(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
2018年10月11日(11.10.2018)

WIPO | PCT

(10) 国際公開番号
WO 2018/186090 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)　*G06Q 10/04* (2012.01)

(21) 国際出願番号：　PCT/JP2018/008465

(22) 国際出願日：　2018年3月6日(06.03.2018)

(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 2017-075790　2017年4月6日(06.04.2017)　JP

(71) 出願人: テンソル・コンサルティング株式会社 (**TENSOR CONSULTING CO. LTD.**) [JP/JP]; 〒1020076 東京都千代田区五番町２番地２４ Tokyo (JP).

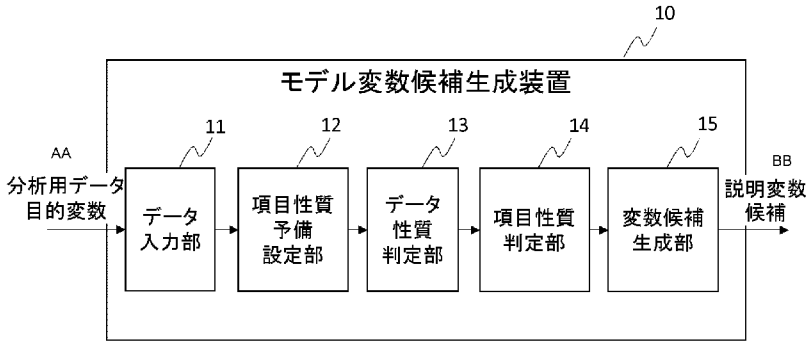
(72) 発明者: 藤本　浩司 (**FUJIMOTO, Koji**); 〒1020076 東京都千代田区五番町２番地２４ テンソル・コンサルティング株式会社内 Tokyo (JP). 柴原 一友 (**SHIBAHARA, Kazutomo**); 〒1020076 東京都千代田区五番町２番地２４ テンソル・コンサルティング株式会社内 Tokyo (JP). 是川 空 (**KOREKAWA, Takashi**); 〒1020076 東京都千代田区五番町２番地２４ テンソル・コンサルティング株式会社内 Tokyo (JP).

(74) 代理人: 特許業務法人ウィルフォート国際特許事務所 (**WILLFORT INTERNATIONAL PATENT FIRM**); 〒1030016 東京都中央区日本橋小網町１９－７ 日本橋ＴＣビル １階 Tokyo (JP).

(54) **Title:** MODEL VARIABLE CANDIDATE GENERATION DEVICE AND METHOD

(54) 発明の名称： モデル変数候補生成装置および方法

図2



10 Model variable candidate generation device
11 Data input unit
12 Item nature preliminarily setting unit
13 Data nature determination unit
14 Item nature determination unit
15 Variable candidate generation unit
AA Objective variable of data for analysis
BB Explanatory variable candidate

(57) **Abstract:** The present invention makes it possible to quickly narrow down the variables of a model in data analysis. A model variable candidate generation device for generating an explanatory variable candidate that is made a candidate for an explanatory variable in generating a prediction model, the device having a data input unit for inputting data for analysis that includes one or more items for each entry and has item values for the items, a first item determination unit for preliminarily setting the natures of the items included in the data for analysis as first item natures, a data nature determination unit for determining data natures that are the natures of the data for analysis on the basis of the first item natures of the items included in the data for analysis, a second item determination unit for determining the natures of the

WO 2018/186090 A1

［続葉有］

(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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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:

— 国際調査報告(条約第21条(3))

items included in the data for analysis as second item natures on the basis of the data natures of the data for analysis, and a variable candidate generation unit for generating an explanatory variable candidate on the basis of the second item natures of the items included in the data for analysis by selecting from the items or processing the items.

(57) 要約:データ分析におけるモデルの変数の絞り込みを迅速に行うことを可能にする。予測モデルの生成において説明変数の候補とする説明変数候補を生成するモデル変数候補生成装置は、エントリ毎に1つ以上の項目を含み、項目に項目値を有する分析用データを入力するデータ入力部と、分析用データに含まれる項目の性質を第1項目性質として予備設定する第1項目判定部と、分析用データに含まれる項目の第1項目性質に基づき、分析用データの性質であるデータ性質を判定するデータ性質判定部と、分析用データのデータ性質に基づき、分析用データに含まれる項目の性質を第2項目性質として判定する第2項目判定部と、分析用データに含まれる項目の第2項目性質に基づいて、項目から選択または項目を加工することより、説明変数候補を生成する変数候補生成部と、を有する。

明　細　書

発明の名称 ： モデル変数候補生成装置および方法

技術分野

[0001] 本発明はデータ分析に用いるモデルの変数候補の抽出に関する。

背景技術

[0002] 収集、蓄積された膨大なデータ間の関係性を解析することにより所望の予測対象の値を予測するビッグデータ解析技術に注目が集まっている。予測対象の値の算出には、蓄積されたデータを基に生成したモデルが利用される。取得されたデータを説明変数としてモデルに入力すると予測対象が目的変数の値として出力される。

[0003] モデル生成は膨大なデータの解析を要する煩雑な作業である。モデル生成の一部の作業はコンピュータにより自動化され、効率化が図られている。データに含まれる変数から効果の高い変数だけを選択し、データを離散化し、そのデータからモデルを構築し、構築したモデルを検証するといった各工程については、ある程度は自動化が可能である。

[0004] 例えば、特許文献１には、予測モデルに追加すべき説明変数を見つける際に行う計算の量を削減することができる情報処理装置が開示されている。

[0005] 特許文献１に開示されている情報処理装置は、複数の目的変数の各々について、当該目的変数の実際の値と当該目的変数の値を予測するための第１の予測モデルによって算出された値との誤差に基づいて複数の目的変数を複数のグループに分類し、当該複数のグループの各々について、当該グループに属する目的変数について算出された誤差を用いて当該誤差の代表値を算出し、複数のグループの各々について、算出した代表値を予測するための第２の予測モデルを、説明変数を変えつつ複数生成し、生成された複数の第２の予測モデルによって算出された値の各々と代表値との差に基づき、当該グループに属する目的変数の第１の予測モデルに追加する説明変数を決定する。

[0006] また、特許文献２には、特定の変数が特定の値をとる確率を計算するモデ

ルを構築するために用いる説明変数を選択する変数選択装置が開示されている。

[0007] 特許文献２に開示されている変数選択装置は、目的変数が第１値及び第２値をもつサンプルの頻度を第１頻度及び第２頻度として計数し、説明変数ごとに、説明変数が第１値であり目的変数が第１値であるサンプルの頻度を第３頻度、説明変数が第１値であり目的変数が第２値であるサンプルの頻度を第４頻度として計数し、第１頻度と、第２頻度と、説明変数ごとに得られた第３頻度及び第４頻度とを用いて、各説明変数の特徴量をそれぞれ算出し、算出された各特徴量に基づき１つ以上の説明変数を選択する。

[0008] また、特許文献３には、説明変数を効率的に選択することができる装置が開示されている。

[0009] 特許文献３に開示されている装置は、線形予測子が、複数の説明変数候補と複数の説明変数候補にそれぞれ対応する複数の係数との線形結合と、定数項との和により表される変数選択用モデルを用いて、複数の説明変数候補から所望の説明変数を選択する際に、複数の係数のうちの少なくとも１つの係数に関して符号条件を取得し、複数のデータを用いて、符号条件の下で、複数の係数の推定値及び定数項の推定値を算出し、推定値が非ゼロと算出された係数に対応する説明変数候補を所望の説明変数として選択する。

[0010] 特許文献４には、精度よく目的変数を予測する予測モデルを生成するデータ分析システムが開示されている。

[0011] 特許文献４に開示されているデータ分析システムは、学習データに基づいて、複数の説明変数から目的変数を予測する予測モデルを生成し、各レコードについて、予測モデルに基づく予測の確からしさを示す信頼度を計算し、複数のレコードのうち、信頼度が所定の範囲内であるレコードからなるサブセットを作成し、サブセットに属するレコードに基づいて、複数の説明変数から、目的変数との相関が高い説明変数の組み合わせを抽出し、抽出した説明変数の組み合わせを新たな説明変数として学習データに追加する。

先行技術文献

特許文献

[0012] 特許文献1：特開２０１３－１５２６５６号公報

特許文献2：特開２００８－１５８７４８号公報

特許文献3：特許第６０６９４６０号公報

特許文献4：特開２０１６－４５２５号公報

発明の概要

発明が解決しようとする課題

[0013] ＳＮＳ（Ｓｏｃｉａｌ Ｎｅｔｗｏｒｋｉｎｇ Ｓｅｒｖｉｃｅ）やＩｏＴ（Ｉｎｔｅｒｎｅｔ ｏｆ Ｔｈｉｎｇｓ）の発達など情報通信の環境変化で、ビッグデータの利用が本格化すると、収集されるデータの種類とデータに含まれる項目が膨大となる。また、収集されるデータの種類およびデータに含まれる項目の変化が頻繁となる。

[0014] データの種類も量も莫大になると、ある目的変数を予測するモデルを生成するとき、説明変数の候補となるデータの項目の大半が目的変数とはほぼ相関がないという状況となる。

[0015] 特許文献１～４に開示された技術はいずれも、予め用意された説明変数の候補から、目的変数を精度よく予測するために効果の高い説明変数を所定のアルゴリズム演算により選択する技術である。

[0016] しかし、上述したような、説明変数の候補となるデータの項目の大半が目的変数とは殆ど相関がないという状況で、データや項目の性質を考慮せずに全てのデータの全ての項目を説明変数の候補としたのでは適切に説明変数を選択し、精度の高いモデルを生成することは困難である。精度の高いモデルを生成するには、説明変数の候補は網羅的であることが望ましい一方で、データや項目の性質に基づく目的変数との関係を考慮してある程度絞り込んでおくことも求められる。

[0017] そのため、膨大なデータの項目の中から、説明変数を選択するアルゴリズム演算に入力する説明変数となりうる適切な候補を生成しておく作業が必要となる。しかし現状では、妥当な説明変数の候補を生成しておく方法は確立

されておらず人間が経験に基づいて行っている。

[0018] ビッグデータの活用が本格化したとき、膨大に存在しかつ頻繁に変化するデータを絞り込む作業を人手に頼っていたのでは、それがボトルネックとなり、モデルの更新が遅れ、目的変数の予測精度が低下する可能性がある。

[0019] 本発明の目的は、データ分析におけるモデルの変数の候補を予め生成しておく技術を提供することである。

課題を解決するための手段

[0020] 本発明の一態様によるモデル変数候補生成装置は、予測モデルの生成において説明変数の候補とする説明変数候補を生成するモデル変数候補生成装置であって、データ入力部と、第１項目判定部と、データ性質判定部と、第２項目判定部と、変数候補生成部と、を有する。

[0021] データ入力部は、エントリ毎に１つ以上の項目を含み、前記項目に項目値を有する分析用データを入力する。第１項目判定部は、前記分析用データに含まれる前記項目の性質を第１項目性質として予備設定する。データ性質判定部は、前記分析用データに含まれる項目の前記第１項目性質に基づき、該分析用データの性質であるデータ性質を判定する。第２項目判定部は、前記分析用データの前記データ性質に基づき、該分析用データに含まれる前記項目の性質を第２項目性質として判定する。変数候補生成部は、前記分析用データに含まれる項目の前記第２項目性質に基づいて、前記項目から選択または前記項目を加工することより、前記説明変数候補を生成する。

発明の効果

[0022] 分析用データの項目の性質を予備設定し、予備設定した項目の性質に基づいて分析用データの性質を判定し、分析用データの性質に基づいて項目の性質を本格的に判定し、その本格的に判定した項目の性質に基づいて説明変数候補を生成するというように、ステップバイステップの手順により説明変数候補を生成する。そのため、分析用データ内に判断材料がないため機械的な判断が困難であった分析用データおよび各項目の性質を判断することが可能となり、分析用データおよび項目を考慮してデータ分析における予測モデル

の変数の絞り込みを迅速に行うことができる。

図面の簡単な説明

[0023] [図1]モデル生成システムのブロック図である。

[図2]本実施形態によるモデル変数候補生成装置のブロック図である。

[図3]本実施形態によるモデル変数候補生成処理の全体の概要フローチャートである。

[図4]対象項目がＩＤの入る項目であるか否か判断する処理のフローチャートである。

[図5]対象項目が登録日の入る項目であるか否か判断する処理のフローチャートである。

[図6]マスターデータの一例を示す図である。

[図7]マスターデータの他の例を示す図である。

[図8]トランザクションデータの一例を示す図である。

[図9]スナップショットデータの一例を示す図である。

発明を実施するための形態

[0024] 本発明の実施形態について図面を参照して説明する。

[0025] 図１は、モデル生成システムのブロック図である。

[0026] モデル生成システム９０は、分析用データを用いて機械学習を行い、目的変数の予測を可能にする予測モデルを生成するシステムである。分析用データは、不図示の分析対象に関連するデータであり、一例として実際に取得されデータである。分析用データは１つ以上のエントリを含む。各エントリは複数の項目で構成されている。機械学習にはある程度、大容量のデータが必要である。分析用データとして、複数のエントリを含む大容量のデータが与えられてもよいし、項目の構成が同一の比較的小容量のデータが多数個与えられてもよい。エントリ毎に各項目には項目値がそれぞれ入っている。目的変数は、予測モデルによって予測する項目であり、予測モデルにより予測値を生成する変数である。

[0027] 図１を参照すると、モデル生成システム９０は、モデル変数候補生成装置

１０およびモデル生成装置２０を有している。モデル変数候補生成装置１０およびモデル生成装置２０は一例として計算機でソフトウェアプログラムを実行することにより実現される装置である。モデル変数候補生成装置１０とモデル生成装置２０は物理的に別々の計算機上に実装されてもよいし、同一の計算機上に実装されてもよい。

[0028] モデル変数候補生成装置１０には分析用データと目的変数を入力とし、目的変数を予測する予測モデルを生成するときに説明変数の候補として用いる説明変数候補を出力する。説明変数候補は、予測モデルの入力となる説明変数の候補である。予測モデルによる予測の精度を高めるためには、目的変数に影響しない変数を除外するとともに、目的変数に影響しうる説明変数の候補を網羅的に抽出しておくことが望ましい。説明変数は説明変数候補の中から選択される。説明変数候補はモデル生成装置２０に与えられる。

[0029] モデル生成装置２０には教師データおよび検証用データが与えられる。教師データは、予測モデルを生成する機械学習に与えられるデータである。検証用データは予測モデルの性能を評価するために教師データとは別に用意されたデータである。教師データおよび検証用データは例えば分析用データの中から選択あるいは分析用データを加工して生成されたデータである。教師データおよび検証用データは複数のエントリを含む。各エントリは複数の項目で構成されている。エントリ毎の各項目には項目値がそれぞれ入っている。教師データおよび検証用データそれぞれの項目には説明変数および／または目的変数が含まれている。なお、教師データにより機械学習した予測モデルの性能の評価は必須ではないので評価が不要であれば検証用データを用意する必要はない。

[0030] モデル生成装置２０は、説明変数候補値および目的変数値を含む教師データを入力とし、予測モデルを生成する。予測モデルは予測装置３０に与えられる。

[0031] 予測装置３０は、一例として、計算機にてソフトウェアプログラムを実行することにより実現される装置である。予測装置３０は、説明変数値あるい

はその元になるデータを入力とし、予測モデルを用いて目的変数値を算出し、目的変数値を予測結果として出力する。

[0032] 図２は、本実施形態によるモデル変数候補生成装置のブロック図である。

[0033] 図２を参照すると、モデル変数候補生成装置１０は、データ入力部１１、項目性質予備設定部１２、データ性質判定部１３、項目性質判定部１４、および変数候補生成部１５を有している。

[0034] データ入力部１１は、与えられる分析用データを受け付け、モデル変数候補生成装置１０内に入力する。分析用データは、複数のエントリを有し、エントリ毎に複数の項目を含み、それぞれの項目に項目値を有する。分析用データの詳細は後述する。

[0035] 項目性質予備設定部１２は、分析用データに含まれる項目の性質を予備設定する。ここでは、項目性質予備設定部１２は、項目の名称および形式とドメイン知識とに基づいて判定できる程度の項目の性質を判定し、設定する。項目の名称および形式とドメイン知識から分析用データの各項目の性質をある程度把握することで分析用データの性質の判定が可能になる。

[0036] なお、項目の性質の判定は、計算機が演算処理により項目の性質を判定し、設定する場合だけでなく、計算機が判定した項目の性質をユーザが修正して設定する場合や、ユーザが項目の性質を判定して設定する場合も含む。予備設定された項目の性質を第１項目性質と呼ぶことにする。

[0037] データ性質判定部１３は、分析用データに含まれる項目の第１項目性質とドメイン知識に基づき、その分析用データの性質であるデータ性質を判定する。データ性質は、様々なデータを複数のカテゴリに分類することを可能にする各データの特徴である。

[0038] なお、データの性質の判定は、計算機が演算処理により判定する場合だけでなく、計算機が判定したデータの性質をユーザが修正して設定する場合や、ユーザが一部のデータの性質を判定して設定する場合も含む。

[0039] 項目性質判定部１４は、データ性質判定部１３により判定された分析用データのデータ性質に基づき、分析用データに含まれる各項目の性質を判定す

る。その際、項目性質判定部１４は、項目性質予備設定部１２により予備設定された項目の性質あるいはメタ情報を利用してもよい。ここでの項目の性質の判定は、その項目にどのような値が入るかを判定することを指す。ここで判定された項目の性質を第２項目性質と呼ぶことにする。また項目の性質の判定には、複数の項目間にどのような関係があるかを判定することを含んでも良い。

[0040] 変数候補生成部１５は、分析用データに含まれる項目の第２項目性質に基づいて、項目から選択または項目を加工することより説明変数候補を生成する。

[0041] 以上説明したように、本実施形態では、分析用データの項目の名称および形式から判定可能な性質を予備設定し、予備設定した項目の性質に基づいて分析用データの性質を判定し、分析用データの性質に基づいて項目の性質を本格的に判定し、その判定した項目の性質に基づいて説明変数候補を生成するというように、ステップバイステップの手順により説明変数候補を生成する。そのため、分析用データ内に判断材料がないため機械的な判断が困難であった各項目の性質を判断することが可能となり、モデル生成において説明変数候補の生成に要する人手の作業工数を低減することができる。その結果、データ分析における予測モデルの変数の絞り込みを迅速に行うことができる。

[0042] また、本実施形態において、第１項目性質は、特定の項目の特徴を表す特定項目特徴であり、データ性質は、分析用データをデータ構造により分類するデータ構造分類であり、第２項目性質は、項目の内容的な特徴を表す項目内容特徴である。このように、まず特定の項目の特徴を判定し、特定の項目の特徴から把握できるデータ構造でデータを分類し、そのデータに含まれる項目の内容的な特徴をデータのデータ構造により判定するというようにステップバイステップで項目の内容を解析していくことで項目の内容的な特徴が把握できるので、説明変数候補を適切に抽出することが可能となる。

[0043] また、項目性質予備設定部１２は、項目の形式に基づいて、特定項目特徴

を備えた項目を判定する。ここでいう項目の形式にはメタ情報に記載された項目名も含まれる。

[0044] 特定の項目には、分析対象を識別する項目、分析対象を識別する項目以外の項目であり分析対象でないものを識別する項目、およびエントリが登録された時期を示す項目が含まれる。分析対象の識別情報が格納された項目、他の何らかのものの識別情報が格納された項目、エントリが登録された時期を示す項目を把握できると、データ構造の推定が可能となり、データの分類を進めることができる。

[0045] また、データ性質判定部１３は、特定項目特徴とドメイン知識に基づいて分析用データのデータ構造分類を判定する。

[0046] その場合、データ構造分類には、予め与えられたデータであるマスターデータ、所定の事象の発生をトリガに取得されたデータであるトランザクションデータ、所定の時刻になったことをトリガに取得されたデータであるスナップショットデータが含まれる。マスターデータ、トランザクションデータ、スナップショットデータというデータ構造が分かると、分析用データに含まれる項目の推定を深めることが可能となり、項目の性質の判定の精度が向上する。

[0047] また、項目性質判定部１４は、項目の形式、データのデータ構造分類、およびドメイン知識に基づいて、項目の項目内容特徴を判定する。ここでいう項目の形式にもメタ情報に記載された項目名が含まれる。より具体的には、項目性質判定部１４は、項目の形式に基づいて項目を大分類に分類し、データ構造分類およびドメイン知識に基づいて、項目を、項目内容特徴を表す小分類に更に分類する。項目を形式により大分類に分類してから内容的な特徴を示す小分類に分類するので、項目の内容的な特徴を容易に判定することができる。

[0048] さらに、項目性質判定部１４は、データ構造分類およびドメイン知識に加えて更に項目間の関係性に基づいて、項目を小分類に更に分類する。項目間の関係性を項目の内容的な性質の推定に利用するので、項目の内容的な特徴

の推定を更に深めることができる。

[0049] 図３は、本実施形態によるモデル変数候補生成処理の全体の概要フローチャートである。以下、モデル変数候補生成装置１０が実行するモデル変数候補生成処理について詳細に説明する。

[0050] モデル変数候補生成装置１０には、分析目的情報と、分析用データおよびそのメタ情報と、ドメイン知識情報とが入力される。

[0051] 分析目的情報には、分析用データの分析方針を定める要素が含まれている。具体的には、分析対象ＩＤ（識別情報）、目的変数、および基準日が含まれる。

[0052] 分析対象ＩＤは、分析対象単位を個々に識別する識別情報である。例えば、クレジットカード加入者というような人の単位で分析を行い、残高不足による引き落とし不能など何らかの事象の発生を予測するとすれば、分析対象単位はクレジットカード加入者という人である。その場合、分析対象ＩＤは各人に対して設定される。また、取引の単位で分析を行い、何らかの予測をするとすれば、分析対象単位は取引である。その場合、分析対象ＩＤは各取引に対して設定される。

[0053] 分析対象ＩＤをモデル変数候補生成装置１０に与える形式は特に限定されない。例えば、複数の分析対象についての分析対象ＩＤをリスト化したファイルをモデル変数候補生成装置１０に入力することにしてもよい。あるいは、モデル変数候補生成装置１０に入力される、ある分析用データのある項目に分析対象ＩＤが記録されているというように指定することにしてもよい。また、分析対象ＩＤは、複数の情報の組み合わせにより表現されてもよい。

[0054] 目的変数は、予測モデルにより予測したい値を表す変数である。例えば、残高不足による引き落とし不能という事象の発生を予測するのであれば、引き落とし不能の発生の有無を目的変数に設定することが考えられる。目的変数は分析対象に対して値が一意に定まるように設定する。

[0055] 基準日は、どの日までのデータを、予測モデルの生成や予測モデルを用いた予測対象の状態の予測など、分析に用いるかを設定する日付である。例え

ば、ある基準日までのデータを分析用データとして用いて予測モデルを生成し、基準日を設定しなおし、新たな基準日までのデータから説明変数に値を入力することにより、目的変数を予測結果として算出することができる。

[0056] 例えば、ある基準日までに取得された分析用データを用い、引き落とし不能という事象の発生の確率を目的変数として出力する予測モデルを生成し、その予測モデルに新たな基準日までのデータから説明変数の値を取得し、予測モデルに入力することにより、引き落とし不能という事象の発生する確率を算出することにしてもよい。

[0057] 基準日は予測対象に対して一意に定まるように設定する。なお、現在時点までに取得したデータを予測モデルに入力して事象を予測するのであれば、現在時点が基準日となり、取得済の全てのデータを予測モデルの生成に用いればよいので、基準日を設定する必要はない。

[0058] 基準日をモデル変数候補生成装置１０に与える形式は特に限定されない。例えば、それぞれの予測対象に対する基準日をリスト化したファイルをモデル変数候補生成装置１０に入力することにしてもよい。あるいは基準日の算出方法（データの加工方法）をモデル変数候補生成装置１０に指定することにしてもよい。具体的には、ある分析用データのある項目に登録された年月の月末の日付を基準日とするというように指定してもよい。

[0059] 分析用データは分析に用いられる各種データである。本実施形態では、分析用データとして、主にＣＳＶ形式など行と列で指定された形式のデータを想定する。行はエントリに対応し、列は項目に対応する。分析用データにはメタ情報が含まれていてもよい。メタ情報には、例えば、各項目の項目名、各項目の説明、項目間の関連性を表わす情報、データフォーマット、文字列長などが記述されていてもよい。モデル変数候補生成装置１０には複数の分析用データが入力されてもよい。ただし、分析用データの形式や構造は特に問わない。分析用データはＷｅｂページのＨＴＭＬファイルのように明確な構造が定まっていないデータであっても、項目の抽出が可能であればよい。

[0060] ドメイン知識情報はドメイン知識を蓄積した情報である。ドメイン知識は

、モデル生成システム９０により生成した予測モデルを適用する対象の領域（対象領域）の専門家が知得している事柄であり、例えば、対象領域において特有な事柄が含まれる。ドメイン知識を前提とすることでその対象領域での具体的な推論や判断が可能となる。本実施形態では分析用データに関するドメイン知識の情報がモデル変数候補生成装置１０に与えられる。ドメイン知識として、分析用データの項目に格納されうる値の属性および／または分布傾向など項目個別の性質を示す項目個別知識と、項目間の関係性を示す項目間知識とが含まれる。

[0061] ドメイン知識の項目個別知識の例として、融資希望者の年齢の分布、クレジットカードの利用限度額の分布などがある。融資希望者の年齢分布がドメイン知識として予め分かっていれば、項目性質予備設定部１２あるいは項目性質判定部１４は、ある項目に格納されている値の分布がその年齢分布と類似したとき、その項目には融資希望者の年齢が格納されていると推定することが可能となる。

[0062] ドメイン知識の項目個別知識の他の例として、項目の名称とその項目の性質との関係などがある。対象領域において、ある項目名の項目に入る値は特定の性質を示すことが予め分かっていれば、項目名から項目の性質を容易に推定することができる。例えば、項目名が「消化率」という項目は０～１００％の値をとることが分かっていれば、「消化率」という項目名から項目に入る値の範囲を容易に推定することができる。

[0063] ドメイン知識の項目間知識の例として、データ構造の傾向がある。対象領域において用いられる場合の多いデータ構造が予め分かっていれば、その傾向を利用してデータ構造を推定し、ある項目の性質から同じ分析用データに含まれる他の項目の性質を推定することができる。

[0064] ドメイン知識の項目間知識の他の例として、数値が入る２つの項目がありそれらの項目の数値の大小関係が予め分かっていれば、分析用データに数値の形式を有する２つの項目がある場合に数値の大小関係に基づきそれぞれの項目を特定することができる。

[0065] モデル変数候補生成装置１０において、項目性質予備設定部１２、データ性質判定部１３、項目性質判定部１４、あるいは変数候補生成部１５はドメイン知識を利用することで各判定の精度を向上させることができる。ドメイン知識は対象とする領域により様々である。各種領域のドメイン知識をモデル変数候補生成装置１０に予め蓄積しておき、予測モデルを生成する担当者がその予測モデルの対象領域のドメイン知識を指定して利用することにしてもよい。

[0066] ＜項目判定処理１＞

図３を参照すると、まず、項目性質予備設定部１２は、分析用データの各項目を対象として項目判定処理１を実行する（ステップＳ１０１）。項目判定処理１により、分析用データの性質の判定が可能となる程度の分析用データの各項目の性質を判定することができる。項目判定処理１では、各項目がどのような性質の値が入る項目かという項目性質の判定と共に、その判定がどの程度の確信度であるかが判定される。

[0067] 本実施形態では、項目性質予備設定部１２は、対象項目の名称および形式に基づいて、対象項目が、分析対象ＩＤ、他のＩＤ、登録時期（登録日時あるいは登録日）を含む所定の事柄を示す値が入る項目であるか否か判定する。

[0068] 分析対象ＩＤは、分析対象を個々に識別する識別情報である。他のＩＤは、分析対象以外の対象を個々に識別する識別情報である。登録時期（登録日時あるいは登録日）は、分析対象データにエントリを追加登録した日時あるいは日である。分析対象データに含まれる項目のうち分析対象ＩＤ、他のＩＤ、および登録時期が分かると、分析対象データの性質の判定が可能となる。

[0069] 図４は、対象項目がＩＤの入る項目であるか否か判断する処理のフローチャートである。

[0070] まず、項目性質予備設定部１２は、対象項目に入っている値と、分析目的情報として予め与えられている分析対象のＩＤがとる値とを対照する（ステ

ップＳ２０１）。対象項目の値と予め与えられている分析対象のＩＤの値とに一定以上の同一性があれば、項目性質予備設定部１２は、対象項目を分析対象ＩＤの項目と推定する（ステップＳ２０９）。

[0071] 対象項目の値と予め与えられている分析対象のＩＤの値とに一定以上の同一性がなければ、項目性質予備設定部１２は、次に、対象項目に主キー制約が付与されているか否か判定する（ステップＳ２０２）。対象項目に主キー制約が付与されていれば、項目性質予備設定部１２は、対象項目はＩＤが入る項目であると推定する（ステップＳ２０７）。

[0072] 対象項目に主キーが付与されていなければ、項目性質予備設定部１２は、次に、メタ情報の項目名に「ＩＤ」という文字列が含まれているか否か判定する（ステップＳ２０３）。メタ情報の項目名に「ＩＤ」という文字列が含まれていれば、項目性質予備設定部１２は、対象項目はＩＤが入る項目であると推定する（ステップＳ２０７）。

[0073] メタ情報の項目名に「ＩＤ」という文字列が含まれていなければ、項目性質予備設定部１２は、次に、対象項目に入っている値が連番になっているか否か判定する（ステップＳ２０４）。対象項目に入っている値が連番になっていれば、項目性質予備設定部１２は、対象項目はＩＤが入る項目であると推定する（ステップＳ２０７）。

[0074] 対象項目に入っている値が連番になっていなければ、項目性質予備設定部１２は、対象項目に入っている値がハッシュ化されているか否か判定する（ステップＳ２０５）。ハッシュ化されている項目はＩＤである可能性が高いからである。項目の値の文字列が０～９およびＡ～Ｆという１６進数字を示す文字のみで表現されており、文字列内の各位置における各１６進数字の出現率が等しければ、その項目の値はハッシュ化されていると推定できる。対象項目に入っている値がハッシュ化されていれば、項目性質予備設定部１２は、対象項目はＩＤが入る項目であると推定する（ステップＳ２０７）。

[0075] 対象項目に入っている値がハッシュ化されていなければ、項目性質予備設定部１２は、次に、対象項目に類似する項目が他の分析用データにあるか否

か判定する（ステップＳ２０６）。項目同士の類似は、文字列長、使われている文字、登場する文字列、などを比較することで判定することができる。ＩＤは分析用データを他の分析用データと結合するキーとなる場合が多く、その場合、そのＩＤは複数の分析用データに共通に存在することとなる。よって、複数の分析用データに共通に存在する項目はＩＤである可能性が高いと言える。対象項目に類似する項目が他の分析用データにあれば、項目性質予備設定部１２は、対象項目はＩＤが入る項目であると推定する（ステップＳ２０７）。

[0076] 対象項目がステップＳ２０６にて対象項目がＩＤと推定された場合、項目性質予備設定部１２は、対象項目に入るＩＤは分析対象のＩＤであるか否か判定する（ステップＳ２０８）。例えば、対象項目の項目名に分析対象を表わす文字列が含まれていたら、対象項目に入るＩＤは分析対象のＩＤであると推定できる。分析対象を表わす文字列には、分析対象の名称、略称、頭文字、英訳などが含まれる。

[0077] 対象項目に入るＩＤが分析対象のＩＤであれば、項目性質予備設定部１２は、対象項目を分析対象ＩＤの項目と推定する（ステップＳ２０９）。対象項目の入るＩＤが分析対象のＩＤでなければ、項目性質予備設定部１２は、対象項目を分析対象以外のもののＩＤであると推定する（ステップＳ２１０）。

[0078] 以上により対象項目が分析対象ＩＤあるいは他のＩＤであるか否かを推定することができる。本実施形態では、項目性質予備設定部１２は、更に、対象項目が分析対象ＩＤあるいは他のＩＤであることの確信度を推定する。

[0079] 項目性質予備設定部１２は、図４のステップＳ２０１～Ｓ２０６の各ステップに予め確信度を付与しておく。ここでは一例として、ステップＳ２０１に確信度Ａ、ステップＳ２０２に確信度Ｂ、ステップＳ２０３に確信度Ｃ、ステップＳ２０４に確信度Ｄ、ステップＳ２０５に確信度Ｅ、ステップＳ２０６に確信度Ｆをそれぞれ付与しておく。そして、項目性質予備設定部１２は、図４の対象項目がＩＤの入る項目であるか否か判断する処理において、

分析対象がＩＤであると判定したとき、ステップＳ２０１～Ｓ２０６のどのステップでＹＥＳとなったかにより確信度を決定する。例えば、対象項目に主キー制約が付与されていれば、ステップＳ２０２でＹＥＳとなり対象項目はＩＤの入る項目と判定されるので、その判定は確信度Ｂとなる。また、対象項目に主キー制約は付与されていないが、対象項目の項目名に「ＩＤ」の文字列が含まれていれば、ステップＳ２０３でＹＥＳとなり対象項目はＩＤの入る項目と判定されるので、その判定は確信度Ｃとなる。

[0080] なお、ここでは、図４に示したように、ステップＳ２０１～Ｓ２０６の各判定処理は、前段ステップの判定結果がＮＯの場合に順次、次段ステップの判定を実行する例を示したが、本発明がこれに限定されることは無い。

[0081] 他の例として、対象項目に対してステップＳ２０１～Ｓ２０６に相当する全ての判定処理を行い、それら全ての判定結果を用いて、対象項目がＩＤの入る項目か否かを総合的に判定することにしてもよい。その場合、ステップＳ２０１～Ｓ２０６に相当する判定処理のそれぞれにスコアを付与しておき、ＹＥＳに該当した判定処理のスコアの合計値を、対象項目がＩＤの入る項目である確信度としてもよい。

[0082] 次に、対象項目が登録日の入る項目か否か判定する処理について説明する。

[0083] 図５は、対象項目が登録日の入る項目であるか否か判断する処理のフローチャートである。

[0084] まず、項目性質予備設定部１２は、対象項目のフォーマットが日付のフォーマットであるか否か判定する（ステップＳ３０１）。対象項目のフォーマットが日付のフォーマットであれば、項目性質予備設定部１２は、次に、対象項目に入っている日付の分布が所定の分布と類似するか否か判定する（ステップＳ３０２）。ステップＳ３０２は、ドメイン知識として予め分布が分かっている日付の項目を除外する処理である。例えばドメイン知識として分析対象の生年月日の分布が予め分かっていれば、その生年月日の分布と類似する日付の項目は登録日ではないと推定する。

[0085] 対象項目の日付の分布が除外すべき日付の項目の分布と類似しなければ、項目性質予備設定部１２は、対象項目の日付が時系列であるか否か判定する（ステップＳ３０３）。対象項目の日付が時系列であれば、項目性質予備設定部１２は、対象項目が登録日の入る項目であると推定する（ステップＳ３０４）。

[0086] なお、ここでは、対象項目が登録日の入る項目であるか否か判定する処理の一例を示したが、他の項目についても同様に判定することができる。また、ここでは、ステップＳ３０２にて、予め分布が分かっている目的外の日付の項目を除外することにしたが、これに限定されることはない。ドメイン知識として目的の日付の項目の分布が予めわかっていれば、対象項目の日付の分布と目的の日付の項目の分布とを比較し、それらが類似すれば、対象項目は目的の日付の入る項目であると推定することができる。

[0087] なお、対象項目について複数の可能性を提示することにしてもよい。例えば、対象項目が、ある確信度で分析対象ＩＤである可能性があり、かつ、ある確信度で他のＩＤである可能性があるというような推定を行うことにしても良い。

[0088] 以上の処理により、分析用データの項目の中で、分析対象ＩＤ、他のＩＤ、および登録時期の項目が分かると、次に説明する分析対象データの性質の判定処理が可能となる。

[0089] ＜データ判定処理＞

図３に戻り、続いて、データ性質判定部１３は、ステップＳ１０２にて、ステップＳ１０１で予備設定（判定）された項目の性質とドメイン知識とに基づいて、分析用データを対象とし、分析用データとしての性質を判定する。本実施形態では分析用データの性質判定の具体的処理として分析用データのデータ構造を判定する。データ構造として、マスターデータ、トランザクションデータ、スナップショットデータという区分がある。それ以外の区分があっても良い。マスターデータ、トランザクションデータ、およびスナップショットデータにはそれぞれ特徴があり、その特徴を利用することにより

、それぞれを判別することが可能である。また、データ性質判定部１３は、どのキーを用いると複数の分析用データのエントリを紐付けることができるかも判定する。

[0090]（マスターデータ）

マスターデータとは、ある対象に関する固定的あるいは半固定的な情報を登録したデータである。マスターデータの例としてクレジットカードの加入者の各種属性を登録したリストデータがある。

[0091] 図６は、マスターデータの一例を示す図である。図６には、ある対象に対して１つずつエントリが登録されるという特徴を有する１対１対応型のマスターデータの例が示されている。図５を参照すると、マスターデータの各エントリに対象に関する識別情報（ＩＤ）および属性情報が登録されている。図５の例では、対象は例えばクレジットカード保有者であり、クレジットカード保有者の固定的な属性情報として性別と生年月日が登録され、半固定的な属性情報として収入と既婚／未婚の属性が登録されている。ＩＤが１回ずつ登場する分析用データは１対１対応型のマスターデータである可能性がある。

[0092] 図７は、マスターデータの他の例を示す図である。図７には、ある対象に対して複数のエントリが登録される１対Ｎ対応型のマスターデータの例が示されている。図７を参照すると、マスターデータの各エントリに対象に関連する識別情報（ＩＤ）およびその他の情報が登録されている。図７の例では、対象は図６の例と同じくクレジットカード保有者であるが、マスターデータにはクレジットカード保有者本人ではなくその家族の情報が登録されている。そのため、対象であるクレジットカード保有者のＩＤが同じ１つ以上のエントリ（家族毎のエントリ）を個々に識別するＩＤとして家族番号（Ｎｏ．）がある。クレジットカード保有者に家族がいる場合といない場合とがある。

[0093] 例えば、図６を参照するとＩＤ＝Ｃ００１のクレジットカード保有者は男性であり、図７を参照するとそのＩＤ＝Ｃ００１のクレジットカード保有者

には配偶者と１人の子供がいる。配偶者は家族Ｎｏ．＝１であり、女性であり、収入が５０万円である。子供は家族Ｎｏ．＝２であり、男性であり、収入の登録が無い。収入の登録が無い場合には収入の情報が取得されていない場合と収入が無い場合とが含まれる。また、図６を参照すると、ＩＤ＝Ｃ００４というクレジットカード保有者が登録されているが、図７を参照すると、そのクレジットカード保有者には家族がいない。

[0094]（トランザクションデータ）

トランザクションデータは何らかの事象が発生したことをトリガに取得されるデータである。通常、トランザクションデータには事象の発生時期（登録日時あるいは発生日）の項目がある。この事象の発生時期（発生日時あるいは発生日）が登録時期（登録日時あるいは登録日）に相当する。

[0095] 図８は、トランザクションデータの一例を示す図である。図８には、ある対象に対して所定の事象が発生したことをトリガに取得されたトランザクションデータの例が示されている。図８を参照すると、トランザクションデータの各エントリに対象のＩＤおよび事象の発生日付とその他の情報が登録されている。図８の例では、対象はクレジットカード保有者であり、クレジットカードの利用という事象をトリガとして取得された事象の属性情報が登録されている。事象の属性情報には利用店業種と利用商品数と利用金額が含まれている。利用店業種はクレジットカードを利用した店舗の業種を示す情報である。利用商品数はクレジットカードで代金支払いをした商品の個数を示す情報である。利用金額はクレジットカードで支払った代金の情報である。トランザクションデータには、対象のＩＤと事象発生日（登録日）とが含まれ、同じＩＤが複数回登場することがあり、時系列な事象発生日の順番通りにエントリが登録されていることが多い特徴がある。ただし、エントリがソートしなおされ、発生日の順番通りでなくなっている可能性もある。

[0096] 例えば、図８の一番上のエントリを参照すると、ＩＤ＝Ｃ００１のクレジットカード保有者が２０１６年１２月１日に飲食店で６点の商品の代金２００００円を支払うのにクレジットカードを利用したというエントリが登録さ

れている。

[0097]（スナップショット）

スナップショットデータは所定の時刻になったことをトリガに取得されたデータである。通常、スナップショットデータには一定時間間隔で取得されるという特徴がある。また、多くの場合、所定時刻になると、複数の対象について同時期にデータが取得されるというのもスナップショットデータの特徴である。そのため、スナップショットデータには、規則性のある時刻に繰り返しエントリが追加される。また、同時期に複数のエントリが追加される。

[0098] 図９は、スナップショットデータの一例を示す図である。図９には、２０１６年１２月３１日になったことをトリガに各対象について取得された対象に関する属性情報を格納したスナップショットデータの例が示されている。具体的には、各クレジットカード保有者の年齢、収入、既婚／未婚の属性情報が取得されている。

[0099] データ性質判定部１３は、分析用データの予備設定された項目の性質とドメイン知識に基づく各データ構造の特徴とを利用して分析用データのデータ構造を判定する。その際、データ性質判定部１３は、分析用データの項目の予備設定された性質として、項目性質予備設定部１２が出力した項目の性質の情報を人間が補正したものを用いることにしてもよい。

[0100]（データ構造判定処理）

データ構造を判定する処理の方法は特に限定されないが、ここでは幾つか例示する。

[0101] データ性質判定部１３は、一例として機械学習により、分析用データと予備設定されたその各項目の性質とを説明変数とし、データ構造を目的変数とするデータ構造判定用モデルを構築しておき、データ性質判定部１３は、そのモデルを用いて分析用データのデータ構造を判定することにしても良い。データ構造判定用モデルは、学習用のデータに対してそのデータがどのデータ構造を有するかの正解を付与し、機械学習を行うことにより構築すること

ができる。機械学習において、予備設定された各項目の性質について判定の確信度を考慮した演算を行なっても良い。また、状況に応じて適切な判断基準が変化することがドメイン知識として予め分かっていれば、機械学習において、ドメイン知識に従い、状況に応じて判断基準を変化させても良い。例えば、状況毎にモデルを構築しておき、状況に応じてモデルを切り替えて用いても良い。

[0102] 他の例として、人間の経験則を用いてデータ構造の判定ルールを作成しておき、データ性質判定部１３は、そのルールを用いて分析用データのデータ構造を判定することにしても良い。ルール作成において、予備設定された各項目の性質について判定の確信度を考慮してもよい。また、状況に応じて適切な判断基準が変化することがドメイン知識として予め分かっていれば、ルールにおいて、ドメイン知識に従い、状況に応じて判断基準を変化させても良い。例えば、状況毎にルールを構築しておき、状況に応じてルールを切り替えて用いても良い。

[0103] また、上述した機械学習によるモデル構築と人間の経験則のルール化とを組み合わせて用いてもよい。例えば、経験則により確信度の高いルールの作成が可能な部分はルール化で対応し、経験則によるルール化が困難な部分を機械学習によりモデルを構築しておき、データ性質判定部１３はルールとモデルを適宜切り替えて用いることにしても良い。

[0104] また、データ性質判定部１３は、どのキーを用いると複数の分析用データのエントリを紐付けることができるかについても判定することにしてもよい。

[0105] 以上説明したデータ判定の処理結果により分析用データの各項目の性質の判定が可能になる。

[0106] ＜項目判定処理２＞

図３に戻り、続いて、ステップＳ１０３にて項目性質判定部１４は、データ性質判定部１３が判定した分析用データの性質に基づいて、分析用データの各項目を対象とし、その各項目の性質を判定する。

[0107] 分析用データの項目の性質の判定では、項目性質判定部１４は、まず各項目をその形式に応じて大分類に分類する。例えば、数値、カテゴリ、日時、日付、などに分類する。更に、項目性質判定部１４は、分析用データの性質およびドメイン知識に基づいて、項目を小分類に分類する。その際、ドメイン知識として、どのような項目がどのデータ構造の分析用データに含まれる可能性があるか、その項目に入る値がどのような形式か、その項目に入りうるのはどのような範囲の値か、その項目の値はどのような分布か、他の項目の値とどのような大小関係か等が利用される。これらドメイン知識を利用すれば、データ構造、値の形式、および値の分布により、項目の分類を絞り込むことができる。

[0108] 例えば、項目が数値を示すものである場合、その項目を更に、例えば、年齢、金額、件数、比率などに細かく分類する。項目がカテゴリを示すものである場合、更に、例えば、ＩＤ、区分レベル（大区分、中区分、小区分など）などに細かく分類する。項目が日時あるいは日付を示すものである場合、更に、生年月日、有効期限などに細かく分類する。ここに示した分類は例であり、他の分類があっても良い。

[0109] また、項目同士の関係性から項目の性質が判明することもある。そのため、本実施形態では、項目性質判定部１４は、分析用データの項目の単独の性質に加え項目同士の関係性を項目の性質の判定に利用する。その際、項目同士の関係性をドメイン知識と共に項目の性質の判定に利用するのが有効である。

[0110] 例えば、日付のフォーマットを有する２つの項目があり、そのうち一方が登録日を表わす日付で他方が有効期限を表わす日付であると推定されるが、どちらが登録日でどちらが有効期限か判断できていないとき、ドメイン知識として登録日と有効期限の間に特定の大小関係が分かっていれば、その大小関係を頼りにどちらが登録日でどちらが有効期限かを判定することができる。

[0111] また、限度額は残高以下であるという大小関係がドメイン知識として予め

分かっていることが考えられる。また、カテゴリを示す項目同士であれば、大区分と中区分のように階層的な包含関係が成立することがドメイン知識として分かっていることが考えられる。これらのドメイン知識も項目の性質を判定するのに利用できる。

[0112] ドメイン知識に基づいて小分類毎に出現頻度のベースとなる分布が予め分かっている場合も考えられる。その場合、分析用データの項目に含まれる値の分布をそのベース分布と比較し、類似度に基づいて項目の分類を絞り込む。例えば、クレジットカードの有効期限の分布は、年、季節、月、日、などにより偏りが少なく、かつ、現在の日付から所定年以上未来の日付は存在しないような所定の分布となる。

[0113] また、２つの小分類の間で値に大小関係があることが予め分かっている場合がある。例えば、クレジットカードの有効期限はクレジットカードの発行日よりも遅いことが予め分かっている。ドメイン知識に基づいて小分類間のベースとなる値の大小関係を知得しておき、項目同士の値の大小関係をベースの大小関係と照合することで分類を絞り込むことができる。例えば、２つの項目のどちらか一方がクレジットカードの有効期限で他方がクレジットカードの発行日であるというように、同じ複数の小分類まで絞り込まれた項目が複数存在し、更に絞り込みたい場合を想定する。そのような場合、クレジットカードの有効期限の方がクレジットカードの発行日よりも遅いので、クレジットカードの有効期限の項目の方がクレジットカードの発行日の項目より大きな値をとる。このように項目同士の値の大小関係から、さらにそれぞれの項目の可能性を絞り込むことができる。

[0114] その他、項目の値の形式に応じて適切なドメイン知識を選択して用いることで項目の分類を詳細に絞り込むことができる。

[0115] また本実施形態では項目性質判定部１４は更に項目の性質の判定の確信度を算出する。例えば、項目性質判定部１４は、項目の分類の判定条件にそれぞれ確信度のスコアを付与しておき、当てはまった判定条件のスコアを積算して確信度を算出することにしてもよい。

[0116] ＜候補生成処理＞

図３に戻り、次に、変数候補生成部１５は、ステップＳ１０４にて説明変数候補を生成する。

[0117] 変数候補生成部１５は、ステップＳ１０３にて項目性質判定部１４により判定された分析用データの各項目の性質、項目間の関係性、およびドメイン知識に基づいて、所望の目的変数を算出する予測モデルの説明変数となりうる項目を説明変数候補とする。例えば、ドメイン知識として、目的変数に影響することが予め想定されるパラメータに分類された所定の項目（候補項目）があるとする。項目性質判定部１４により判定された項目の性質がその候補項目に該当すれば、その項目は説明変数候補にしておくのがよい。また、予測対象の属性に分類された所定の項目は目的変数に影響する可能性が高いことが予めドメイン知識として分かっているとする。その場合、その予測対象の属性を示すと判定された項目は説明変数候補にするとよい。

[0118] また、ステップＳ１０３における項目性質判定部１４による項目の性質の判定の確信度が所定値以下の項目は説明変数候補としないことにしてもよい。ドメイン知識として目的変数に影響することが想定される項目と判定されても、そうでない可能性も十分に高いのであれば、説明変数候補から除外した方がよいこともあるからである。

[0119] 上述した項目性質判定部１４による項目性質の判定とそれに関連する変数候補生成部１５による説明変数候補の決定の具体的な例（判定例）を以下に列挙する。

[0120] （判定例１）

項目性質判定部１４は、マスターデータに含まれている項目のうち、登場する項目値が所定種類（下限閾値）以上で所定種類（上限閾値）以下の項目を、観測対象の属性を示す観測対象属性項目と推定し、変数候補生成部１５は、観測対象属性項目と推定された項目を説明変数候補とすることにしてもよい。適切な下限閾値および上限閾値は対象領域や分析用データのデータ量などに応じて変わるので、下限閾値および上限閾値は適宜設定可能としても

よい。個数が限定された観測対象の属性を示す項目は観測対象をその性質により大きく分類し、予測モデルの効果的な説明変数となることがあるので、説明変数候補としておくのがよい。例えば、観測対象が人間であれば、マスターデータには年齢や性別といった項目がある可能性が高く、また年齢や性別が予測モデルの説明変数となる場合がある。ただし、このような判定を行うかことが妥当かどうかは対象領域のドメイン知識に照らして事前に判断するのがよい。全ての領域において、マスターデータの項目のうち登場頻度の高い項目を全て変数の候補とすべきとは限らない。

[0121] （判定例２）

項目性質判定部１４は、トランザクションデータに含まれている項目のうち、登場する項目値が所定種類以上で所定種類以下の項目を、事象の属性を示す事象属性項目と推定し、変数候補生成部１５は、事象属性項目と推定された項目を説明変数候補とすることにしてもよい。個数が限定された事象の属性を示す項目は事象をその性質により大きく分類し、予測モデルの効果的な説明変数となることがあるので、本例のように説明変数候補としておくのがよい場合がある。なお、事象属性項目には、事象の原因、内容、結果、または事象が発生したときの観測対象の状態の少なくともいずれか１つを示す項目が含まれてもよい。事象の原因、内容、結果、または事象が発生したときの観測対象の状態は事象を分類するので、事象と観測対象の関わりが強い場合に効果的な説明関数となり得る。

[0122] （判定例３）

項目性質判定部１４は、スナップショットデータに含まれる項目のうち、登場する項目値が所定種類以上で所定種類以下の項目は、観測対象の状態を示す観測対象状態項目と推定し、変数候補生成部１５は、観測対象状態項目と推定された項目を説明変数候補とすることにしてもよい。個数が限定された観測対象の状態を示す項目は観測対象の状態をその性質により大きく分類し、予測モデルの効果的な説明変数となることがあるので、本例のように、説明変数候補としておくのがよい場合がある。

[0123] （判定例４）

項目性質判定部１４は、マスターデータである分析用データの項目が単独で示す項目性質である項目個別性質と、項目同士の相互関係を示す項目性質である項目間性質とを判定し、変数候補生成部１５は、データ性質と項目個別性質と項目間性質とに基づいて前記項目を加工することにより前記説明変数候補を生成することにしてもよい。各項目について単独の性質および相互関係を判断し、それらを総合判断して説明変数の候補を生成するので、各項目が有する様々な性質から適切に説明変数の候補を列挙することができる。

[0124] （判定例５）

項目性質判定部１４は、マスターデータである分析用データに含まれる項目のうちから、その分析用データの各エントリを一意に特定するキー項目と、値に順序性のない区分を示す非順序カテゴリ項目と、値に順序性をある区分を示す順序カテゴリ項目とを抽出し、それら項目の属性を項目個別性質とし、変数候補生成部１５は、項目個別性質を説明変数候補の生成に利用することにしてもよい。値に順序性のない区分とは、値の大小に特別な意味を持たない区分である。非順序カテゴリ項目の例として住所の番地がある。値に順序性のある区分とは、値の大小に特別な意味を持つ区分である。順序カテゴリ項目の例として年齢がある。変数候補生成部１５は、例えば、特定キー項目に該当する項目を説明変数候補とすることにしてもよい。また、非順序カテゴリ項目と順序カテゴリ項目の一方あるいは両方を説明変数候補とすることにしてもよい。

[0125] （判定例６）

項目性質予備設定部１２と項目性質判定部１４の両方または一方は、既知項目の値の分布を示す既知項目分布情報が予め与えられ、分析用データにおいて判定対象とする項目である判定対象項目の値の分布と既知項目分布情報における既知項目の値の分布とを比較することにより、判定対象項目の項目性質を判定することにしてもよい。例えば、一般的な年齢分布と類似する分布を有する項目は年齢の項目であると推定ができる。その他に、カード限度

額の項目なども推定できる。

[0126] （判定例７）

項目性質判定部１４は、エントリ数が、データ構造に応じて定まる所定閾値以上の分析用データにおいて、項目値が所定種類以上の項目を候補項目として抽出し、変数候補生成部１５は、候補項目に該当した項目を説明変数候補とすることにしてもよい。エントリ数が多ければ登場する値の種類数が多い項目でも説明変数として目的変数の精度向上に貢献する場合があるので、そのような項目を説明変数候補とするとよい場合がある。なお、項目の性質により、項目が有効な説明変数となるであろうエントリ数がある程度想定できる場合があるので、対象領域とデータ構造に対して適切な閾値を設定すれば、十分なエントリ数がある項目を適切に選択することが可能となる。ただし、項目値の種類がエントリ数とほぼ同数などのように、エントリ数の割に項目値の種類数が多い項目の場合、その項目は目的変数への影響の傾向が抽出されない可能性があるので、項目性質判定部１４は、エントリ数が、データ構造に応じて定まる所定閾値以上の分析用データにおいて、項目値が所定種類（下限閾値）以上で所定種類（上限閾値）以下の項目を候補項目とし、変数候補生成部１５は、候補項目に該当した項目を説明変数候補とすることにしてもよい。

[0127] （判定例８）

項目性質判定部１４は、所定のデータ構造のデータの項目のうち、項目値が所定数値範囲内の項目を候補項目として抽出し、変数候補生成部１５は、候補項目に該当した項目を説明変数候補とすることしてもよい。説明変数候補としたい項目を予めデータ構造と項目値の範囲とで条件付けしておくことにより、その条件を満たす項目を説明変数候補として抽出することができる。

[0128] 以上、本実施形態では、主に、分析用データの項目から説明変数候補を選択することを想定した例を用いたが、本発明がこれに限定されることはない。変数候補生成部１５は、項目性質判定部１４が判定した各項目の性質を利

用して説明変数候補を生成するものであればよく、項目をそのまま説明変数候補とするものでなく、項目を加工して説明変数候補を生成してもよい。

[0129] 例えば、変数候補生成部１５は、ある変数と他の変数を組み合わせた合成変数を網羅的に作成し、得られた様々な合成変数について、目的変数との関連度合いを算出し、関連度合いが上位所定個の合成変数を説明変数候補とすることしてもよい。また、ある変数と他の変数を組み合わせた合成変数にさらに他の変数を組み合わせた合成変数を網羅的に作成し、上記と同様に目的変数との関連度合いで合成変数を絞り込むことで説明変数候補を生成することにしてもよい。

[0130] 上述した本発明の実施形態は、本発明の説明のための例示であり、本発明の範囲をそれらの実施形態にのみ限定する趣旨ではない。当業者は、本発明の要旨を逸脱することなしに、他の様々な態様で本発明を実施することができる。

符号の説明

[0131] １０…モデル変数候補生成装置、１１…データ入力部、１２…項目性質予備設定部、１３…データ性質判定部、１４…項目性質判定部、１５…変数候補生成部、２０…モデル生成装置、３０…予測装置、９０…モデル生成システム

請求の範囲

[請求項1] 予測モデルの生成において説明変数の候補とする説明変数候補を生成するモデル変数候補生成装置であって、

エントリ毎に１つ以上の項目を含み、前記項目に項目値を有する分析用データを入力するデータ入力部と、

前記分析用データに含まれる前記項目の性質を第１項目性質として予備設定する第１項目判定部と、

前記分析用データに含まれる項目の前記第１項目性質に基づき、該分析用データの性質であるデータ性質を判定するデータ性質判定部と、

前記分析用データの前記データ性質に基づき、該分析用データに含まれる前記項目の性質を第２項目性質として判定する第２項目判定部と、

前記分析用データに含まれる項目の前記第２項目性質に基づいて、前記項目から選択または前記項目を加工することより、前記説明変数候補を生成する変数候補生成部と、

を有するモデル変数候補生成装置。

[請求項2] 前記第１項目性質は、特定の項目の特徴を表す特定項目特徴であり、

前記データ性質は、前記データをデータ構造により分類するデータ構造分類であり、

前記第２項目性質は、前記項目の内容的な特徴を表す項目内容特徴である、

請求項１に記載のモデル変数候補生成装置。

[請求項3] 前記第１項目判定部は、前記項目の形式に基づいて、前記特定項目特徴を備えた項目を判定する、

請求項２に記載のモデル変数候補生成装置。

[請求項4] 前記特定の項目は、分析対象を識別する項目、前記分析対象でない

ものを識別する項目、および前記エントリが登録された時期を示す項目を含む、

請求項３に記載のモデル変数候補生成装置。

[請求項5] 前記データ性質判定部は、前記特定項目特徴とドメイン知識に基づいて前記分析用データのデータ構造分類を判定する、

請求項２に記載のモデル変数候補生成装置。

[請求項6] 前記データ構造分類には、予め与えられたデータであるマスターデータ、所定の事象の発生をトリガに取得されたデータであるトランザクションデータ、所定の時刻になったことをトリガに取得されたデータであるスナップショットデータが含まれる、

請求項５に記載のモデル変数候補生成装置。

[請求項7] 前記第２項目判定部は、前記項目の形式、前記データのデータ構造分類、およびドメイン知識に基づいて、前記項目の前記項目内容特徴を判定する、

請求項２に記載のモデル変数候補生成装置。

[請求項8] 前記第２項目判定部は、前記項目の形式に基づいて前記項目を大分類に分類し、前記データ構造分類および前記ドメイン知識に基づいて、前記項目を、前記項目内容特徴を表す小分類に更に分類する、

請求項７に記載のモデル変数候補生成装置。

[請求項9] 前記第２項目判定部は、前記データ構造分類および前記ドメイン知識に加えて更に前記項目間の関係性に基づいて、前記項目を前記小分類に更に分類する、

請求項８に記載のモデル変数候補生成装置。

[請求項10] 予測モデルの生成において説明変数の候補とする説明変数候補を生成するためのモデル変数候補生成方法であって、

コンピュータが備えるデータ入力手段が、エントリ毎に１つ以上の項目を含み、前記項目に項目値を有する分析用データを入力し、

コンピュータが備える第１項目判定手段が、前記分析用データに含

まれる前記項目の性質を第１項目性質として予備設定し、

コンピュータが備えるデータ性質判定手段が、前記分析用データに含まれる項目の前記第１項目性質に基づき、該分析用データの性質であるデータ性質を判定し、

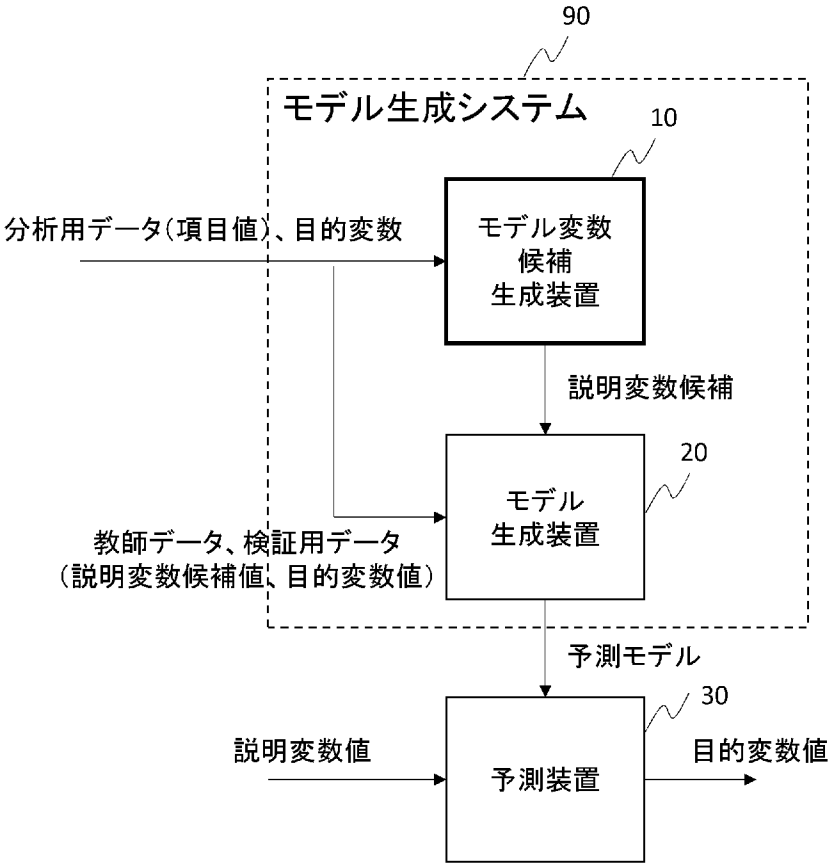
コンピュータが備える第２項目判定手段が、前記分析用データの前記データ性質に基づき、該分析用データに含まれる前記項目の性質を第２項目性質として判定し、

コンピュータが備える変数候補生成手段が、前記分析用データに含まれる項目の前記第２項目性質に基づいて、前記項目から選択または前記項目を加工することより、前記説明変数候補を生成する、

モデル変数候補生成方法。

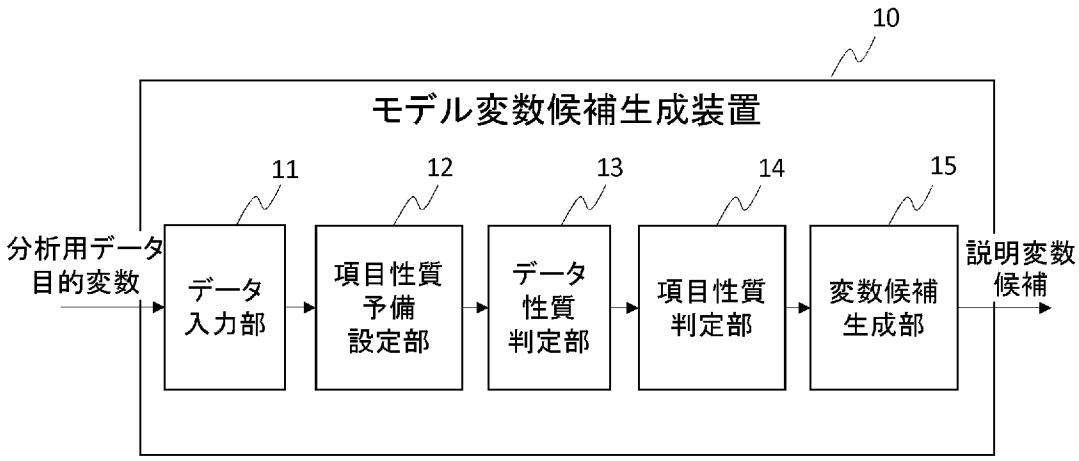
[図1]

図1


90
モデル生成システム
10
分析用データ(項目値)、目的変数
モデル変数
候補
生成装置
説明変数候補
20
モデル
生成装置
教師データ、検証用データ
(説明変数候補値、目的変数値)
予測モデル
30
説明変数値
予測装置
目的変数値

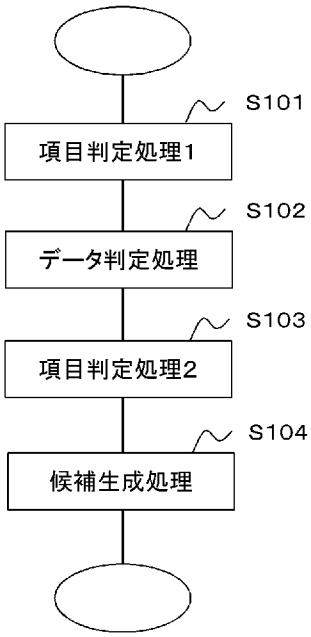
[図2]

図2


10
モデル変数候補生成装置
11
12
13
14
15
分析用データ
目的変数
データ
入力部
項目性質
予備
設定部
データ
性質
判定部
項目性質
判定部
変数候補
生成部
説明変数
候補

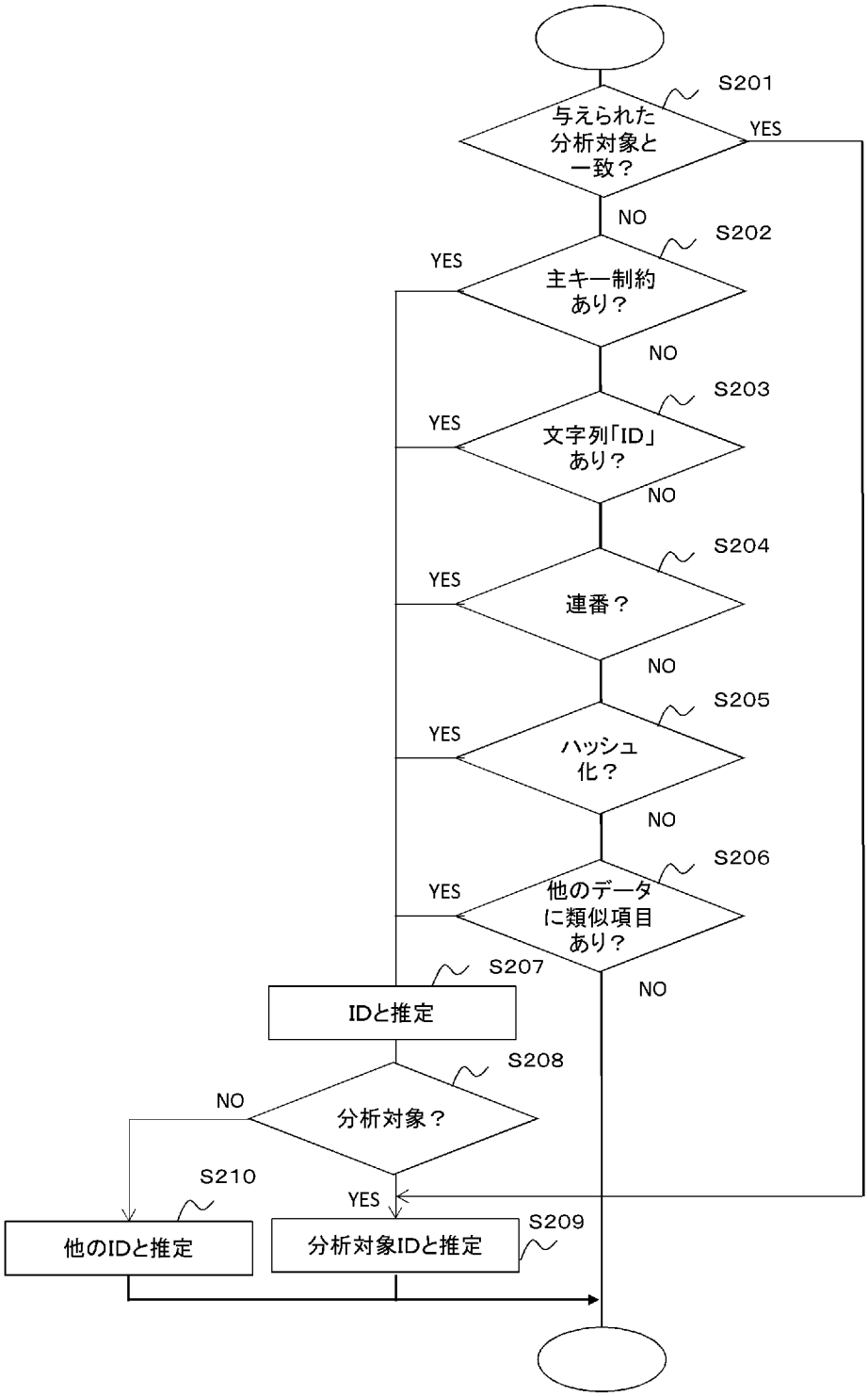
[図3]

図3



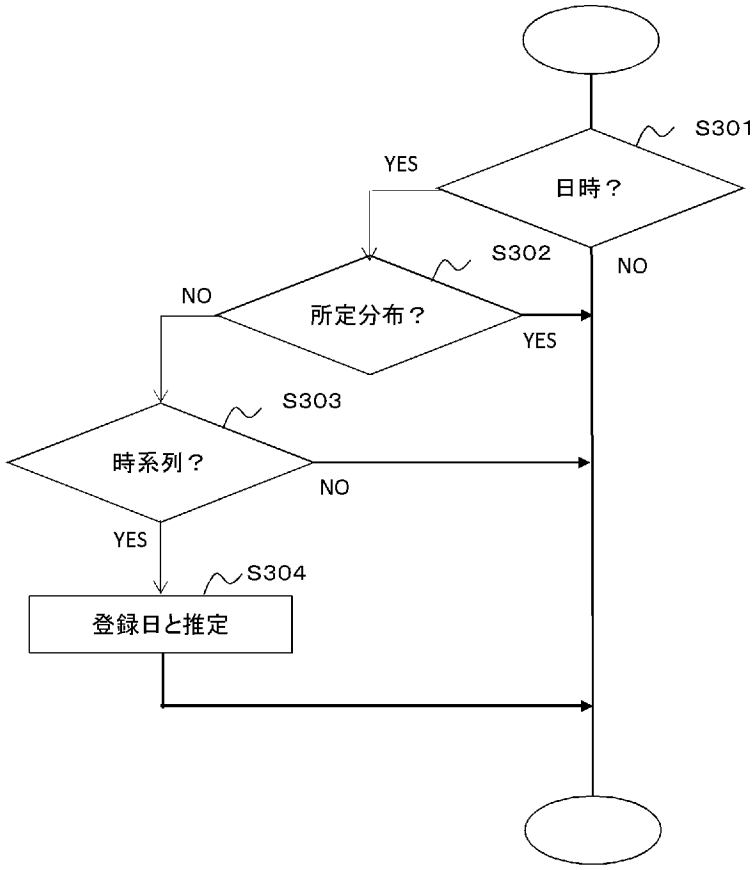
[図4]

図4


S201
与えられた分析対象と一致？
YES
NO
S202
主キー制約あり？
YES
NO
S203
文字列「ID」あり？
YES
NO
S204
連番？
YES
NO
S205
ハッシュ化？
YES
NO
S206
他のデータに類似項目あり？
YES
NO
S207
IDと推定
S208
分析対象？
NO
YES
S209
分析対象IDと推定
S210
他のIDと推定

[図5]

図5



[図6]

図6

ID	性別	生年月日	収入	既婚・未婚
C001	男性	1965/2/8	1200万	既婚
C002	女性	1974/6/9	400万	既婚
C003	男性	1981/12/20	580万	既婚
C004	男性	1976/5/7	600万	未婚
C005	女性	1992/8/7	240万	未婚
C006		1944/6/26	120万	既婚
C007	男性	1956/1/31	600万	既婚
C008	女性	1958/4/15	20万	既婚
C009	女性	1968/10/12	100万	既婚
C010	女性	2000/1/5	5万	未婚

[図7]

図7

ID	家族No	性別	続柄	収入
C001	1	女性	配偶者	50万
C001	2	男性	子供	
C002	1	男性	配偶者	
C003	1	女性	配偶者	
C003	2	女性	子供	
C007	1	女性	配偶者	80万
C008	1	男性	配偶者	700万
C008	2	男性	親	
C008	3	男性	子供	260万
C009	1	男性	配偶者	1000万

[図8]

図8

ID	発生日付	利用店業種	利用商品数	利用金額
C001	2016/12/1	飲食店	6	20,000
C001	2016/12/1	雑貨	4	40,000
C010	2016/12/1	百貨店	4	30,000
C005	2016/12/2	飲食店	2	5,000
C010	2016/12/4	スーパー	2	500
C004	2016/12/6	百貨店	3	50,000
C001	2016/12/7	飲食店	4	15,000
C010	2016/12/7	スーパー	6	5,000
C007	2016/12/8	百貨店	3	15,000
C005	2016/12/9	雑貨	4	4,000
C008	2016/12/16	飲食店	3	3,000
C010	2016/12/16	飲食店	2	2,500
C003	2016/12/17	レジャー	1	20,000
C002	2016/12/20	百貨店	6	60,000
C008	2016/12/20	パチンコ		20,000
C002	2016/12/24	スーパー	7	10,000
C003	2016/12/28	レジャー	1	30,000
C007	2016/12/28	雑貨	4	6,000
C009	2016/12/28	レジャー	1	15,000
C003	2016/12/30	パチンコ		10,000
C010	2016/12/30	雑貨	4	2,000
C003	2016/12/31	スーパー	5	8,000
C007	2016/12/31	スーパー	6	8,000

[図9]

図9

ID	取得日	年令	収入	既婚・未婚
C001	2016/12/31	51	1200万	既婚
C002	2016/12/31	42	400万	既婚
C003	2016/12/31	35	580万	既婚
C004	2016/12/31	40	600万	未婚
C005	2016/12/31	24	240万	未婚
C006	2016/12/31	72	120万	既婚
C007	2016/12/31	60	600万	既婚
C008	2016/12/31	58		既婚
C009	2016/12/31	48	100万	既婚
C010	2016/12/31	16	5万	未婚

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2018/008465

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. G06F17/30(2006.01)i, G06Q10/04(2012.01)i

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. G06F17/30, G06Q10/04

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	1971-2018
Registered utility model specifications of Japan	1996-2018
Published registered utility model applications of Japan	1994-2018

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2002-358411 A (SUMITOMO MITSUI BANKING CORP.) 13 December 2002, entire text, all drawings (Family: none)	1-10
A	JP 2007-329415 A (FUJITSU LIMITED) 20 December 2007, entire text, all drawings & US 2007/0288105 A1	1-10
A	JP 2015-224389 A (NIPPON STEEL & SUMITOMO METAL CORPORATION) 14 December 2015, entire text, all drawings (Family: none)	1-10

☒ 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
31.05.2018	12.06.2018

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2018/008465

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2013-065084 A (FUJITSU LIMITED) 11 April 2013, entire text, all drawings (Family: none)	1-10
A	WO 2010/082322 A1 (HITACHI, LTD.) 22 July 2010, entire text, all drawings & US 2011/0276828 A1	1-10
A	谷岡 日出男, 外２名, 金融業界におけるAI技術, 人工知能学会誌, 01 March 2002, vol. 17, no. 2, pp. 214-221, (TANIOKA, Hideo and 2 other, AI technologies in financial business industry, Journal of Japanese Society for Artificial Intelligence)	1-10
A	藤野 秀則, 因果関係の分析(1)-回帰分析・重回帰分析, ヒューマンインタフェース学会誌, 25 May 2013, vol. 15, no. 2, pp. 141-149, (FUJINO, Hidenori, Analysis of causal relationship (1): regression analysis & multiple regression analysis, Journal of Human Interface Society: human interface)	1-10
A	上田 太一郎, 正準相関分析と回帰分析変数選択基準を用いた判別分析変数選択の試み, 計算機統計学, 31 May 1996, vol. 8, no. 2, pp. 171-175, (UEDA, Taichiro, A function on variable selection in discriminant analysis and its applications, Bulletin of the Computational Statistics of Japan)	1-10
A	藤巻 遼平、 外１名, 分析プロセス自動化・標準化への挑戦, 情報処理学会 デジタルプラクティス, vol. 6, no. 3, 08 November 2016 (received date), pp. 198-206, (FUJIMAKI, Ryohei and 1 other, INFORMATION PROCESSING SOCIETY OF JAPAN, Journal of Digital Practices), non-official translation (Challenge for automation and standardization of analysis process)	1-10

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１８／００８４６５

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i, G06Q10/04(2012.01)i

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））

Int.Cl.　G06F17/30, G06Q10/04

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	１９７１－２０１８年
日本国実用新案登録公報	１９９６－２０１８年
日本国登録実用新案公報	１９９４－２０１８年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2002-358411 A (株式会社三井住友銀行) 2002.12.13, 全文,全図（ファミリーなし）	1-10
A	JP 2007-329415 A (富士通株式会社) 2007.12.20, 全文,全図 & US 2007/0288105 A1	1-10
A	JP 2015-224389 A (新日鐵住金株式会社) 2015.12.14, 全文,全図（ファミリーなし）	1-10

☑　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矛盾するものではなく、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
３１．０５．２０１８	１２．０６．２０１８

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｍ　４７７４
齊藤　貴孝
電話番号　０３－３５８１－１１０１　内線　３５９９

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2013-065084 A（富士通株式会社） 2013.04.11，全文,全図（ファミリーなし）	1-10
A	WO 2010/082322 A1（株式会社日立製作所） 2010.07.22，全文,全図 & US 2011/0276828 A1	1-10
A	谷岡　日出男、外２名，金融業界におけるＡＩ技術，人工知能学会誌，2002.03.01，第１７巻，第２号，ｐ．２１４－２２１	1-10
A	藤野　秀則，因果関係の分析（１）－回帰分析・重回帰分析，ヒューマンインタフェース学会誌，2013.05.25，第１５巻，第２号，ｐ．１４１－１４９	1-10
A	上田　太一郎，正準相関分析と回帰分析変数選択基準を用いた判別分析変数選択の試み，計算機統計学，1996.05.31，第８巻，第２号，ｐ．１７１－１７５	1-10
A	藤巻　遼平、外１名，分析プロセス自動化・標準化への挑戦，情報処理学会　デジタルプラクティス　Ｖｏｌ．６　Ｎｏ．３， 2016.11.08（受入日），第６巻，第３号，ｐ．１９８－２０６	1-10