す情報）、問題設定（分類、回帰、時系列予測等）等を示す情報である。

[0032] 各特徴の情報とは、予測モデルを生成する際にデータセットに対して適用されたアルゴリズム、各特徴の名前、ユニークなデータの数、各特徴のデータ型（テキスト、数値、日付、カテゴリ変数等）、その他特徴を説明するための統計量（平均、分散、尖度等）を示す情報である。これらの情報は、公知の方法により定量化（数値化）できる。例えば、各特徴のデータ型として「テキストデータ」がある場合には、データ型として「テキストデータ」を示す識別子が割り当てられる。そして、「テキストデータ」には、統計量の例として「スペースや区切りの文字数」、「文章の長さの平均」、「言語の種類」等が対応付けられる。また、日付などの「タイムスタンプデータ」の場合には、データ型として「タイムスタンプデータ」を示す識別子が割り当てられる。そして、統計量の例として「時間帯の平均」、「データに含まれる期間」、「タイムスタンプデータのフォーマット」等が対応付けられる。

[0033] 予測モデル生成時間は、予測モデルを生成するのに要した時間である。予測モデル使用メモリは、予測モデルを生成するのに必要としたメモリの容量である。

[0034] アルゴリズムで使用された各パラメータの実験結果は、適用されたアルゴリズムのパラメータの履歴、及び、そのパラメータで予測モデルを生成した際の結果を示す情報である。本項目には、パラメータ設定名が記述されている。パラメータ設定名には、図４に示すように、予測モデルを予測する際に使用されるアルゴリズム名と具体的なパラメータの値が対応付けられている。なお、アルゴリズムが変更されて予測モデルが生成される場合もあれば、同じアルゴリズムであってもパラメータの値が変更されて予測モデルが生成される場合もある。かかる場合には、その全てが履歴として記述される。

[0035] 図３では、例えば、モデル名「Ａローンの貸し倒れ予測モデル」の予測モデルを生成する際に、アルゴリズムとして「分類用決定木」が使用され、そ

のパラメータとして「決定木モデルパラメータＡ」に対応するパラメータ及びその値が使用されたことが示されている。更に図３では、当該パラメータ及びパラメータの値を使用して予測モデルを生成した結果、精度として「0.82」、再現率として「0.6」、Ｆ値として「0.2」であったことが示されている。

[0036] 予測モデル生成条件は、予測モデルを生成する際に優先すべき処理項目を示す条件である。係る処理項目は、ユーザの操作入力によって設定される。処理項目とは、例えば「性能優先」、「速度優先」、「メモリ優先」の何れかである。「性能優先」とは、予測モデルの精度を優先する設定である。「速度優先」は、予測モデルを生成する速度を優先する設定である。「メモリ優先」とは、予測モデルを生成する際に使用されるメモリの容量ができるだけ小さいものを優先する設定である。

[0037] 予測メモリ生成条件には、ユーザによって回答された補助情報の内容が含まれる。補助情報とは、第１データセットに基づいて予測モデルを効率的に生成するための情報である。具体的には、補助情報は、第１データセットに含まれる時系列なデータのうち予測モデルの生成に使用されるデータの期間、第１データセットに含まれるテキストデータのうち予測モデルの生成に使用されるテキストデータの指定、及び、第１データセットに含まれる所定のデータの正確性に関する情報の少なくとも一つである。情報処理装置１は、情報処理装置１からユーザに対してなされる質問に対して入力されたユーザの回答に基づいて補助情報を取得する。

[0038] 以上がデータベース情報の一例である。なお、上述したデータベース情報の項目の区別は便宜上のものであり、適宜、変更可能である。

[0039] ［情報処理装置の動作例］

（動作例Ａ１）

続いて、情報処理装置１の複数の動作例に関する説明がなされる。始めに、情報処理装置１の動作例Ａ１に関する説明がなされる。なお、特に断らない限り、以下に説明する情報処理装置１の動作（他の動作例を含む）は、制

御部１１の制御に従って行われる。

[0040]「手順Ｂ１」

始めに、ユーザが情報処理装置１の操作部１５を使用して、予測モデルを生成するためのプロジェクトを開始し、予測モデルの生成に使用する表形式データを選択して情報処理装置１に読み込ませる。そして、ユーザは、表形式データにおけるどの特徴を、予測モデルを生成する処理に使用するかを指定する。かかる指定により、読み込まれた表形式データに基づく第１データセットが生成される。かかる処理は、以下の説明において「手順Ｂ１」と適宜、称される。

[0041] 図５は、予測モデルを生成するための新規プロジェクトを設定するための表示例を示す図である。図５に示す表示例は、例えば、情報処理装置１の表示部１３に表示される。表示部１３には、プロジェクト名を入力可能な矩形状の表示枠１０１と、適宜な説明やメモを入力可能な矩形状の表示枠１０２と、キャンセルボタン１０３と、決定ボタン１０４とが表示される。ユーザは、操作部１５を使用して各表示の箇所に情報を入力する。

[0042] 具体的には、ユーザは、適宜なプロジェクト名（図示の例では「顧客データからの売り上げ予測」）を表示枠１０１の枠内に入力する。また、ユーザは、必要に応じて適宜な説明（図示の例では「2000年11月から2013年12月のデータを用いて次期の売り上げ予測を検証」）を表示枠１０２の枠内に操作部１５を使用して入力する。

[0043] 図６は、表形式データを選択して情報処理装置１に読み込ませるための表示例を示す図である。ユーザは、操作部１５を使用して表形式データを選択する。表示部１３には、選択された表形式データの格納元のアドレス情報１０５が表示される。これまでになされたプロジェクト名の入力、プロジェクト名に付随する説明の入力、及び、表形式データの選択を完了させる場合は、ユーザは、決定ボタン１０４をクリックする。プロジェクト名を修正する場合等は、ユーザは、キャンセルボタン１０３をクリックして再度の入力を行う。

[0044] 決定ボタン１０４が押されると、表示部１３の表示内容が図７に示す表示内容に遷移する。図７は、選択された表形式データの中で、予測モデルを生成する処理にどの特徴（本例では、表形式データにおける項目）を用いるかを設定するための画面例を示す図である。図７に示すように、表示部１３には、表形式データにおける各項目の名称である項目名１０７が表示される。各項目の左側にはチェックボックス１０８が表示される。ユーザは、例えば、予測モデルを生成するのに用いる特徴に対応するチェックボックスにチェックを付し、予測モデルを生成するのに用いない特徴に対応するチェックボックスからチェックを外す。なお、少なくとも一つのチェックボックスにチェックが付されていれば良く、全てのチェックボックスにチェックが付されていても良い。また、図７では、特徴毎のデータ形式１０９が設定可能とされている。また、図７に示す画面を使用して、予測モデルの結果である予測タイプ１１０（出力形式であり、例えば、二値分類、多値分類、数値分類）を設定することも可能とされている。

[0045] 各特徴に関する設定を完了させる場合には、ユーザにより決定ボタン１０４がクリックされる。これにより、表形式データに基づく第１データセットの作成が完了する。

[0046] 「手順Ｂ２」

第１データセットの作成が完了すると、第１データセットに対して「データセットの情報」及び「各特徴の情報」（図２参照）を求める演算がなされる。判別部１１Ａは、演算結果に基づいて、データベース１４に記憶されている複数の第２データセットの中から第１データセットに類似する第２データセットを検索して判別する。判別部１１Ａは、例えば、データセットの情報が第１データセットと同じ、若しくは、第１データセットと第２データセットとの各情報同士の差分値を積算した値が一定以下であるデータセットを、第１データセットに類似する第２データセットと判別する。また、判別部１１Ａは、各特徴の情報を参照し、類似する特徴を多く有しているデータセットを、第１データセットに類似する第２データセットと判別しても良いし

、これらを組み合わせた方法によって第１データセットに類似する第２データセットを判別しても良い。本例では、第１データセットに類似するデータセットとして１個の第２データセットが判別部１１Ａにより判別される。

[0047] 「手順Ｂ３」

手順３では、手順Ｂ２で判別された第２データセットに適用されたアルゴリズム等が判別部１１Ａにより判別される。判別部１１Ａは、データベース情報を参照することにより、第２データセットに適用されたアルゴリズム等を取得する。そして、第２データセットに適用されたアルゴリズム等となるように、各種の設定がチューニングされる。チューニング中に表示される画面の一例が、図８により示されている。

[0048] 「手順Ｂ４」

手順Ｂ３で各種の設定に関するチューニングが終了すると、予測モデル生成部１１Ｂが、第１データセットに対して、チューニングされたアルゴリズム等を適用することにより予測モデルを生成する。そして、生成された予測モデルが表示部１３に表示される。

[0049] 図９は、生成された予測モデルの表示例を示す図である。表示部１３には、売り上げの予測を示すグラフ１１３が表示される。また、ユーザによって設定された予測タイプの情報１１１（図示の例では、数値分類）が表示される。また、予測モデルの精度に関する情報１１２が表示される。なお、予測モデルを生成する処理の内容（アルゴリズム等や予測モデルの精度等）は、新たなデータベース情報としてデータベース１４に記憶される。

[0050] 以上が、情報処理装置１の動作例Ａ１で行われる処理の内容である。上述したように、予測モデルを生成する際に設定された第１データセットと類似する第２データセットを検索し、検索された第２データセットに適用されたアルゴリズム等を第１データセットに適用する。これにより、第１データセットに基づく予測モデルを生成する際に効果的なアルゴリズム等を一から探索する必要がなくなる。これにより、第１データセットに基づく予測モデルを効率的に生成することができる。また、ユーザは、表形式データに基づい

て第１データセットを設定するだけで良いので、専門的な知識やスキルを持たないユーザであっても所望の予測モデルを生成することが可能となる。

[0051] なお、手順Ｂ２において、第１データセットと類似する複数の第２データセットが判別されても良い。例えば、第１データセットと一定以上の類似度を有する複数の第２データセットが判別部１１Ａにより判別されても良い。例えば、第１データセットと一定以上の類似度を有する１００個の第２データセットが検索されたとする。その中で、最も多くの第２データセットに対して適用されたアルゴリズム等が手順Ｂ４で適用されても良い。また、第１データセットと一定以上の類似度を有する１０個程度の第２データセットを検索し、それぞれのデータセットに対して適用されたアルゴリズム等を順次、第１データセットに適用しても良い。そして、その結果、生成された予測モデル（１０個の予測モデル）が順次、表示部１３に表示されるようにしても良い。

[0052] また、所定の基準に従って、第１データセットに複数のアルゴリズム等を適用して検証するようにしても良い。例えば、図１０に示すように、アルゴリズム毎の特徴（例えば、性能への影響の平均、性能の分散、データベースの記録数（アルゴリズムの適用数）等がデータベース１４に記録されていても良い。例えば、平均的にプラスとなるアルゴリズムを優先して検証する、という基準が設定されている場合には、参照符号Ｃ１で囲まれた箇所の性能がプラス方向に最も大きいことから、参照符号Ｃ１に対応するアルゴリズム（欠損値を削る）を優先した検証が行われる。また、例えば、分散が大きいアルゴリズムを優先して検証する、という基準が設定されている場合には、参照符号Ｃ２で囲まれた箇所の分散が最も大きいことから、参照符号Ｃ２に対応するアルゴリズム（三角関数で変換）を優先した検証が行われる。また、例えば、Upper Confidence bound（探索回数が少なく、性能がプラスになる確信が得られていないもの）、という基準が設定されている場合には、参照符号Ｃ３で囲まれた箇所の性能がプラス方向になるものの適用数であるデータベースの記録数が最も小さいことから、参照符号Ｃ３に対応するアルゴ

リズム（20の区間に区切る）を優先した検証が行われる。基準の内容は予め決められていても良いし、ユーザによって設定されても良い。

[0053]（動作例Ａ２）

続いて、動作例Ａ２に関する説明がなされる。なお、動作例Ａ１で説明された処理や表示例と同一又は同質の処理や表示例については同一の参照符号が付され、重複した説明が適宜、省略される。動作例Ａ２は、ユーザにより設定された優先すべき処理項目（例えば、「速度優先」「性能優先」等）に基づいてアルゴリズム等が選択され、選択されたアルゴリズム等に基づいて予測モデルが生成される動作である。

[0054]「手順Ｂ２１」

手順Ｂ２１では、基本的に手順Ｂ１と同様の処理が行われる。手順Ｂ２１では、優先すべき処理項目についても設定可能な点が手順Ｂ１と異なる。図１１は、優先すべき処理項目を設定できる画面の一例を示す図である。図１１に示す表示例では、図７に示した画面の内容に加え、優先すべき処理項目を設定できる処理項目設定表示１２１が表示されている。

[0055] 処理項目設定表示１２１は、例えば、半円状のインジケータによって表示される。インジケータの左端が速度優先に対応し、インジケータの右側が性能優先に対応している。インジケータの針を適宜な箇所に設定することにより、速度又は性能のどちらをどの程度優先できるかが設定可能とされている。具体例としては、処理項目設定表示１２１におけるインジケータの針が左端に設定された場合には「完全に速度優先」という内容の処理項目が設定される。また、インジケータの針が中央と左端との間に設定された場合には「やや速度優先」という内容の処理項目が設定される。また、処理項目設定表示１２１におけるインジケータの針が右端に設定された場合には「完全に性能優先」という内容の処理項目が設定される。処理項目設定表示１２１におけるインジケータの針が中央と右端との間に設定された場合には「やや性能優先」という内容の処理項目が設定される。

[0056]「手順Ｂ２２」

手順Ｂ２２では、基本的に手順Ｂ２及び手順Ｂ３と同様の処理が行われる。概略的には、第１データセットと類似するデータセットが選択される。そして、選択されたデータセットの中から、更に、ユーザによって設定された優先すべき処理項目に対応するデータセットが選択され、選択されたデータセットが第２データセットとして設定される。

[0057] 処理項目設定表示１２１で「完全に速度優先」が設定された場合には、第１データセットと類似するデータセットの中から、例えば処理時間が少ない速度（図３における予測モデル生成時間）上位１％のデータセットが選択され、選択されたデータセットが第２データセットして設定される。そして、例えば、設定された第２データセットの中で最も多く使用されているアルゴリズム等が第１データセットに適用されるアルゴリズム等として設定される。設定された第２データセットのそれぞれに適用されたアルゴリズム等の全てが第１データセットに適用され検証が行われても良い。処理項目設定表示１２１で「やや速度優先」又は「やや性能優先」が設定された場合には、第１データセットと類似するデータセットの中から、例えば速度上位１０％に入り、且つ、性能（図３における精度）上位１０％に入るデータセットが選択され、選択されたデータセットが第２データセットして設定される。そして、設定された第２データセットの中で最も多く使用されているアルゴリズム等が第１データセットに適用されるアルゴリズム等として設定される。処理項目設定表示１２１で「完全に性能優先」が設定された場合には、第１データセットと類似するデータセットの中から、性能が高い上位１％のデータセットが選択され、選択されたデータセットが第２データセットして設定される。そして、設定された第２データセットの中で最も多く使用されているアルゴリズム等が第１データセットに適用されるアルゴリズム等として設定される。図１２には、第１データセットと類似するデータセットに基づいてアルゴリズム等を検索した結果の一例を示す図である。

[0058] 「手順Ｂ２３」

手順Ｂ２３では手順Ｂ３と同様の処理が行われる。概略的には、予測モデ

ル生成部１１Ｂが、第１データセットに対して、チューニングされたアルゴリズム等を適用することにより予測モデルを生成する。そして、生成された予測モデルが表示部１３に表示される。

[0059] 本例によれば、ユーザによって設定された優先すべき処理項目に基づいて、予測モデルを生成することができる。なお、速度優先や性能優先の他にもメモリ優先等にかかる設定ができるようにされても良く、優先すべき処理項目の内容や個数に応じて処理項目設定表示１２１の表示態様は適宜、変更され得る。

[0060] （動作例Ａ３）

続いて、動作例Ａ３に関する説明がなされる。なお、動作例Ａ１、Ａ２で説明された処理や表示例と同一又は同質の処理や表示例については同一の参照符号が付され、重複した説明が適宜、省略される。

[0061] 本例では、情報処理装置１を使用して、ある店舗の毎時間の利用者データから翌週の売り上げを予測する予測モデルを生成する例を想定している。通常、ある時点での売り上げ予測を行う場合には「直前ｘ週間の累積売り上げ」や「昨年の同じ時期の売り上げ」のような情報に基づいて予測を行うことが効果的である場合が多い。しかし「一週間前」「二週間前」…「一年前」等の期間をすべて検証してどれが効果的であるかを判断することは非効率である。そこで、本例では、過去のデータベース情報からは絞りこめない情報（本例の場合、どの期間の累積データを特徴に加えると予測に効果があるか）をユーザに質問するダイアログを出し、ユーザから処理に必要なヒントとしての補助情報を受け取る。補助情報に基づく処理が第１データセットに適用されることにより予測モデルが生成される。

[0062] 「手順Ｂ３１」

手順Ｂ３１では、手順Ｂ１及び手順Ｂ２と同一の処理が行われる。

[0063] 「手順Ｂ３２」

手順Ｂ３２では、ユーザに対して補助情報を問いかける報知がなされる。図１３は、補助情報をユーザに問いかける表示例を示す図である。表示部１

３には、例えば「売り上げ予測に有効と思われる期間はいつですか？」という質問１３１が表示される。また、表示部１３には、質問に対する回答候補１３２が表示される。また、表示部１３には、回答内容をキャンセルするためのキャンセルボタン１３３が表示される。図示の例では、３個の回答候補１３２が表示されている。なお、ユーザが質問に回答している間も、バックグラウンドでは、売り上げの期間が適宜、変更されて予測モデルのパラメータのチューニングが続行される。

[0064] 「手順Ｂ３３」

質問に対して「予測したいタイミングの直前１ヶ月の累積売り上げ」が売り上げ予測に有効である、というユーザの回答の補助情報を予測モデル生成部１１Ｂが得たとする。予測モデル生成部１１Ｂは、補助情報に基づく処理を適用する。例えば、第１データセットの特徴（例えば、売り上げ）に「直前１ヶ月」という特徴を追加する。これにより、全ての売り上げのデータが直前１ヶ月のデータに絞り込まれる。なお、追加された特徴に基づいて、第１データセットと類似するデータセットが再度検索され、検索結果に基づいて、第２データセットが再設定されても良い。

[0065] 「手順Ｂ３４」

手順Ｂ３４では、手順Ｂ４と同様の処理が行われる。特徴が追加された第１データセットに対して所定のアルゴリズム等が予測モデル生成部１１Ｂにより適用されることで、予測モデルが生成される。生成された予測モデルが表示される。

[0066] 本例によれば、予測分析に有効若しくは予測分析を効率的に行うための情報である補助情報を得ることができるようにしているので、一層、効率的な予測分析を行うことが可能となる。

[0067] （動作例Ａ４）

続いて、動作例Ａ４に関する説明がなされる。なお、動作例Ａ１～Ａ３で説明された処理や表示例と同一又は同質の処理や表示例については同一の参照符号が付され、重複した説明が適宜、省略される。本例では、補助情報の

内容が上述した動作例Ａ３と異なっている。

[0068] 本例では、具体例として、商品のレビューの文章からユーザの満足度を予測する例を想定している。従って、第１データセットは、少なくともテキストデータを含む。テキストデータの場合、例えば、予測に必要ない単語（例えば、「です」、「ます」等）をデータから除外する処理を行う前処理が考えられる。これらの処理も繰り返し予測モデルを生成しながら予測に寄与する度合いを見ることで自動的に行うことが可能となり得るが、非常に長い時間を要するため効率的ではない。このような場合、情報処理装置１が、ユーザからヒントとしての補助情報を受け取ることでこれらの検証を行うための時間を削減することができる。

[0069]「手順Ｂ４１」

手順Ｂ４１では、手順Ｂ１及び手順Ｂ２と同一の処理が行われる。

[0070]「手順Ｂ４２」

手順Ｂ４２では、補助情報を問いかける表示が表示部１３に表示される。例えば、図１４に示すように、第１データセットに含まれ、一定数以上検索された複数の単語（単語群１４１）が表示部１３に表示される。単語群１４１の各単語にはチェックボックスが表示されており、例えば予測に不必要な単語にチェックを付すことにより、当該単語は予測分析に不要な単語として設定される。例えば、図１４に示す例では、「です」及び「ます」という単語が予測に不要な単語として設定されている。また、表示部１３には、設定内容をキャンセルするためのキャンセルボタン１４１Ａが表示される。

[0071]「手順Ｂ４３」

手順Ｂ４３では、手順Ｂ４と同様の処理が行われる。また、予測モデル生成部１１Ｂが予測モデルを生成する際には補助情報に基づく処理が適用される。具体的には、テキストデータから「です」、「ます」が除外された第１データセットに対して所定のアルゴリズム等が適用されることにより、予測モデルが生成される。生成された予測モデルが表示される。

[0072] なお、補助情報は、上述したデータの期間や予測に不要な単語に関する情

報に限られない。補助情報は、例えば、表記ゆれによって同一の物を指しているのに異なる単語として扱われてしまう単語を名寄せする情報であっても良い。図１５は、かかる補助情報をユーザに問いかける表示例を示す図である。図１５に示す例では、補助情報を得るための質問として「以下の単語の中で「東京」と同じ単語はありますか？」との質問１４２が表示される。そして、質問１４２の下側に例えば４個の単語（「Ｔｏｋｙｏ」、「東京都」、「ＴＯＫＩＯ」、「ＴＯＫＹＯＵ」）を含む単語群１４３が表示される。単語群１４３の個々の単語の横にはチェックボックスが表示されている。また、表示部１３には、設定内容をキャンセルするためのキャンセルボタン１４３Ａが表示される。ユーザは、例えば、「東京」と同じ単語にチェックをつける。そして、予測モデル生成部１１Ｂは、予測モデルを生成する際には、「Ｔｏｋｙｏ」及び「東京都」との単語が「東京」と同じ単語して扱われるようにして予測モデルを生成する。

[0073] 補助情報は、はずれ値であるか否か、換言すれば、第１データセットに含まれるデータの正確性がユーザによって確認された情報であっても良い。例えば、売り上げや在庫数は、通常、正の値である。しかしながら、第１データセットの特徴、具体的には、売り上げや在庫数に対応するデータに負の値がある場合には、当該データは異常なデータである虞が高い。一方で、当該データが異常であるかを検証する処理がなされると予測分析が非効率となる。そこで、他のデータと比較して異質なデータが、異常なデータであるか否かがユーザに確認される。図１６は、かかる補助情報をユーザに問いかける表示例を示す図である。図１６に示す例では、例えば「以下のデータは正常なデータですか？」という質問１４４が表示される。そして、正常ではないと思われる具体的なデータの内容１４５（図示の例では「店名：渋谷店、売り上げ：－１、在庫数：－１」）が表示される。また、図１６には、正常なデータと思われる他のデータの内容１４６（図示の例では「店名：東京店、売り上げ：１２００万円、在庫数：２００」）が表示されることにより、正常と思われるデータと異常と思われるデータとをユーザが対比することがで

きるようにされている。ユーザは、表示されたデータが異常である場合には「取り除く」と表示されたボタン１４７Ａをクリックすることにより補助情報を入力する。この場合には、渋谷店の売り上げ及び在庫数に関するデータは、予測モデルが生成される際に用いられる第１データセットから除外される。ユーザは、表示されたデータを予測モデルを生成する処理に使用する場合には「使用する」と表示されたボタン１４７Ｂをクリックすることにより補助情報を入力する。この場合には、渋谷店の売り上げ及び在庫数に関するデータは、予測モデルが生成される際に用いられる第１データセットから除外されずに使用される。

[0074] 本例によれば、予測分析に有効若しくは予測分析を効率的に行うための情報である補助情報を得ることができるようにしているので、一層、効率的な予測分析を行うことが可能となる。

[0075] （動作例Ａ５）

本例は、予測モデルを生成した結果を確認したユーザにヒントを求める例である。具体的には、情報処理装置１が手入力された売り上げデータに基づいて需要予測を行うことで予測モデルを生成したものの、予測モデルの性能があまり良くない場合にユーザからのフィードバックを受け付ける処理が想定される。そして、フィードバックに基づいて、アルゴリズム等が再設定される。

[0076] 「手順Ｂ５１」

手順Ｂ５１では、手順Ｂ１～手順Ｂ４が行われ、予測モデルが生成される。そして、手順Ｂ５１では、情報処理装置１が、ユーザが予測分析に利用すると設定した各特徴について、第１データセットに基づく予測モデルの生成時にどれだけ役に立ったかを示す有用度を判定する。例えば、情報処理装置１の制御部１１は、予測モデルを生成する演算に、特徴に対応するデータをどれだけ用いたかに基づいて、特徴毎の有用度を判定する。勿論、他の公知の方法により特徴毎の有用度が判定されても良い。

[0077] 判定された特徴毎の有用度が表示部１３に表示される。図１７は、特徴毎

の有用度の表示例を示す図である。特徴である項目名１５１が表示されるとともに、各項目名の右側に有用度１５２が表示されている。有用度１５２は、例えば、矩形状の枠で表示され、枠内の黒い箇所が大きいほど有用度１５２が高いことが示されている。勿論、有用度１５２の表示態様は適宜、変更することができる。例えば、具体的なスコアによって有用度１５２が表示されても良い。また、表示部１３には、有用度が所定以下である特徴に関するコメント１５３が表示される。図１７に示す例では、特徴の一つである「購入金額」に関する有用度が著しく低い。そこで、コメント１５３としては、例えば「「購入金額（円）」が予測の際にほとんど利用されませんでした」という内容のコメントが表示される。また、表示部１３には、有用度が低い特徴である「購入金額（円）」に関する現在の認識結果１５４が表示される。

[0078] 「手順Ｂ５２」

手順Ｂ５２では、ユーザが、表示された有用度１５２を確認する。ユーザは、有用度１５２に基づいて、売り上げに関係していると想定していた「購入金額(円)」のデータが予測モデルを生成する上で役に立っていない（有用度が低い）ことを認識する。また、ユーザは、認識結果１５４に基づいて、「購入金額（円）」にカンマ・円・￥などの記号が混在しているため、「購入金額（円）」が数値データではなく文字列として処理されていることを認識する。ユーザは、かかる認識に基づいて、「購入金額(円)」のデータ形式を数値データに設定する（図７参照）。そして、ユーザは、ボタン１５５をクリックする。

[0079] 「手順Ｂ５３」

ボタン１５５がクリックされると、「購入金額（円）」が数値データとして扱われた上で、上述した手順Ｂ２～Ｂ４の処理が行われる。そして、再度、予測モデル生成部１１Ｂによる予測モデルが生成され、生成された予測モデルが表示部１３に表示される。

[0080] なお、手順Ｂ５２において、有用度１５２が低い場合でも予測モデルを修

正する必要がない場合もあり得る。かかる場合には、ユーザは、表示部１３に表示された、「正しい」とのボタン１５６をクリックすれば良い。

[0081] 本例によれば、予測モデルを生成する上での設定ミスをユーザが気づきやすくすることができる。そして、ユーザからのフィードバックがなされることにより、正確な予測モデルを生成することができる。

[0082] 以上説明した本実施形態によれば、繰り返し予測モデルを生成するツール上や、類似するデータセットで繰り返し予測モデルの性能の検証を行う環境において短時間で性能が高い予測モデルの生成ができるようになる。また、アルゴリズム等を探索中にユーザが質問に答えることで、より短時間に予測モデルの生成を行うことが可能になる。また、ユーザの性能優先・速度優先などの設定に沿った予測モデルを過去に適用されたアルゴリズム等の履歴を用いてより高速に生成することができる。

[0083] ＜変形例＞

以上、本開示の一実施形態について具体的に説明したが、本開示の内容は上述した一実施形態に限定されるものではなく、本開示の技術的思想に基づく各種の変形が可能である。以下、変形例について説明する。

[0084] 上述した実施形態において、予測モデルの生成時間や予測モデルを生成する際に使用されるメモリ容量の制限、予測モデルの生成時間等に関する具体的な値や範囲がユーザによって指定されることにより第１データセットの内容が設定されても良い。また、上述した実施形態では、各種の設定や生成された予測モデルが表示によって報知されたが、音声等によって報知されても良い。表形式データは、ユーザによって入力されたデータでも良い。

[0085] 情報処理装置１で行われる処理の一部が、クラウド上の機器やスマートフォン等の外部機器で行われても良い。また、上述した実施形態における動作例の内容は、適宜、組み合わせることも可能である。

[0086] 一実施形態にかかる情報処理装置１の構成は適宜、変更することができる。例えば、情報処理装置１が、サーバ装置等と通信を行うための通信部や、音声を再生するためのスピーカ等を有していても良い。

[0087] 本開示は、装置、方法、プログラム、システム等により実現することもできる。例えば、上述した実施形態で説明した機能を行うプログラムをダウンロード可能とし、実施形態で説明した機能を有しない装置が当該プログラムをダウンロードしてインストールすることにより、当該装置において実施形態で説明した制御を行うことが可能となる。本開示は、このようなプログラムを配布するサーバにより実現することも可能である。また、各実施形態、変形例で説明した事項は、適宜組み合わせることが可能である。

[0088] なお、本開示中に例示された効果により本開示の内容が限定して解釈されるものではない。

[0089] 本開示は、以下の構成も採ることができる。

（１）

複数のデータを含む第１データセットが入力される入力部と、

前記第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別する判別部と、

前記第１データセットに対して、前記判別部により判別された処理を適用することにより前記第１データセットに基づく予測モデルを生成する予測モデル生成部と

を有する情報処理装置。

（２）

前記判別部は、前記第２データセットに基づく予測モデルを生成する際に適用されたアルゴリズム及び当該アルゴリズムにおけるパラメータの値を判別する

（１）に記載の情報処理装置。

（３）

所定のデータに対するユーザ入力に応じて前記第１データセットの内容が設定される

（１）又は（２）に記載の情報処理装置。

（４）

ユーザ入力に応じて、前記第１データセットに含ませるデータの特徴、前記予測モデル生成部により生成される予測モデルの値、前記予測モデル生成部による予測モデルの生成に要する時間、及び、前記予測モデル生成部による予測モデルの生成に要するメモリ容量の少なくとも一つが設定されることにより、前記第１データセットの内容が設定される

（３）に記載の情報処理装置。

（５）

前記ユーザ入力に応じて設定された前記第１データセットに含ませるデータの特徴のそれぞれに対して、前記第１データセットに基づく予測モデルを生成する際の前記特徴の有用度が報知される

（４）に記載の情報処理装置。

（６）

前記予測モデル生成部により予測モデルを生成する際に優先すべき処理項目が設定可能とされる

（１）から（５）までの何れかに記載の情報処理装置。

（７）

前記判別部は、前記第１データセットと類似し、前記設定された処理項目に対応する前記第２データセットに基づく予測モデルを生成する際に適用された処理を判別する

（６）に記載の情報処理装置。

（８）

前記予測モデルを生成する際の補助情報をユーザに対して問いかける報知がなされる

（１）から（７）までの何れかに記載の情報処理装置。

（９）

前記補助情報は、前記第１データセットに含まれる時系列なデータのうち前記予測モデルの生成に使用されるデータの期間、前記第１データセットに含まれるテキストデータのうち前記予測モデルの生成に使用されるテキスト

データの指定、及び、前記第１データセットに含まれる所定のデータの正確性に関する情報の少なくとも一つである

（８）に記載の情報処理装置。

（１０）

前記予測モデル生成部は、前記判別部により判別された処理及びユーザの応答により得られた補助情報に基づく処理を適用することにより前記第１データセットに基づく予測モデルを生成する

（７）又は（８）に記載の情報処理装置。

（１１）

前記第１データセットは前記入力部に対して現在入力されたデータセットであり、前記第２データセットは前記入力部に対して過去に入力されたデータセットである

（１）から（１０）までの何れかに記載の情報処理装置。

（１２）

判別部が、入力部に入力された複数のデータを含む第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別し、

予測モデル生成部が、前記第１データセットに対して、前記判別部により判別された処理を適用することにより前記第１データセットに基づく予測モデルを生成する

を有する情報処理方法。

（１３）

判別部が、入力部に入力された複数のデータを含む第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別し、

予測モデル生成部が、前記第１データセットに対して、前記判別部により判別された処理を適用することにより前記第１データセットに基づく予測モデルを生成する

を有する情報処理方法をコンピュータに実行させるプログラム。

符号の説明

[0090] １・・・情報処理装置

１１・・・制御部

１１Ａ・・・判別部

１１Ｂ・・・予測モデル生成部

１２・・・入力部

請求の範囲

[請求項1] 複数のデータを含む第１データセットが入力される入力部と、

前記第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別する判別部と、

前記第１データセットに対して、前記判別部により判別された処理を適用することにより前記第１データセットに基づく予測モデルを生成する予測モデル生成部と

を有する情報処理装置。

[請求項2] 前記判別部は、前記第２データセットに基づく予測モデルを生成する際に適用されたアルゴリズム及び当該アルゴリズムにおけるパラメータの値を判別する

請求項１に記載の情報処理装置。

[請求項3] 所定のデータに対するユーザ入力に応じて前記第１データセットの内容が設定される

請求項１に記載の情報処理装置。

[請求項4] ユーザ入力に応じて、前記第１データセットに含ませるデータの特徴、前記予測モデル生成部により生成される予測モデルの値、前記予測モデル生成部による予測モデルの生成に要する時間、及び、前記予測モデル生成部による予測モデルの生成に要するメモリ容量の少なくとも一つが設定されることにより、前記第１データセットの内容が設定される

請求項３に記載の情報処理装置。

[請求項5] 前記ユーザ入力に応じて設定された前記第１データセットに含ませるデータの特徴のそれぞれに対して、前記第１データセットに基づく予測モデルを生成する際の前記特徴の有用度が報知される

請求項４に記載の情報処理装置。

[請求項6] 前記予測モデル生成部により予測モデルを生成する際に優先すべき処理項目が設定可能とされる

請求項１に記載の情報処理装置。

[請求項7] 前記判別部は、前記第１データセットと類似し、前記設定された処理項目に対応する前記第２データセットに基づく予測モデルを生成する際に適用された処理を判別する

請求項６に記載の情報処理装置。

[請求項8] 前記予測モデルを生成する際の補助情報をユーザに対して問いかける報知がなされる

請求項１に記載の情報処理装置。

[請求項9] 前記補助情報は、前記第１データセットに含まれる時系列なデータのうち前記予測モデルの生成に使用されるデータの期間、前記第１データセットに含まれるテキストデータのうち前記予測モデルの生成に使用されるテキストデータの指定、及び、前記第１データセットに含まれる所定のデータの正確性に関する情報の少なくとも一つである

請求項８に記載の情報処理装置。

[請求項10] 前記予測モデル生成部は、前記判別部により判別された処理及びユーザの応答により得られた補助情報に基づく処理を適用することにより前記第１データセットに基づく予測モデルを生成する

請求項７に記載の情報処理装置。

[請求項11] 前記第１データセットは前記入力部に対して現在入力されたデータセットであり、前記第２データセットは前記入力部に対して過去に入力されたデータセットである

請求項１に記載の情報処理装置。

[請求項12] 判別部が、入力部に入力された複数のデータを含む第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別し、

予測モデル生成部が、前記第１データセットに対して、前記判別部により判別された処理を適用することにより前記第１データセットに基づく予測モデルを生成する

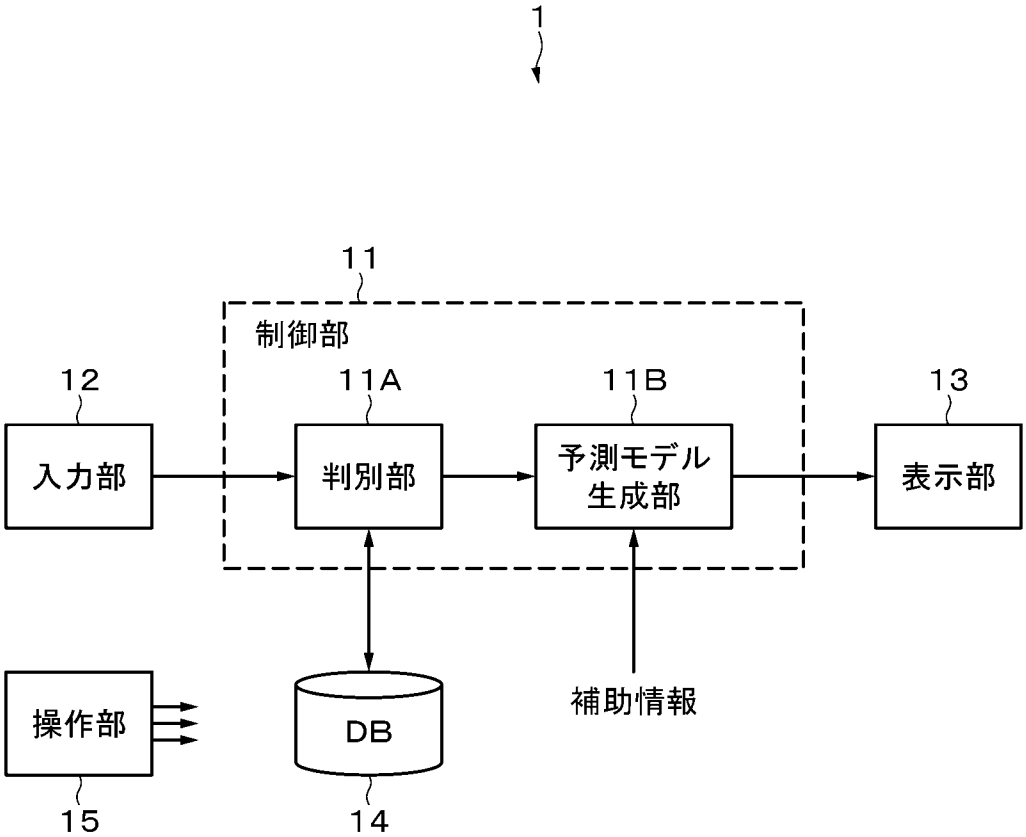
を有する情報処理方法。

[請求項13] 判別部が、入力部に入力された複数のデータを含む第１データセットと類似する第２データセットに基づく予測モデルを生成する際に適用された処理を判別し、

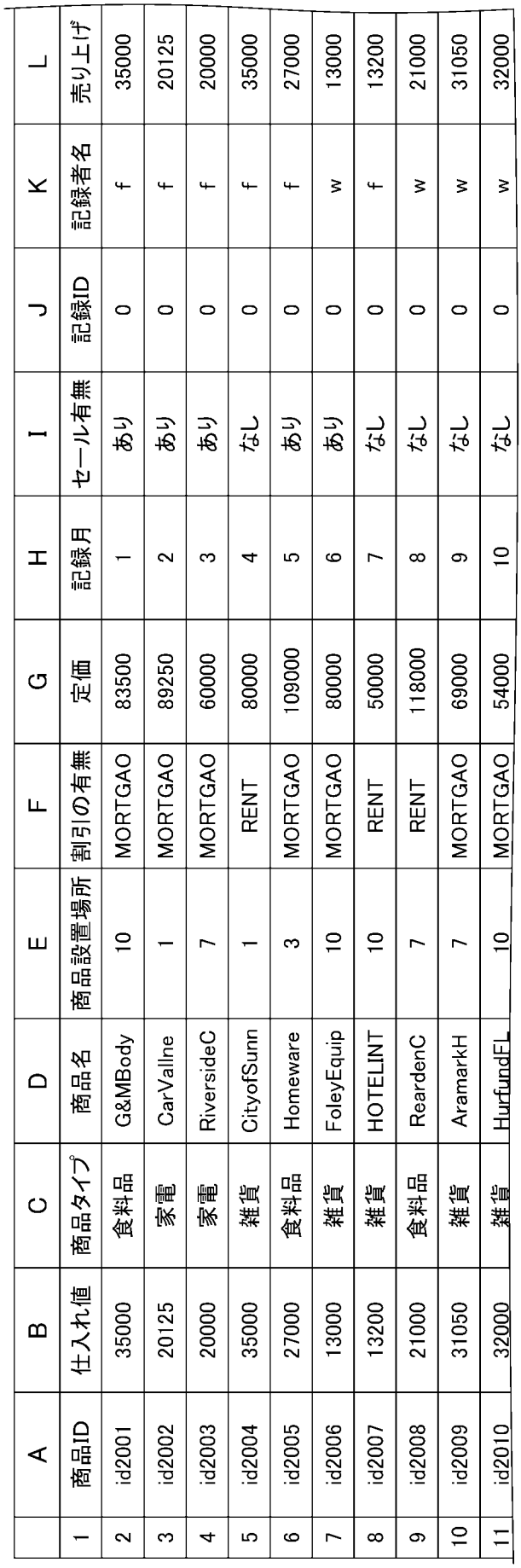
予測モデル生成部が、前記第１データセットに対して、前記判別部により判別された処理を適用することにより前記第１データセットに基づく予測モデルを生成する

を有する情報処理方法をコンピュータに実行させるプログラム。

[図1]



[図2]



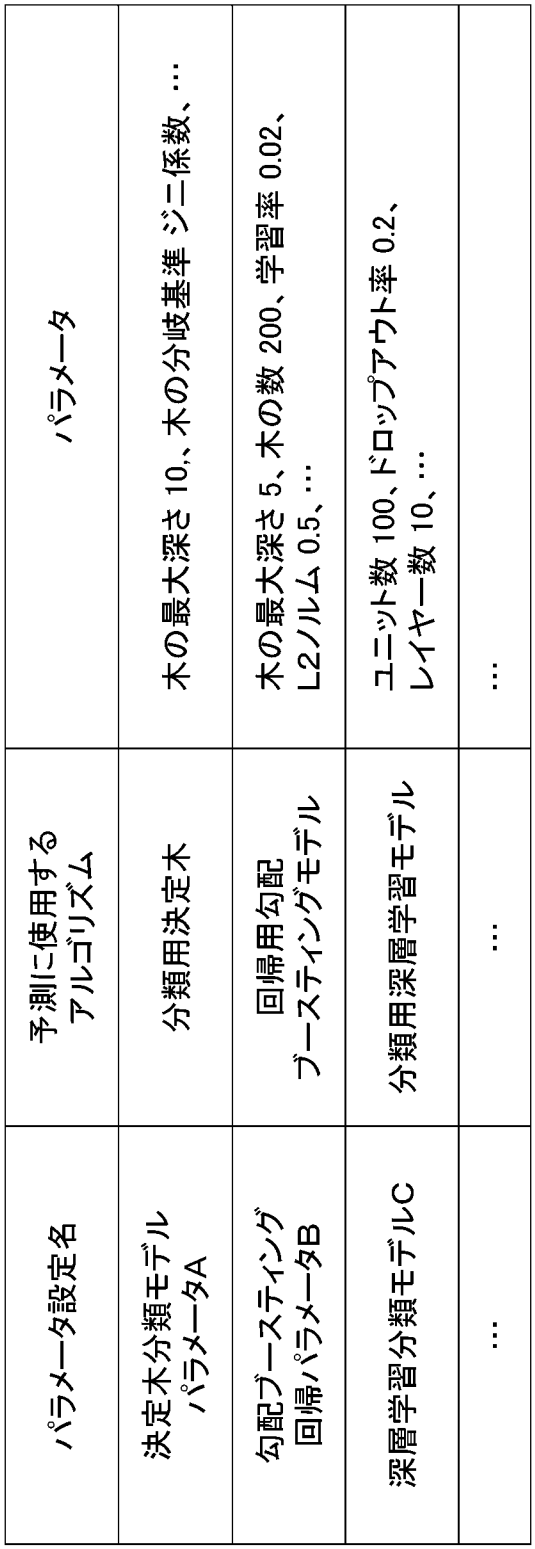
	A	B	C	D	E	F	G	H	I	J	K	L
1	商品ID	仕入れ値	商品タイプ	商品名	商品設置場所	割引の有無	定価	記録月	セール有無	記録ID	記録者名	売り上げ
2	id2001	35000	食料品	G&MBody	10	MORTGAO	83500	1	あり	0	f	35000
3	id2002	20125	家電	CarVallne	1	MORTGAO	89250	2	あり	0	f	20125
4	id2003	20000	家電	RiversideC	7	MORTGAO	60000	3	あり	0	f	20000
5	id2004	35000	雑貨	CityofSunn	1	RENT	80000	4	なし	0	f	35000
6	id2005	27000	食料品	Homeware	3	MORTGAO	109000	5	あり	0	f	27000
7	id2006	13000	雑貨	FoleyEquip	10	MORTGAO	80000	6	あり	0	w	13000
8	id2007	13200	雑貨	HOTELINT	10	RENT	50000	7	なし	0	f	13200
9	id2008	21000	食料品	ReardenC	7	RENT	118000	8	なし	0	w	21000
10	id2009	31050	雑貨	AramarkH	7	MORTGAO	69000	9	なし	0	w	31050
11	id2010	32000	雑貨	HurfundFL	10	MORTGAO	54000	10	なし	0	w	32000

[図3]



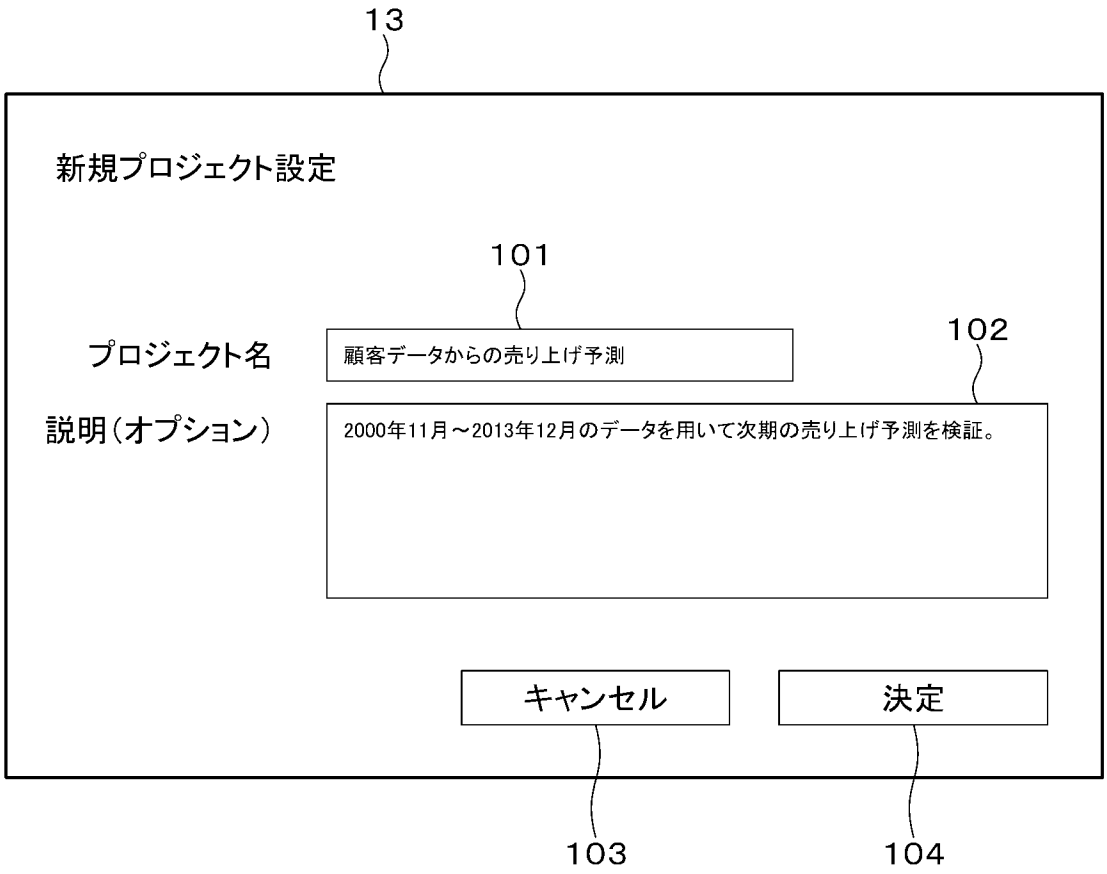
モデル名	表形式データファイル名	データセットの情報	各特徴の情報	予測モデル生成時間	予測モデル使用メモリ	探索した各パラメータの実験結果	予測モデル生成条件
Aローンの貸し倒れ予測モデル	2018/11/12-貸し倒れ.csv	- 特徴数 100 - データ数 20000 - 数値型の特徴数 10 - 文字列型の特徴数 70 - テキストの特徴数 20 - 欠損の割合 30% - 時系列データ - …	特徴ごとの情報（アルゴリズム、特徴のデータ型、平均・分散・尖度などの統計量、…）のリスト	1220秒	12GB	<決定木モデル パラメータA、 精度0.82、 再現率0.6、 F値0.2、…>, <パラメータ設定2、 精度0.73、 再現率0.5、 F値0.4、…>,	- 速度優先 - メモリ優先 - テキストの処理方法 -
B店の廃棄量予測モデル	2018/11/12-salesB.csv	- 特徴数 230 - データ数 490000 - 数値型の特徴数 100 - カテゴリ方の特徴数 10 - テキストの特徴数 120 - 欠損の割合 30% - 時系列データ - …		23000秒	30GB	<パラメータ設定3、 誤差平均0.82、…>, <パラメータ設定4、 誤差平均0.4、…>,	- 回帰問題 - 性能優先 -
…	…	…	…	…	…	…	…

[図4]

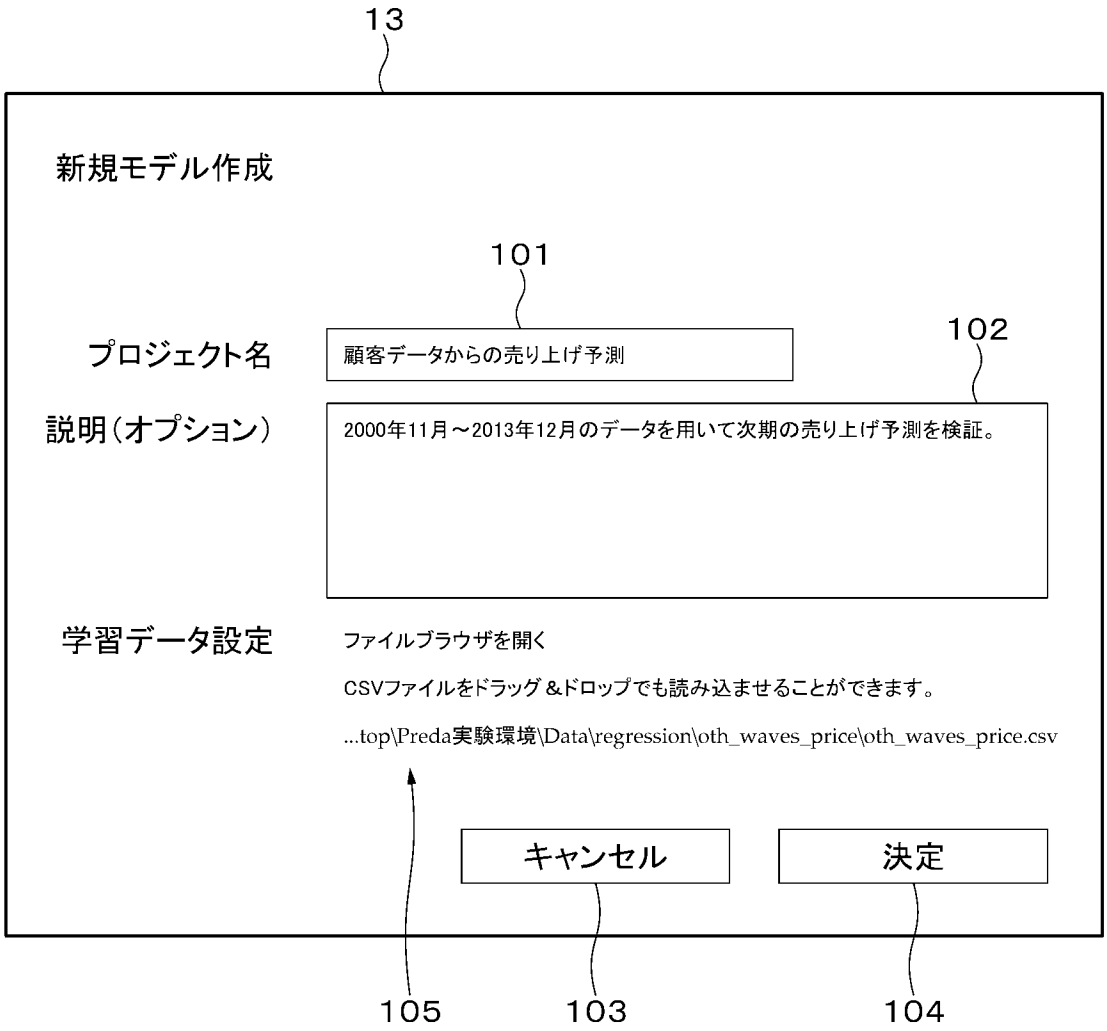


パラメータ設定名	予測に使用する アルゴリズム	パラメータ
決定木分類モデル パラメータA	分類用決定木	木の最大深さ 10,、木の分岐基準 ジニ係数、…
勾配ブースティング 回帰パラメータB	回帰用勾配 ブースティングモデル	木の最大深さ 5、木の数 200、学習率 0.02、 L2ノルム 0.5、…
深層学習分類モデルC	分類用深層学習モデル	ユニット数 100、ドロップアウト率 0.2、 レイヤー数 10、…
…	…	…

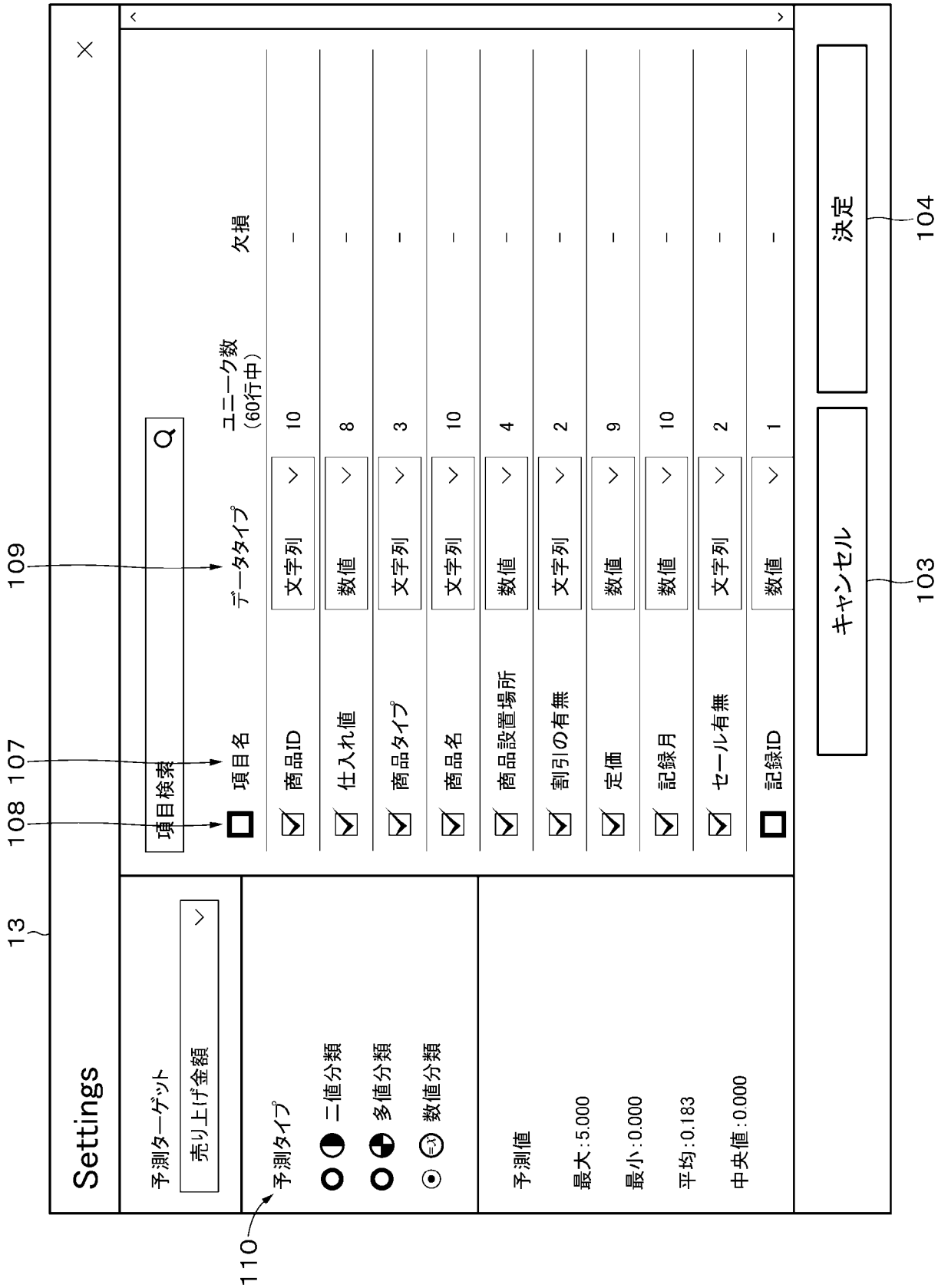
[図5]


13
新規プロジェクト設定
101
プロジェクト名
顧客データからの売り上げ予測
102
説明（オプション）
2000年11月～2013年12月のデータを用いて次期の売り上げ予測を検証。
キャンセル
決定
103
104

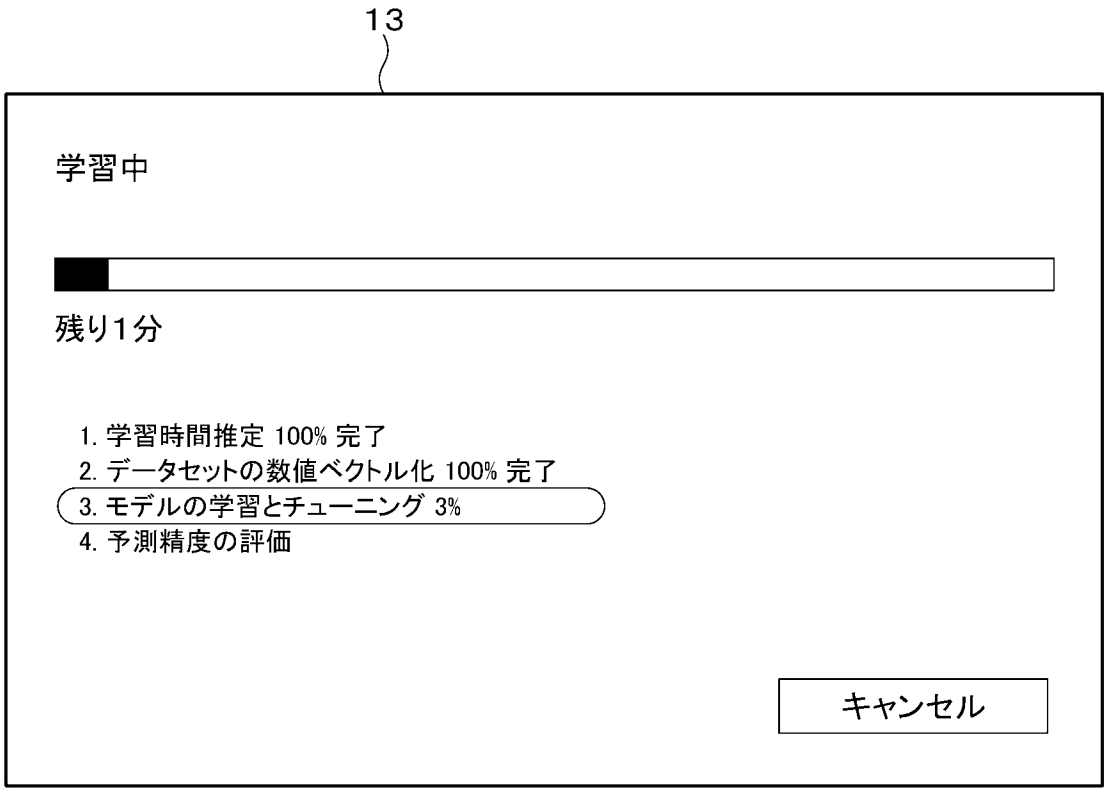
[図6]


13
新規モデル作成
101
プロジェクト名
顧客データからの売り上げ予測
102
説明（オプション）
2000年11月～2013年12月のデータを用いて次期の売り上げ予測を検証。
学習データ設定
ファイルブラウザを開く
CSVファイルをドラッグ＆ドロップでも読み込ませることができます。
...top\Preda実験環境\Data\regression\oth_waves_price\oth_waves_price.csv
キャンセル
決定
105
103
104

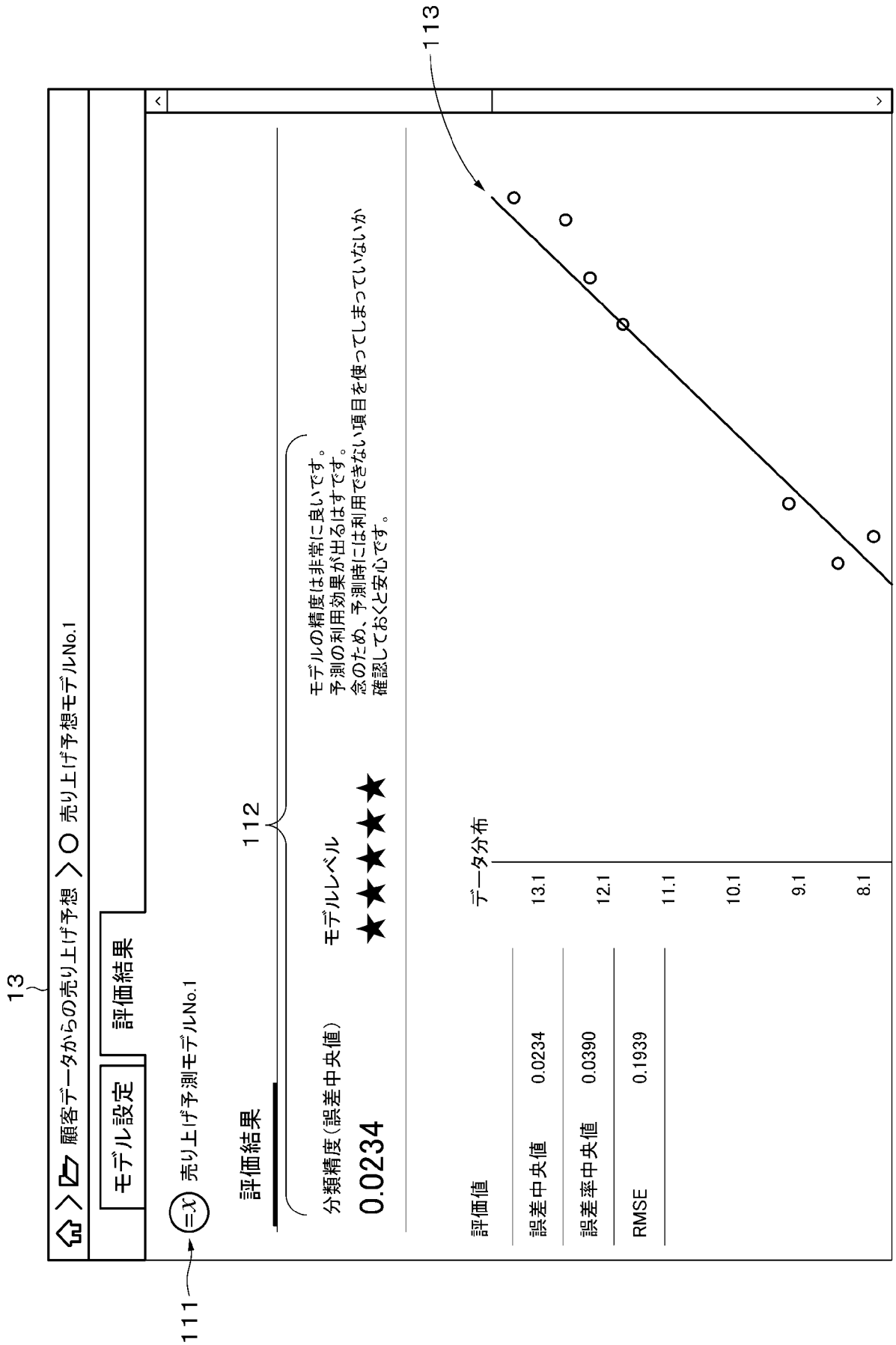
[図7]


Settings
予測ターゲット
売り上げ金額
予測タイプ
二値分類
多値分類
数値分類
予測値
最大:5.000
最小:0.000
平均:0.183
中央値:0.000
項目検索
項目名
データタイプ
ユニーク数
(60行中)
欠損
商品ID 文字列 10 –
仕入れ値 数値 8 –
商品タイプ 文字列 3 –
商品名 文字列 10 –
商品設置場所 数値 4 –
割引の有無 文字列 2 –
定価 数値 9 –
記録月 数値 10 –
セール有無 文字列 2 –
記録ID 数値 1 –
キャンセル
決定
13
108
107
109
110
103
104

[図8]


13
学習中
残り1分
1. 学習時間推定 100% 完了
2. データセットの数値ベクトル化 100% 完了
3. モデルの学習とチューニング 3%
4. 予測精度の評価
キャンセル

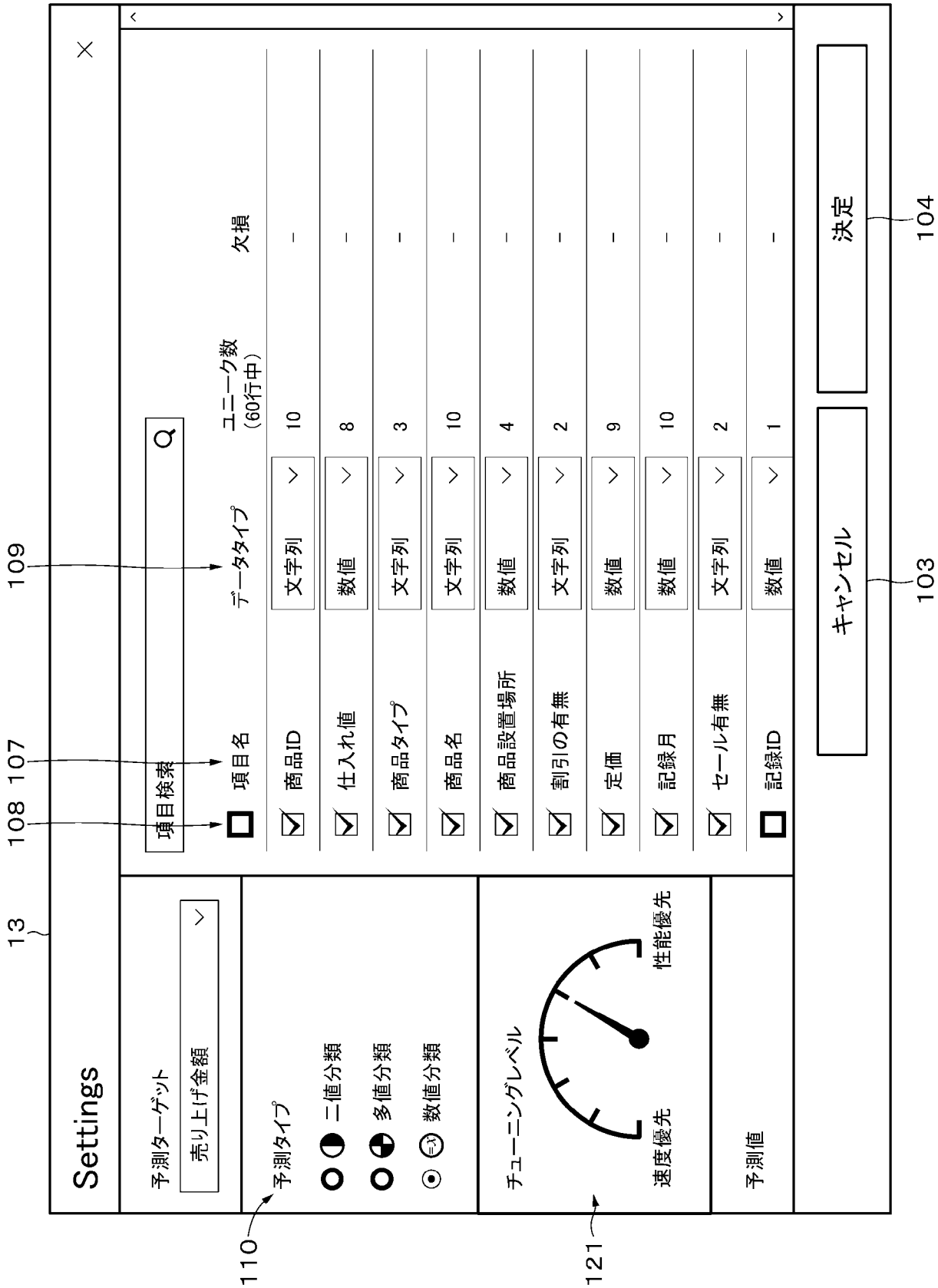
[図9]


13
顧客データからの売り上げ予想 > 売り上げ予想モデルNo.1
モデル設定
評価結果
売り上げ予測モデルNo.1
111
評価結果
112
分類精度（誤差中央値）
0.0234
モデルレベル
★★★★★
モデルの精度は非常に良いです。
予測の利用効果が出るはずです。
念のため、予測時には利用できない項目を使ってしまっていないか
確認しておくと安心です。
113
評価値
誤差中央値 0.0234
誤差率中央値 0.0390
RMSE 0.1939
データ分布
13.1
12.1
11.1
10.1
9.1
8.1

[図10]

アルゴリズム / アルゴリズムの特徴	C2 三角関数で変換	連続値に変換	C3 20の区間に区切る	C1 欠損値を削る	二乗する
性能への影響の平均（＋だと良い）	−0.1	−1.2	+0.01	+0.11	−0.1
性能の分散	0.82	0.02	0.11	0.02	0.45
データベースの記録数	129	19	4	100	3

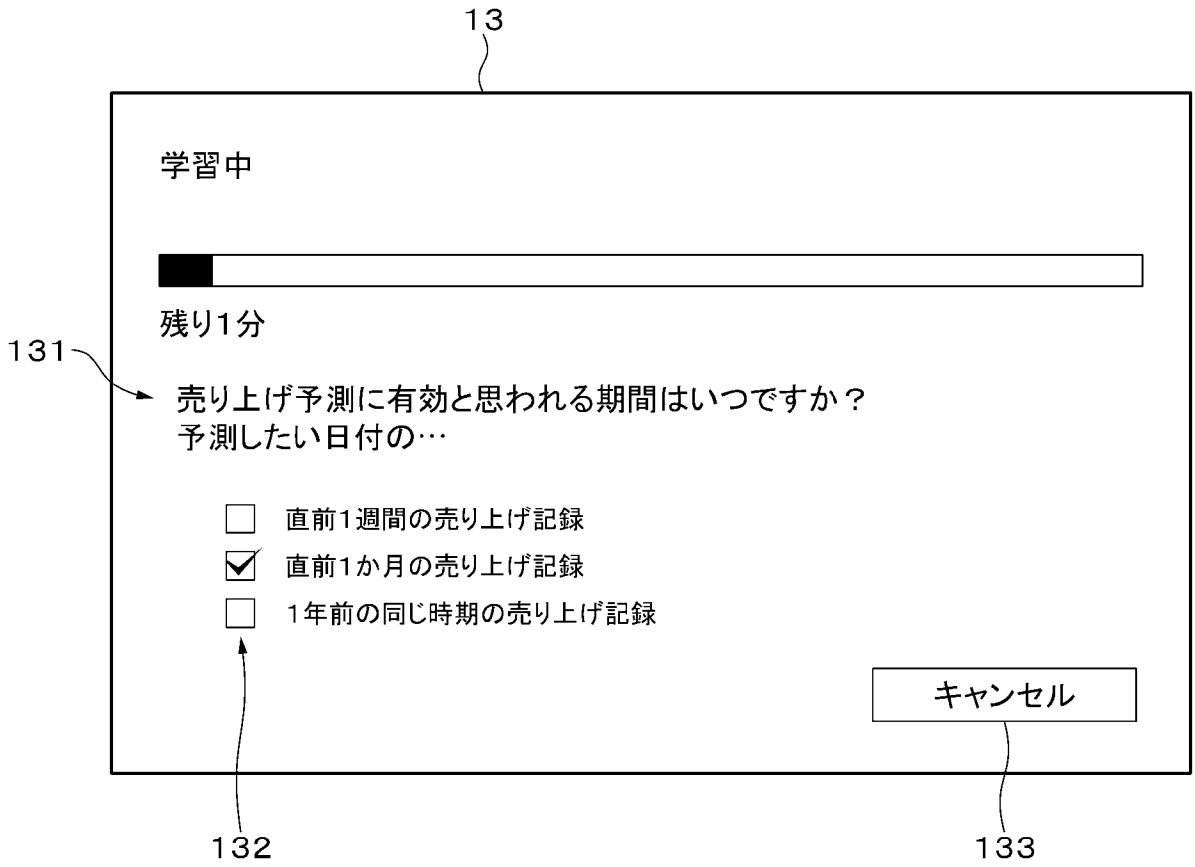
[図11]


Settings
予測ターゲット
売り上げ金額
予測タイプ
二値分類
多値分類
数値分類
チューニングレベル
速度優先
性能優先
予測値
項目検索
項目名
データタイプ
ユニーク数
(60行中)
欠損
商品ID
文字列
10
仕入れ値
数値
8
商品タイプ
文字列
3
商品名
文字列
10
商品設置場所
数値
4
割引の有無
文字列
2
定価
数値
9
記録月
数値
10
セール有無
文字列
2
記録ID
数値
1
キャンセル
決定
13
107
108
109
110
121
103
104

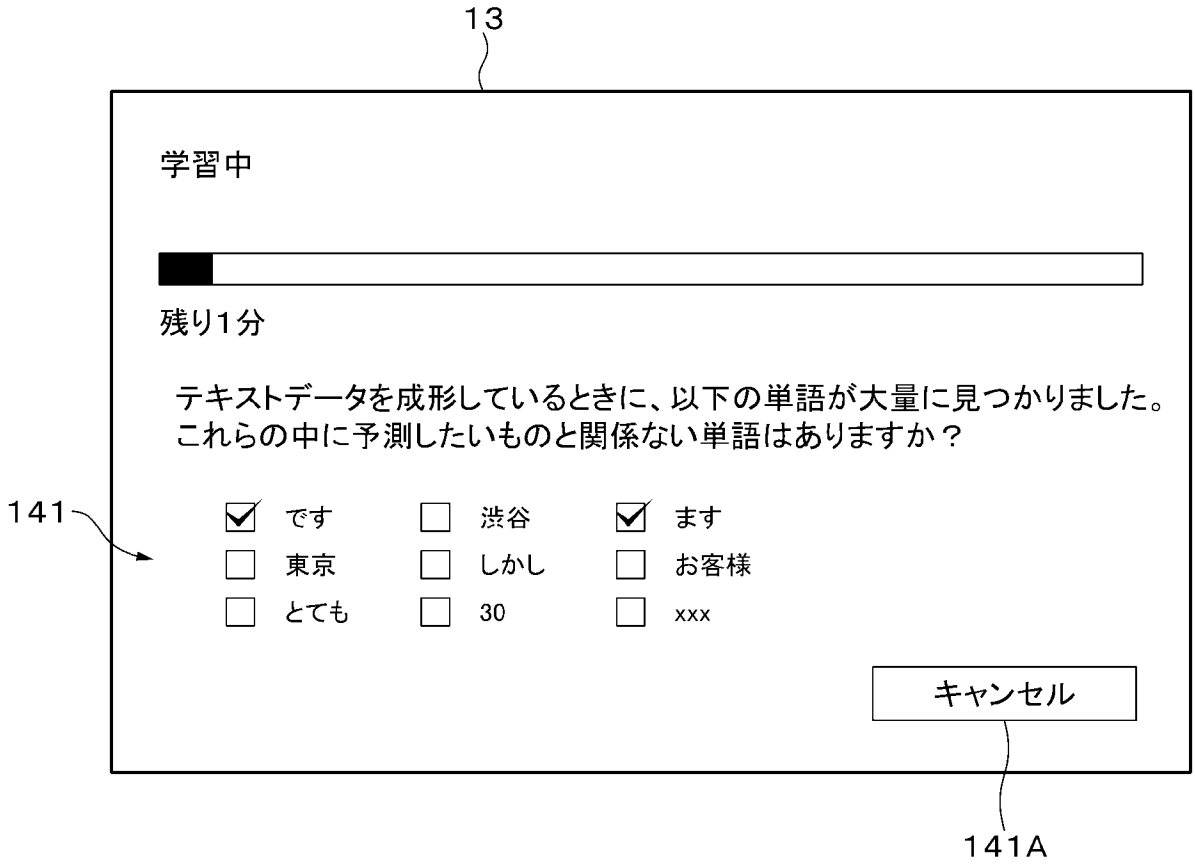
[図12]

特徴名	適用アルゴリズム	性能への寄与	計算時間	メモリ	次元数
Age	スケーリング	+.12	0.2	2	1
Price	20のビンに区切る	+.08	1.1	10	20
amount	スケーリング	+.11	0.1	1	1
…	…				

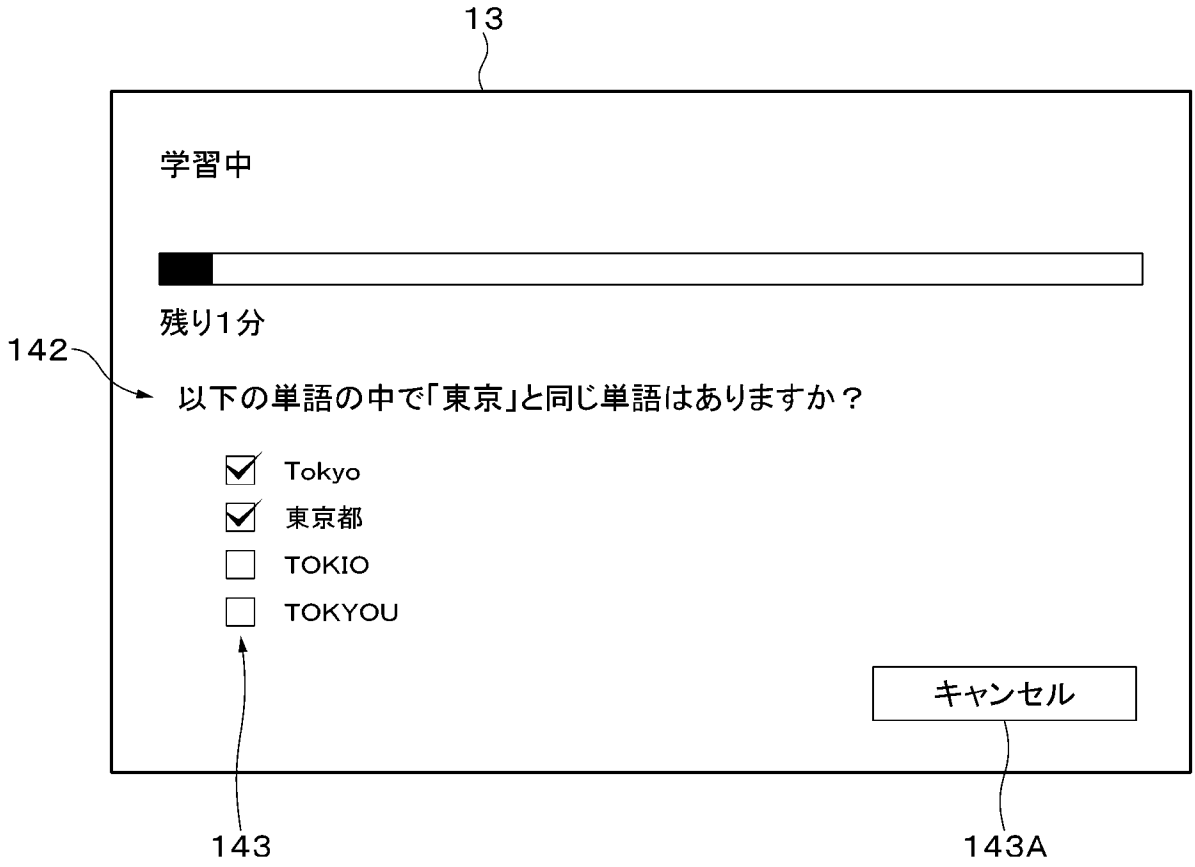
[図13]



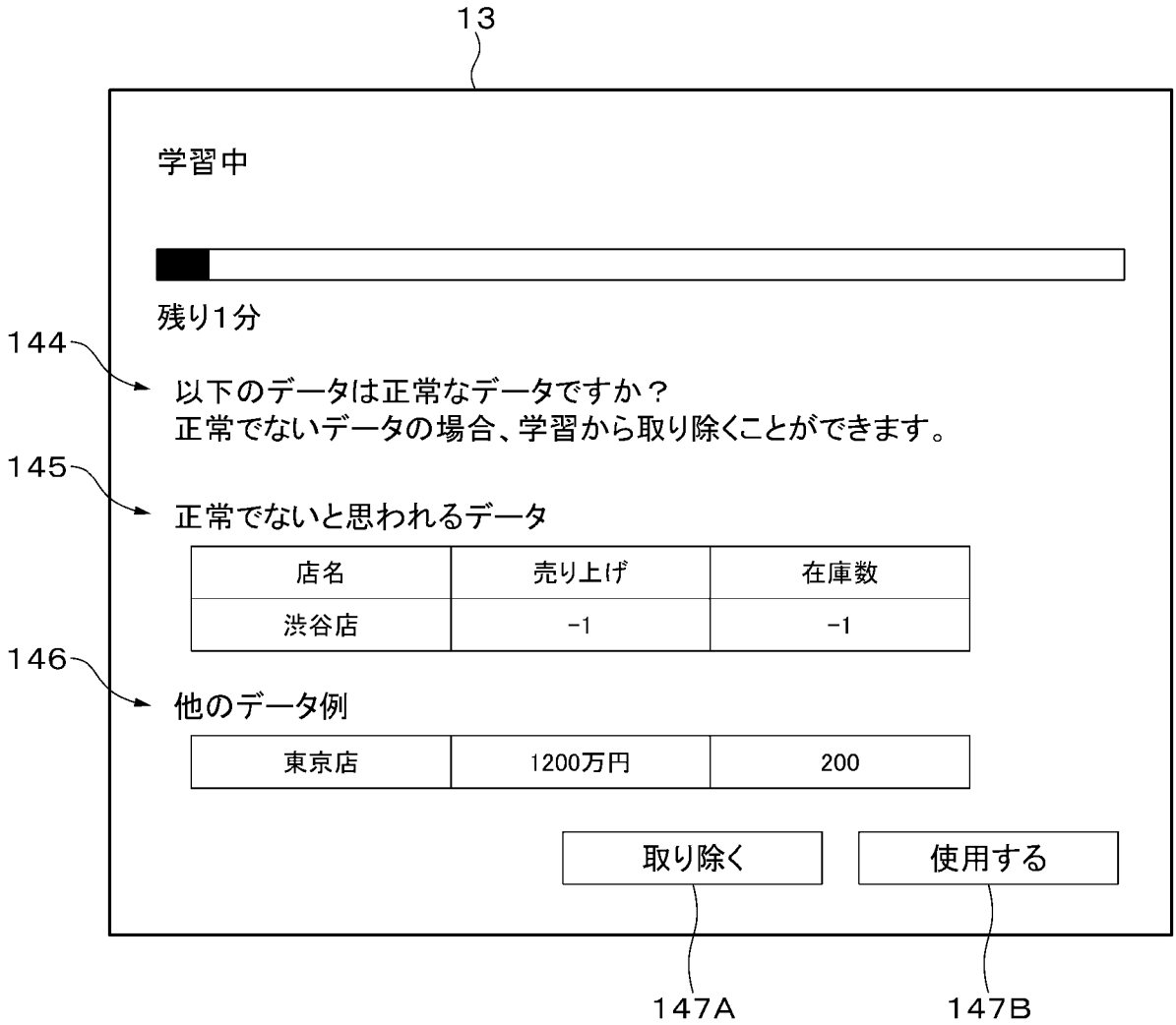
[図14]


13
学習中
残り1分
テキストデータを成形しているときに、以下の単語が大量に見つかりました。
これらの中に予測したいものと関係ない単語はありますか？
141
です
渋谷
ます
東京
しかし
お客様
とても
30
xxx
キャンセル
141A

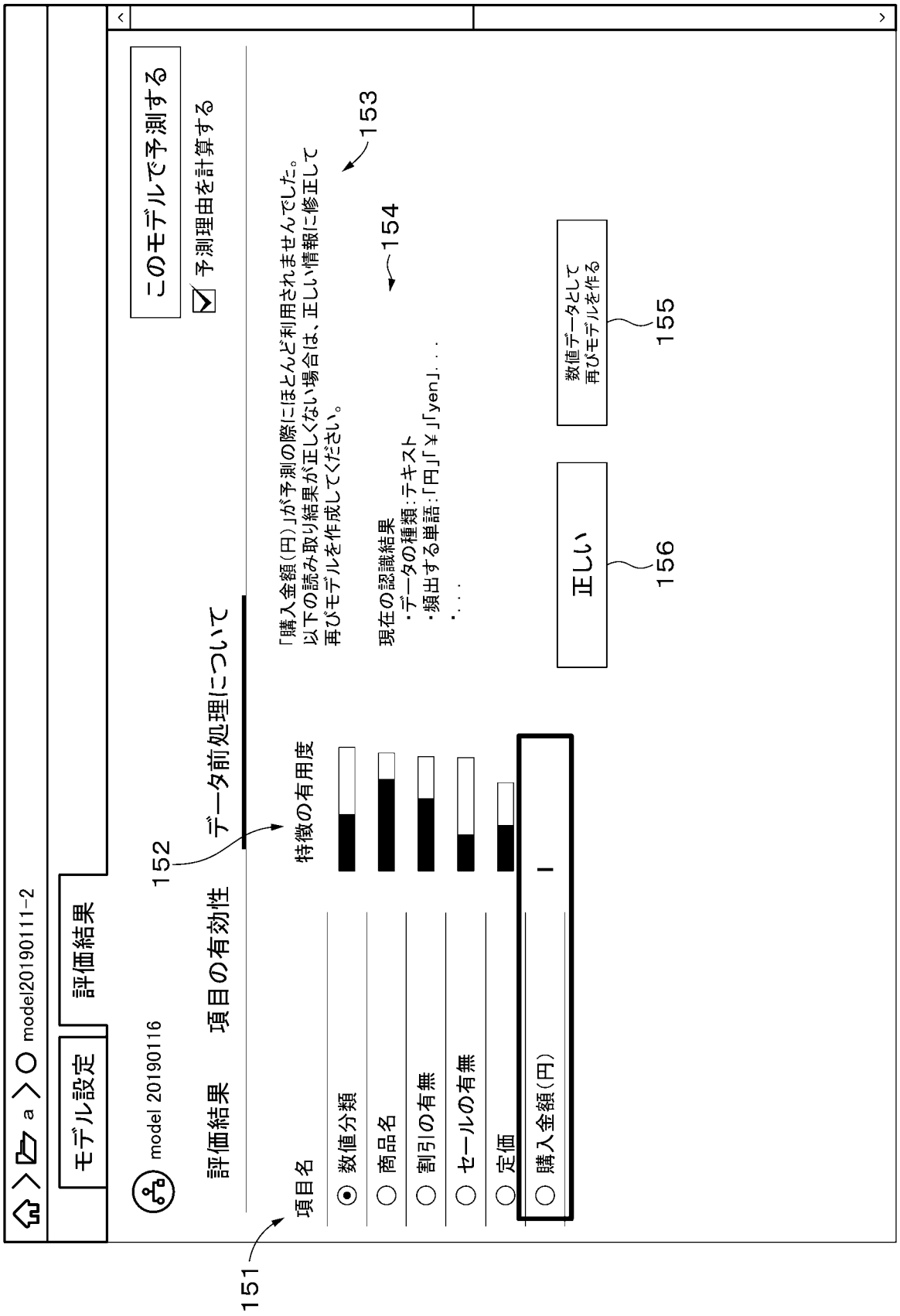
[図15]


13
学習中
残り1分
142
以下の単語の中で「東京」と同じ単語はありますか？
Tokyo
東京都
TOKIO
TOKYOU
キャンセル
143
143A

[図16]


13
学習中
残り1分
144
以下のデータは正常なデータですか？
正常でないデータの場合、学習から取り除くことができます。
145
正常でないと思われるデータ
店名
売り上げ
在庫数
渋谷店
-1
-1
146
他のデータ例
東京店
1200万円
200
取り除く
使用する
147A
147B

[図17]


model20190111-2
モデル設定
評価結果
model 20190116
このモデルで予測する
予測理由を計算する
評価結果　項目の有効性
データ前処理について
152
「購入金額(円)」が予測の際にほとんど利用されませんでした。
以下の読み取り結果が正しくない場合は、正しい情報に修正して
再びモデルを作成してください。
153
154
現在の認識結果
・データの種類:テキスト
・頻出する単語:「円」「￥」「yen」...
...
項目名
特徴の有用度
数値分類
商品名
割引の有無
セールの有無
定価
購入金額(円)
正しい
156
数値データとして
再びモデルを作る
155
151

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/018400

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. G06Q50/06(2012.01)i
FI: G06Q50/06

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. G06Q50/06

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 2019-16361 A (NANTOMICS LLC) 31.01.2019 (2019-01-31), paragraph [0034]	1, 12, 13
A	JP 2019-75159 A (NEC CORPORATION) 16.05.2019 (2019-05-16), entire text	1-13

☐ 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 27.07.2020	Date of mailing of the international search report 11.08.2020
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/JP2020/018400

JP 2019-16361 A	31.01.2019	US 2018/0190381 A1 paragraph [0039] CN 108292329 A paragraph [0039]
JP 2019-75159 A	16.05.2019	US 2018/0181875 A1 entire text

国際調査報告

国際出願番号
PCT/JP2020/018400

A.　発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06Q 50/06(2012.01)i
FI: G06Q50/06

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））

G06Q50/06

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 2019-16361 A (ナントミクス, エルエルシー) 31.01.2019 (2019-01-31) ００３４	１，１２，１３
A	JP 2019-75159 A (日本電気株式会社) 16.05.2019 (2019-05-16) 全文	１－１３

☐ 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
27.07.2020	11.08.2020

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）

石川　正二 5R 8524

電話番号 03-3581-1101 内線 3502

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/JP2020/018400

引用文献	公表日	パテントファミリー文献	公表日
JP 2019-16361 A	31.01.2019	US 2018/0190381 A1 0039 CN 108292329 A 0039	
JP 2019-75159 A	16.05.2019	US 2018/0181875 A1 全文	