(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7559762号
(P7559762)

(45)発行日　令和6年10月2日(2024.10.2)

(24)登録日　令和6年9月24日(2024.9.24)

(51)国際特許分類
Ｇ０６Ｎ　20/00　(2019.01)

ＦＩ
Ｇ０６Ｎ　20/00

請求項の数　18　（全30頁）

(21)出願番号　特願2021-545233(P2021-545233)
(86)(22)出願日　令和2年9月1日(2020.9.1)
(86)国際出願番号　PCT/JP2020/032996
(87)国際公開番号　WO2021/049365
(87)国際公開日　令和3年3月18日(2021.3.18)
審査請求日　令和5年7月6日(2023.7.6)
(31)優先権主張番号　62/898,649
(32)優先日　令和1年9月11日(2019.9.11)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　000002185
ソニーグループ株式会社
東京都港区港南１丁目７番１号
(74)代理人　110003339
弁理士法人南青山国際特許事務所
(72)発明者　飯田 紘士
東京都港区港南１丁目７番１号 ソニー株式会社内
審査官　大倉 崚吾

最終頁に続く

(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

(57)【特許請求の範囲】

【請求項１】

予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得する取得部と、

前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定する推定処理部と

を具備し、

前記部分データセットの特徴量は、前記部分データセットの内容に応じた第１の特徴量を含み、

前記取得部は、前記部分データセットを解析することで前記第１の特徴量を算出する

情報処理装置。

【請求項２】

請求項１に記載の情報処理装置であって、

前記推定処理部は、前記精度情報として、前記部分データセットを用いて生成される前記予測モデルの予測精度に対する前記全データセットを用いて生成される前記予測モデルの予測精度の変化を推定する

情報処理装置。

【請求項３】

請求項２に記載の情報処理装置であって、

前記推定処理部は、前記予測精度の変化を推定する推定モデルを用いて構成される

情報処理装置。
【請求項４】
請求項３に記載の情報処理装置であって、
前記推定モデルは、所定のデータセットの一部のデータセットの特徴量と、所定の予測モデルを前記所定のデータセットの全部及び一部を用いて生成した場合に生じる予測精度の変化との関係を学習したモデルである
情報処理装置。
【請求項５】
請求項３に記載の情報処理装置であって、
前記推定モデルは、前記予測精度の変化量を複数のレベルに分類する分類モデルである
情報処理装置。
【請求項６】
請求項３に記載の情報処理装置であって、
前記推定モデルは、前記予測精度の変化量を複数のレベルに分類する分類モデルをルールベースで近似したモデルである
情報処理装置。
【請求項７】
請求項３に記載の情報処理装置であって、
前記推定モデルは、前記予測精度の変化量を推定する回帰モデルである
情報処理装置。
【請求項８】
請求項１に記載の情報処理装置であって、
前記第１の特徴量は、前記部分データセットに含まれるデータの数、前記データに含まれる特徴量の数、前記データの数と前記データに含まれる特徴量の数との比率の少なくとも１つを含む
情報処理装置。
【請求項９】
請求項１に記載の情報処理装置であって、
前記部分データセットの特徴量は、前記部分データセットを用いて生成される前記予測モデルに応じた第２の特徴量を含み、
前記取得部は、前記部分データセットを用いた前記予測モデルの生成処理を実行することで前記第２の特徴量を算出する
情報処理装置。
【請求項１０】
請求項９に記載の情報処理装置であって、
前記部分データセットは、互いに用途の異なる複数のデータグループを含み、
前記第２の特徴量は、前記複数のデータグループの各々に対する前記部分データセットを用いて生成される前記予測モデルの予測値を評価する評価値、又は前記評価値を比較した比較値の少なくとも一方を含む
情報処理装置。
【請求項１１】
請求項１０に記載の情報処理装置であって、
前記複数のデータグループは、学習データのグループと、検証データのグループと、テストデータのグループとを含む
情報処理装置。
【請求項１２】
請求項１０に記載の情報処理装置であって、
前記評価値は、前記部分データセットを用いて生成される前記予測モデルの予測値に関する誤差中央値、平均二乗誤差、及び誤差率中央値の少なくとも１つを含む
情報処理装置。

【請求項１３】
請求項１０に記載の情報処理装置であって、
前記比較値は、前記複数のデータグループのうち２つのデータグループについて算出された前記評価値の差分又は比率の少なくとも一方を含む
情報処理装置。
【請求項１４】
請求項１に記載の情報処理装置であって、さらに、
前記精度情報を提示する画面を生成する画面生成部を具備する
情報処理装置。
【請求項１５】
請求項１４に記載の情報処理装置であって、
前記推定処理部は、前記精度情報として、前記部分データセットを用いて生成される前記予測モデルの予測精度に対する前記全データセットを用いて生成される前記予測モデルの予測精度の変化を推定し、
前記画面生成部は、前記予測精度の変化量を複数のレベルにわけて提示する画面、または前記予測精度の変化量の値を提示する画面の少なくとも一方を生成する
情報処理装置。
【請求項１６】
請求項１４に記載の情報処理装置であって、
前記画面生成部は、前記部分データセットを用いた前記予測モデルの生成処理の実行を選択するための選択画面を生成し、
前記取得部は、前記生成処理の実行が選択された場合に、前記生成処理を実行して前記部分データセットの特徴量を算出し、
前記推定処理部は、前記部分データセットの特徴量に基づいて前記精度情報を推定する
情報処理装置。
【請求項１７】
予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得するステップと、
前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定するステップと
をコンピュータシステムが実行する情報処理方法であって、
前記部分データセットの特徴量は、前記部分データセットの内容に応じた第１の特徴量を含み、
前記部分データセットの特徴量を取得するステップは、前記部分データセットを解析することで前記第１の特徴量を算出することを含む
情報処理方法。
【請求項１８】
予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得するステップと、
前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定するステップと
をコンピュータシステムに実行させるプログラムであって、
前記部分データセットの特徴量は、前記部分データセットの内容に応じた第１の特徴量を含み、
前記部分データセットの特徴量を取得するステップは、前記部分データセットを解析することで前記第１の特徴量を算出することを含む
プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】

本技術は、機械学習を用いた予測モデルの学習処理に適用可能な情報処理装置、情報処理方法、及びプログラムに関する。

【背景技術】

【０００２】

従来、機械学習を用いて予測モデルを構築する技術が開発されている。予測モデルを適正に構築することで様々な予測処理を行うことが可能となる。予測モデルは、多数のデータを学習させることで構築されるが、その学習処理に時間がかかる場合がある。

【０００３】

例えば特許文献１には、深層学習の学習処理の最中にハードウェアリソースを追加することが可能なシステムについて記載されている。このシステムでは、学習処理の進捗状況とともに、ハードウェアリソースを追加するための追加ボタンがユーザに提示される。これにより、例えば学習処理の進捗状況が捗っていない場合には、ユーザはハードウェアリソースを追加して学習処理の速度を向上させることが可能となっている（特許文献１の明細書段落［００３０］［００３４］［００３５］、図４等）。

【先行技術文献】

【特許文献】

【０００４】

【文献】特開２０１７－１８２１１４号公報

【発明の概要】

【発明が解決しようとする課題】

【０００５】

予測モデルの学習処理では、上記したように演算リソースを増やすことで多数のデータを短時間で学習させることが可能となる。一方で、学習させるデータ数を増やしても予測精度の向上が見込めない場合等には、時間や費用が無駄になってしまう可能性がある。このため、予測モデルを効率的に学習させることが可能な技術が求められている。

【０００６】

以上のような事情に鑑み、本技術の目的は、予測モデルを効率的に学習させることが可能な情報処理装置、情報処理方法、及びプログラムを提供することにある。

【課題を解決するための手段】

【０００７】

上記目的を達成するため、本技術の一形態に係る情報処理装置は、取得部と、推定処理部とを具備する。

前記取得部は、予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得する。

前記推定処理部は、前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定する。

【０００８】

この情報処理装置では、全データセットのうち、部分データセットの特徴量が取得される。この特徴量に基づいて、全データセットを用いて予測モデルを生成した場合の予測精度を表す精度情報が推定される。これにより、例えば全データセットを用いるべきか否かを判断することが可能となり、予測モデルを効率的に生成することが可能となる。

【０００９】

前記推定処理部は、前記精度情報として、前記部分データセットを用いて生成される前記予測モデルの予測精度に対する前記全データセットを用いて生成される前記予測モデルの予測精度の変化を推定してもよい。

【００１０】

前記推定処理部は、前記予測精度の変化を推定する推定モデルを用いて構成されてもよい。

【００１１】

前記推定モデルは、所定のデータセットの一部のデータセットの特徴量と、所定の予測

モデルを前記所定のデータセットの全部及び一部を用いて生成した場合に生じる予測精度の変化との関係を学習したモデルであってもよい。

【００１２】

前記推定モデルは、前記予測精度の変化量を複数のレベルに分類する分類モデルであってもよい。

【００１３】

前記推定モデルは、前記予測精度の変化量を複数のレベルに分類する分類モデルをルールベースで近似したモデルであってもよい。

【００１４】

前記推定モデルは、前記予測精度の変化量を推定する回帰モデルであってもよい。

【００１５】

前記部分データセットの特徴量は、前記部分データセットの内容に応じた第１の特徴量を含んでもよい。この場合、前記取得部は、前記部分データセットを解析することで前記第１の特徴量を算出してもよい。

【００１６】

前記第１の特徴量は、前記部分データセットに含まれるデータの数、前記データに含まれる特徴量の数、前記データの数と前記データに含まれる特徴量の数との比率の少なくとも１つを含んでもよい。

【００１７】

前記部分データセットの特徴量は、前記部分データセットを用いて生成される前記予測モデルに応じた第２の特徴量を含んでもよい。この場合、前記取得部は、前記部分データセットを用いた前記予測モデルの生成処理を実行することで前記第２の特徴量を算出してもよい。

【００１８】

前記部分データセットは、互いに用途の異なる複数のデータグループを含んでもよい。この場合、前記第２の特徴量は、前記複数のデータグループの各々に対する前記部分データセットを用いて生成される前記予測モデルの予測値を評価する評価値、又は前記評価値を比較した比較値の少なくとも一方を含んでもよい。

【００１９】

前記複数のデータグループは、学習データのグループと、検証データのグループと、テストデータのグループとを含んでもよい。

【００２０】

前記評価値は、前記部分データセットを用いて生成される前記予測モデルの予測値に関する誤差中央値、平均二乗誤差、及び誤差率中央値の少なくとも１つを含んでもよい。

【００２１】

前記比較値は、前記複数のデータグループのうち２つのデータグループについて算出された前記評価値の差分又は比率の少なくとも一方を含んでもよい。

【００２２】

前記情報処理装置は、さらに、前記精度情報を提示する画面を生成する画面生成部を具備してもよい。

【００２３】

前記推定処理部は、前記精度情報として、前記部分データセットを用いて生成される前記予測モデルの予測精度に対する前記全データセットを用いて生成される前記予測モデルの予測精度の変化を推定してもよい。この場合、前記画面生成部は、前記予測精度の変化量を複数のレベルにわけて提示する画面、または前記予測精度の変化量の値を提示する画面の少なくとも一方を生成してもよい。

【００２４】

前記画面生成部は、前記部分データセットを用いた前記予測モデルの生成処理の実行を選択するための選択画面を生成してもよい。この場合、前記取得部は、前記生成処理の実行が選択された場合に、前記生成処理を実行して前記部分データセットの特徴量を算出し

てもよい。また、前記推定処理部は、前記部分データセットの特徴量に基づいて前記精度情報を推定してもよい。

【００２５】

本技術の一実施形態に係る情報処理方法は、コンピュータシステムにより実行される情報処理方法であって、予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得することを含む。

前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報が推定される。

【００２６】

本技術の一実施形態に係るプログラムは、コンピュータシステムに以下のステップを実行させる。

予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得するステップ。

前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定するステップ。

【図面の簡単な説明】

【００２７】

【図１】本技術の一実施形態に係るモデル生成システムの構成例を示すブロック図である。

【図２】図１に示す端末装置の構成例を示すブロック図である。

【図３】推定モデルの生成処理の概要を示す模式図である。

【図４】モデル生成システムの概要を説明するための模式図である。

【図５】メタ特徴量の具体例を示す表である。

【図６】モデル生成システムの基本的な動作例を示すフローチャートである。

【図７】設定画面の一例を示す模式図である。

【図８】選択エリアのインターフェースの一例を示す模式図である。

【図９】第１の予測モデルに関する評価画面の一例を示す模式図である。

【図１０】向上幅αの表示エリアのインターフェースの一例を示す模式図である。

【図１１】サーバ装置での演算を含む学習処理の一例を示すタイムチャートである。

【発明を実施するための形態】

【００２８】

以下、本技術に係る実施形態を、図面を参照しながら説明する。

【００２９】

［システムの構成］

図１は、本技術の一実施形態に係るモデル生成システムの構成例を示すブロック図である。モデル生成システム１００は、機械学習の手法を用いて予測処理を行う予測モデルを生成するシステムである。予測モデルにより予測対象についての予測分析が可能となる。

モデル生成システム１００では、予測モデルを生成するためのアプリケーション（以下、予測分析ツールと記載する）が動作する。ユーザは、予測分析ツールを用いることで、所望の予測処理を行う予測モデルを生成することが可能となる。

予測モデルの種類や予測対象等は限定されず、ユーザが任意に設定可能である。

【００３０】

モデル生成システム１００は、端末装置１０と、サーバ装置３０とを有する。端末装置１０及びサーバ装置３０は、通信ネットワーク３１を介して相互に通信可能に接続される。

端末装置１０は、ユーザが直接操作する情報処理装置であり、予測分析ツールの操作端末として機能する。端末装置１０としては、ＰＣ（Personal Computer）等が用いられる。あるいは、タブレット端末やスマートフォン等が端末装置１０として用いられてもよい。

サーバ装置３０は、端末装置１０にリモート接続する情報処理装置である。サーバ装置３０は、例えば端末装置１０で指定された所定の処理（例えば予測モデルの学習処理等）を実行し、その処理結果を端末装置１０に送信する。サーバ装置３０は、例えば所定のネ

ットワークで接続可能なネットワークサーバや、クラウド接続可能なクラウドサーバ等が用いられる。ここでは、従量課金制のサーバ装置３０が用いられる場合を想定する。
通信ネットワーク３１は、端末装置１０とサーバ装置３０とを通信可能に接続するネットワークであり、例えばインターネット回線等が用いられる。あるいは、専用のローカルネットワーク等が用いられてもよい。
【００３１】
図２は、図１に示す端末装置１０の構成例を示すブロック図である。端末装置１０は、表示部１１と、操作部１２と、通信部１３と、記憶部１４と、制御部１５とを有する。
【００３２】
表示部１１は、各情報を表示するディスプレイであり、例えば予測分析ツールのＵＩ（User Interface）画面等を表示する。表示部１１としては、例えば液晶ディスプレイ（ＬＣＤ：Liquid Cristal Display）や有機ＥＬ（Electro-Luminescence）ディスプレイ等が用いられる。表示部１１の具体的な構成は限定されず、例えば操作部１２として機能するタッチパネル等を搭載したディスプレイ等が用いられてもよい。また表示部１１としてＨＭＤ（Head Mounted Display）が用いられてもよい。
【００３３】
操作部１２は、ユーザが各種の情報を入力するための操作装置を含む。操作部１２としては、例えばマウスやキーボード等の情報入力が可能な装置が用いられる。この他、操作部１２の具体的な構成は限定されない。例えば操作部１２として、タッチパネル等が用いられてもよい。また操作部１２として、ユーザを撮影するカメラ等が用いられ、視線やジェスチャによる入力が可能であってもよい。
【００３４】
通信部１３は、端末装置１０と他の装置（例えばサーバ装置３０）との通信処理を行うモジュールである。通信部１３は、例えばＷｉ－Ｆｉ等の無線ＬＡＮ（Local Area Network）モジュールや、有線ＬＡＮモジュールにより構成される。この他、Ｂｌｕｅｔｏｏｔｈ（登録商標）等の近距離無線通信や、光通信等が可能な通信モジュールが用いられてよい。
【００３５】
記憶部１４は、不揮発性の記憶デバイスであり、例えばＨＤＤ（Hard Disk Drive）やＳＳＤ（Solid State Drive）等が用いられる。この他、記憶部１４して用いられる記録媒体の種類等は限定されず、例えば非一時的にデータを記録する任意の記録媒体が用いられてよい。
記憶部１４には、本実施形態に係る制御プログラム１６が記憶される。制御プログラム１６は、例えば端末装置１０全体の動作を制御するプログラムである。
【００３６】
また記憶部１４には、予測モデルの生成に用いる学習データセット１７が記憶される。学習データセット１７は、予測モデルの機械学習に用いられる複数のデータを含むデータセットである。学習データセット１７は、予測モデル５０の対象（予測項目）に合わせて適宜生成され、記憶部１４に格納される。予測モデルを構築する際には、学習データセット１７に含まれるデータが適宜読み込まれて用いられる。
【００３７】
学習データセット１７のデータは、例えば複数の属性値（特徴量）とそれらに対応する正解ラベルとが対応づけられたデータである。この場合、正解ラベルの項目を予測する予測モデルの学習が可能となる。
例えば顧客データを学習データセット１７として、顧客が好む商品を予測するモデルを生成するとする。この場合、例えば顧客データのうち、顧客が好む商品を表す項目（例えば顧客が購入した商品や閲覧した商品）が正解ラベルとなる。また他の属性（顧客の年齢、性別、商品の購入頻度等）についての項目は、予測モデルを学習させるための入力項目となる。
この他、学習データセット１７の種類等は限定されず、予測モデルに応じた任意のデー

タセットが用いられてよい。

【００３８】

端末装置１０では、後述するように、学習データセット１７の一部のデータセットを用いた処理が実行される。以下では、学習データセット１７の一部であるデータセットを部分データセット１８と記載する。

部分データセット１８は、例えば学習データセット１７からサンプリングされた複数のデータにより構成される。部分データセット１８となるデータは、例えば部分データセット１８が必要となるたびに適宜サンプリングされる。あるいは、部分データセット１８となるデータが予め設定されていてもよい。

本実施形態では、学習データセット１７は、予測モデルの生成に用いる全データセットに相当し、部分データセット１８は、全データセットの一部である部分データセットに相当する。

【００３９】

制御部１５は、端末装置１０が有する各ブロックの動作を制御する。制御部１５は、例えばＣＰＵやメモリ（ＲＡＭ、ＲＯＭ）等のコンピュータに必要なハードウェア構成を有する。ＣＰＵが記憶部１４に記憶されているプログラムをＲＡＭにロードして実行することにより、種々の処理が実行される。制御部１５としては、例えばＦＰＧＡ（Field Programmable Gate Array）等のＰＬＤ(Programmable Logic Device)、その他ＡＳＩＣ（Application Specific Integrated Circuit）等のデバイスが用いられてもよい。

【００４０】

本実施形態では、制御部１５のＣＰＵが本実施形態に係るプログラムを実行することで、機能ブロックとして、ＵＩ生成部２０と、予測モデル生成部２１と、メタ特徴量算出部２２と、精度推定部２３とが実現される。そしてこれらの機能ブロックにより、本実施形態に係る情報処理方法が実行される。なお、各機能ブロックを実現するために、ＩＣ（集積回路）等の専用のハードウェアが適宜用いられてもよい。

【００４１】

ＵＩ生成部２０は、ユーザと端末装置１０（あるいはサーバ装置３０）との情報のやり取りを行うためのＵＩを生成する。具体的には、ＵＩ生成部２０は、予測モデル５０を生成する際に表示部１１に表示されるＵＩ画面（図７及び図９参照）等を生成する。このＵＩ画面が、上記した予測分析ツールの画面となる。

ＵＩ画面には、例えばユーザに提示するための情報や、ユーザが情報を入力するための入力欄等が表示される。ユーザはＵＩ画面を見ながら、操作部（キーボード等）を操作して各種の設定や値等を指定することが可能である。ＵＩ生成部２０は、このようにＵＩ画面を介してユーザが指定した情報を受け付ける。

本実施形態では、ＵＩ生成部は、画面生成部に相当する。

【００４２】

予測モデル生成部２１は、予測モデルの生成処理を実行する。本実施形態では、予測モデル生成部２１により、部分データセット１８を用いた予測モデルの生成処理が実行される。この処理は、全ての学習データセット１７を用いた予測モデルの生成処理と比べ、短時間で実行可能な処理となる。なお、全ての学習データセット１７を用いた生成処理は、例えばサーバ装置３０により実行される。

以下では、部分データセット１８を用いて生成される予測モデルを第１の予測モデルと記載する。また全ての学習データセット１７を用いて生成される予測モデルを第２の予測モデルと記載する。

【００４３】

予測モデルの生成処理には、予測モデルを構築するために必要な一連の処理が含まれる。例えば予測分析ツールでは、予測モデルの生成処理として、予測モデルを学習させる学習処理（予測モデルのトレーニング）、予測モデルの状態（学習の傾向等）を検証する検証処理、予測モデルの予測精度等を確認するテスト処理等が適宜実行される。

従って、予測モデル生成部２１では、学習処理、検証処理、及びテスト処理等が部分デ

ータセット１８を用いてそれぞれ実行される。
予測モデルに用いられる機械学習のアルゴリズム等は限定されず、例えば予測モデルの処理内容に応じた任意のアルゴリズムが用いられてよい。アルゴリズムの種類等に係わらず本技術は適用可能である。
以下では、予測モデルの生成処理のことを指して、単に学習処理と記載する場合がある。
【００４４】
メタ特徴量算出部２２は、予測モデルの生成に用いる学習データセット１７の一部である部分データセット１８の特徴量を取得する。
ここで、部分データセット１８の特徴量とは、部分データセット１８自身の性質等を表す特徴量である。以下では、このようなデータセット自身の特徴量をメタ特徴量と記載する。すなわち、メタ特徴量算出部２２では、部分データセット１８のメタ特徴量が取得される。
なお、メタ特徴量は、部分データセット１８を構成するデータに記録された属性値（以下、データ特徴量と記載する）とは異なる。例えば、データセットに含まれるデータの数やデータ特徴量の数といったデータセットそのものが持つ特徴量は、メタ特徴量となる。
【００４５】
メタ特徴量には、部分データセット１８を解析して得られる特徴量（第１の特徴量）が含まれる。この特徴量は、部分データセット１８を解析することで算出される。
またメタ特徴量には、実際に部分データセット１８を使用することで得られる特徴量（第２の特徴量）が含まれる。この特徴量は、上記した予測モデル生成部２１により生成される予測モデルを用いて算出される。
メタ特徴量については、図５等を参照して後に詳しく説明する。
本実施形態では、予測モデル生成部２１とメタ特徴量算出部２２とが共動することで、取得部が実現される。
【００４６】
精度推定部２３は、部分データセット１８の特徴量（メタ特徴量）に基づいて、学習データセット１７を用いて生成される予測モデル（第２の予測モデル）の予測精度を表す精度情報を推定する。
精度情報は、第２の予測モデルの予測精度を表すことが可能な情報である。この精度情報を参照することで、全ての学習データセット１７を使って第２の予測モデルを構築した場合にどの程度の予測精度が実現できるかといったことを判断することが可能となる。
【００４７】
本実施形態では、精度推定部２３により、精度情報として、部分データセット１８を用いて生成される予測モデル（第１の予測モデル）の予測精度に対する学習データセット１７を用いて生成される予測モデル（第２の予測モデル）の予測精度の変化が推定される。
機械学習では、学習させるデータの数が多いほど予測精度が向上すると考えられる。しかしながら、データの数を増やしたからといって予測精度が十分に向上するとは限らない、
精度推定部２３は、部分データセット１８で学習した第１の予測モデルを基準として、全ての学習データセット１７を用いて第２の予測モデルを生成した場合に予想される予測精度の向上幅を推定する。この予測精度の向上幅が、上記した予測精度の変化に対応する。
【００４８】
精度推定部２３は、予測精度の変化を推定する推定モデルを用いて構成される。推定モデルは、部分データセット１８のメタ特徴量を入力として、第１の予測モデルに対する第２の予測モデルの予測精度の変化を出力するように学習した学習モデルである。このように、精度推定部２３は、推定モデルを実装したモジュール（推定モジュール）であるともいえる。
推定モデルは、例えば、ウェブ等から入手できる多数のデータセットのメタ特徴量から学習を行うことで構成される。推定モデルを生成する方法については、後に詳しく説明する。
【００４９】

予測分析ツールでは、まず、予測精度の向上幅を推定する推定モデル（推定モジュール）が構成される。推定モデルは、例えば予測モデルの種類に合わせて生成される。あるいは種類の異なる予測モデルに対応可能な汎用性のある推定モデルが生成されてもよい。推定モデルのデータは、例えば予め記憶部１４に格納され、精度推定部２３を動作させるたびに適宜読み込まれて使用される。

精度推定部２３では、このように構成された推定モデルを用いて、実際に使用する学習データセット１７について、全てのデータセットを学習に用いた際の推定精度の向上幅（推定精度の変化）が推定される。

【００５０】

［推定モデルの生成処理］

図３は、推定モデルの生成処理の概要を示す模式図である。以下では、図３を参照して、予測精度の向上幅を推定する推定モデル４０を生成する方法について説明する。

近年、機械学習に用いることが可能な多数のデータセットがウェブ等から入手できるようになっている。これらのデータセットのメタ特徴量から学習した情報を用いることで、新規のデータセットの性質を予測するといったことが可能である。

本実施形態では、推定モデル４０を構築するために、このように既に存在する多数のデータセットが用いられる。以下では、推定モデル４０を構築するために用いられるデータセットを、推定用データセットと記載する。

【００５１】

例えば、推定用データセットを用いてある予測モデル（以下、推定用予測モデルと記載する）を生成するとする。この場合、推定用データセットの一部を用いて学習したモデルと、全ての推定用データセットを用いて学習したモデルとでは、予測精度が異なる。このような予測精度の違いを、複数の推定用データセットごとに学習させることで、予測精度の向上幅を推定する推定モデル４０が構築される。

なお推定用予測モデルは、例えば推定用データセットに合わせて任意に設定されてよい。

本実施形態では、推定用データセットは、所定のデータセットに相当する。また推定用予測モデルは、所定の予測モデルに相当する。

【００５２】

図３に示すように、推定モデルの生成処理では、入力データ２５と、入力データ２５に対応する解答データ２６とのセットが用いられる。入力データ２５及び解答データ２６のセットは、推定用データセットごとにそれぞれ生成される。推定モデルの生成処理に用いられる推定用データセットの数（入力データ２５及び解答データ２６のセットの数）は、例えば数百セット程度である。

【００５３】

入力データ２５は、推定用データセットのメタ特徴量である。具体的には、対象となる推定用データセットに含まれる一部のデータセット(例えば１０％等)のメタ特徴量が、入力データ２５として用いられる。

入力データ２５に含まれるメタ特徴量としては、例えばデータの個数、データ特徴量の個数、あるいは後述する学習データ（train）/検証データ（validation）/テストデータ（test）に対する予測評価値等が挙げられる。これらのメタ特徴量の数や種類は、例えば推定モデル４０が向上幅を推定する際に実際に参照するメタ特徴量と同様に設定される（図５参照）。

また入力データ２５を生成する際には、メタ特徴量算出部２２が部分データセット１８のメタ特徴量を算出する方法と同様の方法が用いられる。

【００５４】

解答データ２６は、推定モデル４０が学習するべき項目（予測精度の向上幅）の正解ラベルである。具体的には、推定用データセットの一部(例えば１０％等)を使用して学習を行った際の推定用予測モデルの予測精度と、推定用データセットの全部を使用して学習を行った際の推定用予測モデルの予測精度との差分（向上幅）が正解ラベルとして用いられる。

従って解答データ２６は、推定用データセットの一部又は全部を用いて実際に推定用予測モデルを学習させることで算出される。なお、この時生成された推定用予測モデルから、上記した入力データ２５の一部が算出される。

【００５５】

推定モデル４０の生成処理では、上記した入力データ２５及び解答データ２６が生成される。すなわち、複数の推定用データセットに対して、それらのメタ特徴量（入力データ２５）と実際に全データで学習した場合の予測精度の向上幅（解答データ２６）とがそれぞれ算出される。

このように算出された入力データ２５及び解答データ２６に基づいて機械学習が実行される。具体的には、予測精度の向上幅を正解ラベルとし、メタ特徴量を特徴量として学習処理等が実行される。

この処理は、例えば全てのデータを使用した際に予測精度の向上幅が大きくなるようなデータセットの特徴を学習させる処理であるともいえる。すなわち、推定モデル４０は、データ数を増やしたときに予測精度が向上するデータセットの特徴（メタ特徴量）を学習することになる。これにより、データセットのメタ特徴量から、予測精度の向上幅を推定する学習済みの推定モデル４０が構築される。

【００５６】

このように、推定モデル４０は、推定用データセットの一部のデータセットの特徴量と、推定用予測モデルを推定用データセットの全部及び一部を用いて学習させた場合に生じる予測精度の変化との関係を学習したモデルである。

推定モデル４０を用いることで、未知の学習データセット１７が用いられる場合であっても、予測モデルの予測精度の向上幅を精度よくかつ容易に推定することが可能となる。

なお、推定モデル４０は、メタ特徴量と正解ラベルから学習して得られた学習モデルでもよいし、その学習モデルを近似したモデルであってもよい。以下、推定モデル４０の種類について説明する。

【００５７】

例えば、推定モデル４０は、予測精度の変化量を複数のレベルに分類する分類モデルである。この場合、例えば正解ラベル（解答データ２６）は、予測精度の変化量を表す各レベルを２値分類したものとなる。

変化量を表すレベルとしては、例えば全データセットで学習した場合に、一部のデータセットで学習した時よりも予測精度が変化する度合いを表すレベルが設定される。例えば、予測精度が「大幅に向上する(５％以上)」場合、「ある程度向上する(２－５％)」場合、あるいは「殆ど向上しない(２％未満)」場合といった３段階のレベルにわけて正解ラベルが設定される。これにより、予測精度の向上幅を複数の段階に分けて推定することが可能となる。

【００５８】

また例えば、推定モデル４０は、予測精度の変化量を複数のレベルに分類する分類モデルをルールベースで近似したモデルであってもよい。この場合、推定モデル４０は、分類モデルを簡易化したルールベースの分類器となる。

例えば上記した分類モデルを所定のアルゴリズムで近似することで、最終的な推定モデル４０が算出される。分類モデルを近似するアルゴリズムとしては、決定木のアルゴリズムや、決定木をランダムに組み合わせたランダムフォレスト、あるいは分類モデルによる処理をルールの集合に置き換えるルールフィット等が用いられる。

ルールベースのモデルを用いることで、向上幅の推定に要する演算量や演算時間を抑制することが可能である。また推定処理の内容等を、ユーザにも理解できるように説明するといったことも可能となる。

【００５９】

また例えば、推定モデル４０は、予測精度の変化量を推定する回帰モデルであってもよい。この場合、例えば正解ラベル（解答データ２６）は、予測精度の変化量の値（例えば向上幅Ｘ％等）に設定される。

このように、予測精度の変化量（向上幅）を具体的な数値として直接回帰するような推定モデル４０が構築されてもよい。これにより、ユーザに対して、予測精度の向上幅の具体的な推定値を提示することが可能となる。

この他、推定モデル４０の具体的な構成は限定されない。

【００６０】

［モデル生成システムの概要］

図４は、モデル生成システム１００の概要を説明するための模式図である。ここでは、上記した推定モデル４０を用いて予測精度の向上幅を推定し、その推定結果を提示するまでの処理の流れが模式的に図示されている。

図４には、予測モデル５０（第１の予測モデル５１）の生成処理（ステップ１）、メタ特徴量の算出処理（ステップ２）、向上幅の推定処理（ステップ３）、及びＵＩの提示処理（ステップ４）が含まれる。以下順番に説明する。

【００６１】

［予測モデルの生成処理］

予測精度の向上幅を推定する場合、部分データセット１８を用いて第１の予測モデル５１を生成する処理が実行される。この処理は、学習データセット１７全体での学習（第２の予測モデル５２の生成処理）を行う前に実行される予備的な生成処理である。

具体的には、予測モデル生成部２１により、学習データセット１７に含まれる一部のデータセット（部分データセット１８）がサンプリングされる。そして、この部分データセット１８を用いた機械学習が実行される。

【００６２】

この生成処理では、部分データセット１８は、互いに用途の異なる複数のデータグループに分けて用いられる。すなわち、部分データセット１８には、互いに用途の異なる複数のデータグループが含まれるとも言える。

一つのデータグループには、少なくとも１つのデータが含まれ、各グループは、それぞれ別の目的で使用される。なおデータグループを設定する方法は限定されない。

【００６３】

本実施形態では、複数のデータグループは、学習データのグループと、検証データのグループと、テストデータのグループである。

学習データ（training data）は、予測モデル５０の学習処理を行う際に用いられるデータであり、予測モデル５０が実際に学習（トレーニング）するデータである。この学習データが多いほど、予測モデル５０の精度が向上する傾向がある。

検証データ（validation data）は、予測モデル５０の学習の状態（学習の傾向等）を検証する検証処理を行う際に用いられるデータである。従って検証データは、予測モデル５０の学習をチェックするためのデータであると言える。

テストデータ（test data）は、学習データで学習した予測モデル５０の最終的な予測精度等を確認するテスト処理を行う際に用いるデータである。従ってテストデータは、予測モデル５０を評価するためのデータであると言える。

なお、学習の種類や設定によっては、これらのデータのうち検証データが不要な場合もある。この場合、検証データのグループはなくてもよい。

【００６４】

予測モデル生成部２１では、これらのデータグループを使って、上記した学習処理、検証処理、テスト処理等が適宜実行される。これにより、部分データセット１８から学習した学習済みの予測モデル５０（第１の予測モデル５１）が生成される。

各データグループの情報や、第１の予測モデル５１のデータは、メタ特徴量算出部２２に出力される。

【００６５】

［メタ特徴量の算出処理］

メタ特徴量算出部２２により、第１の予測モデル５１の生成に用いた部分データセット１８のメタ特徴量Ｆが算出される。

まず、メタ特徴量を算出するために必要なデータが適宜読み込まれる。具体的には、部分データセット１８に含まれる各データグループと、部分データセット１８を用いて生成された第１の予測モデル５１とが読み込まれる。図４には、部分データセット１８の学習データ及びテストデータのグループと、第１の予測モデル５１とが模式的に図示されている。また図示を省略したが、検証データのグループも適宜読み込まれる。

このように、予測精度の向上幅を知りたいデータセット（学習データセット１７）について、そこからサンプリングした部分データセット１８（学習データ、検証データ、テストデータ）と、部分データセット１８で学習済みの第１の予測モデルとが用意される。

以下では、これらのデータをもとに算出されるメタ特徴量Ｆについて具体的に説明する。

【００６６】

図５は、メタ特徴量の具体例を示す表である。図５には、複数のメタ特徴量について、各メタ特徴量の項目とその具体的な内容とが示されている。これらのメタ特徴量は、例えば予測モデルとして回帰モデルを用いる場合に使用される。

ここでは、各メタ特徴量に番号（Ｆ１～Ｆ１６）を付けて説明する。なお、図５に示す表は一例であって、メタ特徴量の数や種類等は限定されない。

【００６７】

部分データセット１８のメタ特徴量には、部分データセット１８の内容に応じた第１の特徴量が含まれる。第１の特徴量とは、部分データセット１８そのものが持つ特徴量である。

本実施形態では、メタ特徴量算出部２２により、部分データセット１８を解析することで第１の特徴量が算出される。図５に示す表では、メタ特徴量Ｆ１～Ｆ４及びＦ９が、第１の特徴量に相当する。

【００６８】

メタ特徴量Ｆ１（データ数）は、部分データセット１８に含まれるデータの数である。例えば、部分データセット１８に含まれるデータの総数がメタ特徴量として算出される。あるいは部分データセット１８に含まれる学習データの総数が用いられてもよい。

メタ特徴量Ｆ２（特徴量数）は、部分データセット１８のデータに含まれるの特徴量（データ特徴量）の数である。例えば、各データに設定されたデータ特徴量の総数がメタ特徴量として算出される。またデータごとに特徴量の数（種類）が異なる場合には、延べ総数等が算出されてもよい。

メタ特徴量Ｆ３（特徴量数／データ数）は、部分データセット１８に含まれるデータの数とデータに含まれるデータ特徴量の数との比率である。例えば、上記したメタ特徴量Ｆ２をメタ特徴量Ｆ１で除算した値が新たなメタ特徴量として算出される。

メタ特徴量Ｆ４（展開後の特徴量数）は、所定の前処理を済ませた後の学習データに使用するデータ特徴量の数である。例えばＯｎｅＨｏｔエンコーディング等の前処理を行う場合、ダミー変数が用いられることでデータ特徴量の数が処理の前後で変化する。この処理後のデータ特徴量の総数がメタ特徴量として算出される。

メタ特徴量Ｆ９（正解値の分散）は、予測対象ラベル（正解ラベル）の分散である。例えば、回帰モデルの予測対象となる予測対象ラベルの値についての分散値（例えば標準偏差等）がメタ特徴量として算出される。

【００６９】

また、部分データセット１８のメタ特徴量には、部分データセット１８を用いて生成される予測モデル５０（第１の予測モデル５１）に応じた第２の特徴量が含まれる。すなわち第２の特徴量は、部分データセット１８を実際に使用することで得られる特徴量であるといえる。

本実施形態では、予測モデル生成部２１により、部分データセット１８を用いた第１の予測モデル５１の生成処理が実行されることで第２の特徴量が算出される。図５に示す表では、メタ特徴量Ｆ５～Ｆ８及びＦ１０～Ｆ１６が、第２の特徴量に相当する。

【００７０】

ここでは、第２の特徴量として、複数のデータグループの各々に対する部分データセッ

ト１８を用いて生成される第１の予測モデル５１の予測値を評価する評価値が用いられる。ここで評価値とは、あるデータグループ（学習データ、検証データ、及びテストデータのグループ）を入力とした場合に、第１の予測モデル５１から出力される予測値を評価することが可能なパラメータである。

予測値を評価するパラメータとしては、例えば予測値に関する誤差中央値（ＭＡＥ：Mean Absolute Error）、平均二乗誤差（ＲＭＳＥ：Root Mean Squared Error）、誤差率中央値（ＭＡＰＥ：Mean Absolute Percentage Error）等が用いられる。あるいは予測値の分散等が評価値として用いられてもよい。評価値として用いるパラメータは限定されず、他の指標が用いられてもよい。

【００７１】

メタ特徴量Ｆ５（Ｉｔｅｒａｔｉｏｎ数に応じたテストデータの誤差中央値の変化）は、テストデータに対するＩｔｅｒａｔｉｏｎ処理における誤差中央値の変化量である。Ｉｔｅｒａｔｉｏｎ処理は、例えばテストデータの選び方を変えてモデルの予測精度を複数回にわたって検証する処理（交差検証）であり、テストデータの選び方による評価の偏りを低減する効果がある。具体的には、Ｉｔｅｒａｔｉｏｎが収束した時の回数の半分の回数における誤差中央値と、最終的に収束した誤差中央値との差がメタ特徴量として算出される。メタ特徴量Ｆ５は、テストデータに対する評価値の一例である。

【００７２】

メタ特徴量Ｆ６（学習／検証／テストデータの誤差中央値）は、学習済みの第１の予測モデル５１で予測した際の、学習データ、検証データ、テストデータの各グループに対する誤差中央値（ＭＡＥ）の値である。

メタ特徴量Ｆ７（学習／検証／テストデータの平均二乗誤差）は、学習済みの第１の予測モデル５１で予測した際の、学習データ、検証データ、テストデータの各グループに対する平均二乗誤差（ＲＭＳＥ）の値である。

メタ特徴量Ｆ８（学習／検証／テストデータの誤差率中央値）は、学習済みの第１の予測モデル５１で予測した際の、学習データ、検証データ、テストデータの各グループに対する誤差率中央値（ＭＡＰＥ）の値である。

メタ特徴量Ｆ１０（予測値の分散）は、学習済みの第１の予測モデル５１で予測した予測値の分散（標準偏差等）である。

これらの評価値の全部、又は一部が用いられてよい。

【００７３】

また、第２の特徴量として、上記した評価値を比較した比較値が用いられる。ここで比較値とは、各データグループ（学習データ、検証データ、及びテストデータのグループ）について算出された評価値をグループ間で比較した値である。

具体的には、複数のデータグループのうち２つのデータグループについて算出された評価値の差分又は比率の少なくとも一方が比較値として用いられる。

【００７４】

メタ特徴量Ｆ１１（学習データとテストデータとの誤差中央値の差）は、学習データに対する誤差中央値と、テストデータに対する誤差中央値との差である。

メタ特徴量Ｆ１２（学習データとテストデータとの誤差中央値の比率）は、学習データに対する誤差中央値と、テストデータに対する誤差中央値との比率である。

メタ特徴量Ｆ１３（検証データとテストデータとの誤差中央値の差）は、検証データに対する誤差中央値と、テストデータに対する誤差中央値との差である。

メタ特徴量Ｆ１４（検証データとテストデータとの誤差中央値の比率）は、検証データに対する誤差中央値と、テストデータに対する誤差中央値との比率である。

メタ特徴量Ｆ１５（学習データと検証データとの誤差中央値の差）は、学習データに対する誤差中央値と、検証データに対する誤差中央値との差である。

メタ特徴量Ｆ１６（学習データと検証データとの誤差中央値の比率）は、学習データに対する誤差中央値と、検証データに対する誤差中央値との比率である。

【００７５】

これらのメタ特徴量Ｆ１１～Ｆ１６は、例えば上記したメタ特徴量Ｆ６の結果をもとに算出される。

なお、差分及び比率を算出する際の基準は任意に設定されてよい。例えばメタ特徴量Ｆ１１において、学習データに対する誤差中央値からテストデータに対する誤差中央値を引いてもよいしその逆でもよい。あるいは差分の絶対値が用いられてもよい。また例えばメタ特徴量Ｆ１２において、学習データに対する誤差中央値をテストデータに対する誤差中央値で割って比率を算出してもよいし、その逆でもよい。

また、誤差中央値に代えて、平均二乗誤差や誤差率中央値を比較した比較値等がメタ情報として用いられてもよい。

【００７６】

このように、誤差中央値等は学習済みの第１の予測モデルと、その学習に用いた部分データセット１８が与えられれば計算可能である。他の特徴量についてもサンプリングした学習データ・検証データ・テストデータ・第１の予測モデル５１を使えば全て計算可能である。

なお、これらの値は殆どが第１の予測モデル５１を作成する過程で計算している値であり、追加の計算は必要としない。

【００７７】

［向上幅の推定処理］

図４に戻り、精度推定部２３により、上記のように算出されたメタ特徴量Ｆに基づいて、予測モデル５０における予測精度の向上幅αが算出される。

予測精度の向上幅αの推定には、メタ特徴量から学習する事で構築した、予測精度の向上幅を推定する推定モデル４０が用いられる（図３参照）。具体的には、部分データセット１８のメタ特徴量Ｆが推定モデル４０に入力データとして入力される。そして推定モデル４０を用いた演算が実行され、向上幅αの分類値や値が出力される。

【００７８】

例えば、推定モデル４０が分類モデルや、分類モデルを近似したルールベースのモデルである場合、向上幅αを複数のレベルに分類した分類結果が出力される。この場合、出力値は「大幅に向上する(５％以上)」、「ある程度向上する(２－５％)」、「殆ど向上しない(２％未満)」といった各レベルについての予測確立となる。すなわち向上幅が５％以上となる確率等が算出される。

また例えば、推定モデル４０が回帰モデルである場合、予測精度の向上幅αの値が回帰問題を解くことで直接推定される。この場合、出力値は向上幅αを具体的に表す値（例えばα＝４％等）となる。

推定モデル４０の出力は、ＵＩ生成部２０に出力される。

【００７９】

［ＵＩの提示処理］

ＵＩ生成部２０により、推定した予測精度の向上幅αが表示される。具体的には、ＵＩ生成部２０により、予測精度の向上幅α（予測精度の変化）の推定結果を提示する画面が生成される。そして生成された画面が、表示部１１に表示される。

これにより、全ての学習データセット１７を使って予測モデル５０（第２の予測モデル５２）を生成した場合に想定される予測精度の向上幅αがユーザに提示され、第２の予測モデル５２の生成を行うか否かの判断を支援するといったことが可能となる。

【００８０】

このように、モデル生成システム１００（予測分析ツール）では、学習データセット１７の一部を用いて短時間で学習を行い、その時の情報から全データセットを用いて学習した際にどれくらい予測精度が向上するかを推定することが可能である。すなわち、一部のデータセットから１回だけ学習することで、全データを学習に使用した際の予測精度の向上幅αが推定可能である。

【００８１】

本発明者は、予測精度の向上幅αを推定する推定モデル４０を実際に構築し、その精度

を検証した。その結果、向上幅αを分類する推定モデル４０のＡＵＣ(分類問題に対する評価指標)は０．７５となり、向上幅αを高い精度で予測できていることが分かった。これは、メタ特徴量から、予測精度が向上するデータセットを適正に予測できることを意味する。

【００８２】

また本発明者は、実際の実験結果から、データ数を増やした際に精度が向上するデータセットの傾向についての知見を得た。具体的には、学習データとテストデータに対する予測値の評価指標（例えば上記した評価値）に大きな差があるデータセットほど、全データで学習した時に精度の向上幅が大きい傾向にあることを見出した。例えば学習データとテストデータの評価指標の差は、予測モデル５０がどの程度学習データに過学習しているかを表す指標となっており、これらの差が大きい場合にはデータ数の増加と共に精度の向上を見込めることが多い。

【００８３】

従ってメタ特徴量の中でも、学習データとテストデータに対する予測値の評価指標（評価値）や、評価指標を比較した値（比較値）は、特に重要な特徴量となる。このようなメタ特徴量を入力とする推定モデル４０を用いることで、予測精度の向上幅αを精度よく推定することが可能となる。

【００８４】

［モデル生成システムの基本動作］

図６は、モデル生成システムの基本的な動作例を示すフローチャートである。図６に示す処理は、例えば端末装置１０を使用するユーザが予測分析ツールで予測モデル５０を生成する際に実行される処理である。

まず、予測モデル５０の各設定値が読み込まれる（ステップ１０１）。具体的には、ＵＩ生成部２０により、予測モデル５０に関する設定画面が生成され、表示部１１に出力される。そしてユーザが設定画面を介して入力した内容（設定値）が読み込まれる。

【００８５】

図７は、設定画面の一例を示す模式図である。図７に示すように、設定画面３５には、複数の設定欄が設けられる。ここでは、商品の購入記録を含む顧客データを学習データセット１７として、商品の購入の有無を予測する予測モデル５０を生成する場合について説明する。

【００８６】

「入力項目」の設定欄（画面右側）では、学習データセット１７に含まれる項目のうち、予測モデル５０の学習に用いる項目（データ特徴量）を指定可能である。ここでは、顧客に関する"年齢"、"性別"、"顧客ランク"、"過去購入額"、"クーポン利用回数"、"メールアドレス登録"、"オプション購入"等の項目が選択可能に提示される。また、各項目についてのデータタイプやユニーク数等が合わせて表示される。

【００８７】

「予測タイプ」の設定欄では、予測モデル５０のタイプを指定可能である。ここでは、「二値分類」、「多値分類」、「数値予測」（回帰予測）の項目が選択可能に表示される。ここでは、二値分類が予測モデル５０のタイプとして選択される。

「予測値」の設定欄では、予測モデル５０の予測対象（対象項目）を指定することが可能である。ここでは、"購入あり"及び"購入なし"の項目のうち"購入あり"が予測対象として選択される。なお項目名（"購入あり"及び"購入なし"）の隣には、学習データセット１７における各項目の割合が表示される。

【００８８】

設定画面３５において点線で示したエリアは、部分データセット１８を用いた学習を選択するための選択エリア３６である。図７に示す例では、選択エリア３６には、「使用するデータの割合」の設定欄が設けられる。この設定欄では、部分データセット１８として用いられるデータの割合をいくつかの候補から選択して指定することが可能である。

例えば学習データセット１７に対する部分データセット１８の割合が０％～１００％の範囲で選択可能に提示される（ここでは１０％が選択される）。このＵＩでは、部分デー

タセット１８の割合が０％より大きい有限値である場合、部分データセット１８を用いた学習が選択されることになる。なお、部分データセット１８の割合が０％である場合には、部分データセット１８を用いた学習は選択されない。

【００８９】

各設定欄に必要な情報を入力した後で、"学習及び評価を実行"と書かれた実行ボタンを押すと、予測モデル５０についての学習処理等が開始される。また"キャンセル"と書かれたキャンセルボタンを押すと、設定画面３５での各設定値の入力がキャンセルされ、ひとつ前の画面が表示される。

【００９０】

図８は、選択エリア３６のインターフェースの一例を示す模式図である。

図８Ａに示す選択エリア３６には、「使用するデータの割合」の設定欄が設けられる。この設定欄では、部分データセット１８として用いられるデータの割合を０％～１００％の範囲で自由に入力して指定することが可能である。この場合、入力値が０よりも大きい場合に、部分データセット１８を用いた学習が選択される。

【００９１】

図８Ｂに示す選択エリア３６には、「学習モード」の設定欄が設けられる。この設定欄では、"クイックモード"という項目と"全データで学習"という項目とがそれぞれ選択可能に提示される。

ここで、クイックモードとは、部分データセット１８を用いた学習を行い、本番の学習の前に予測精度の向上幅を短時間で算出するモードである。クイックモードでは、例えば予め設定されたデフォルトの割合で部分データセット１８が選択されて用いられる。なお部分データセット１８の割合が選択可能であってもよい。このように、学習モードを選択させることで、部分データセット１８での学習の有無が設定されてもよい。

【００９２】

図８Ｃに示す選択エリア３６には、「学習を行う端末」の設定欄が設けられる。この設定欄では、"この端末で学習"という項目と"クラウド上で学習"という項目とがそれぞれ選択可能に提示される。

"この端末で学習"という項目は、部分データセット１８（一部データ）を用いた学習を端末装置１０で実行する場合に選択される。また"クラウド上で学習"という項目は、全ての学習データセット１７を用いた学習をサーバ装置３０で実行する場合に選択される。このように、学習処理を行う装置を選択させることで、部分データセット１８での学習の有無が設定されてもよい。

【００９３】

このように、ＵＩ生成部２０は、部分データセット１８を用いた第１の予測モデル５１の生成処理の実行を選択するための設定画面３５を生成する。本実施形態では、設定画面３５は、選択画面に相当する。

これにより、ユーザは部分データセット１８での学習を行うか否か、すなわち向上幅αを推定するか否かを適宜選択することが可能となる。

【００９４】

図６に戻り、設定画面３５から入力された設定値が読み込まれると、部分データセット１８での学習処理を開始するか否かが判定される（ステップ１０２）。

例えば選択エリア３６に表示されたＵＩにおいて、部分データセット１８での学習を行う旨が選択されたとする。この状態で、図７に示す実行ボタンが押された場合、部分データセット１８での学習を行うと判定され（ステップ１０２のＹｅｓ）、部分データセット１８での学習及びその学習結果を用いた向上幅αの推定処理が開始される。

また例えば、部分データセット１８での学習を行う旨が選択されていない状態で実行ボタンが押された場合、部分データセット１８での学習は行わないと判定され（ステップ１０２のＮｏ）、後述するステップ１０７が実行される。

【００９５】

部分データセット１８での学習を行うと判定された場合、予測モデル生成部２１により

、部分データセット１８を用いた第１の予測モデル５１の生成処理が実行される（ステップ１０３）。この処理は、図４を参照して説明したステップ１の予測モデルの生成処理に相当する。

例えば、設定画面３５の設定値で選択された入力項目から予測値を出力するようなモデルが構成され、部分データセット１８を用いた学習処理・検証処理・テスト処理等が実行され、学習済みの第１の予測モデル５１が構築される。

【００９６】

第１の予測モデル５１が構築されると、メタ特徴量算出部２２により、部分データセット１８のメタ特徴量Ｆが算出される（ステップ１０４）。この処理は、図４を参照して説明したステップ２のメタ特徴量の算出処理に相当する。

例えば、第１の予測モデル５１のデータと、その学習に用いられた部分データセット１８とが読み込まれ、既に用意されている推定モデル４０の入力となるメタ特徴量Ｆがそれぞれ算出される。

【００９７】

第１の予測モデル５１が構築されると、メタ特徴量算出部２２により、部分データセット１８のメタ特徴量Ｆが算出される（ステップ１０４）。この処理は、図４を参照して説明したステップ２のメタ特徴量の算出処理に相当する。

例えば、第１の予測モデル５１のデータと、その学習に用いられた部分データセット１８とが読み込まれ、既に用意されている推定モデル４０の入力となるメタ特徴量Ｆがそれぞれ算出される。

このように、本実施形態では、第１の予測モデル５１の生成処理の実行が選択された場合に、その生成処理を実行して部分データセット１８のメタ特徴量Ｆが算出される。

【００９８】

メタ特徴量Ｆが算出されると、精度推定部２３により、部分データセット１８のメタ特徴量Ｆに基づいて予測精度の向上幅α（精度情報）が推定される（ステップ１０５）。この処理は、図４を参照して説明したステップ３の向上幅の推定処理に相当する。

例えば、推定モデル４０に対して前のステップで算出された各メタ特徴量Ｆが入力され、予測精度の向上幅αの分類レベルや値が算出される。

【００９９】

向上幅αが算出されると、ＵＩ生成部２０により向上幅αを提示する画面が生成される（ステップ１０６）。この処理は、図４を参照して説明したステップ４のＵＩの提示処理に相当する。

本実施形態では、第１の予測モデル５１の評価結果とともに、向上幅αを提示する評価画面が生成され、表示部１１に表示される。

【０１００】

図９は、第１の予測モデル５１に関する評価画面の一例を示す模式図である。

図９に示す評価画面３７の左側には、モデルの選択エリア３６が設けられる。選択エリアには、既に評価を行った第１の予測モデル５１が、その評価値、生成日時、使用したデータ名とともに、選択可能に提示される。

また評価画面３７の右側には、選択エリア３６で選択された第１の予測モデル５１の予測精度のレベルを示す「予測精度レベル」の表示欄と、「項目の寄与度」の表示欄とが設けられる。また評価画面３７の右側には、向上幅αの推定結果を提示する表示エリア３８が設けられる。

【０１０１】

「予測精度レベル」の表示欄には、第１の予測モデル５１の性能を示す評価指標として、例えばＲＯＣ（Receiver Operating Characteristic）曲線のＡＵＣ（Area Under the Curve）が表示される。ＡＵＣは、分類モデルの分類精度を示す指標である。この他、評価指標に関連する説明項目（モデルの精度についてのコメント等）が表示される。

また「項目の寄与度」の表示欄には、分類に影響した項目ごとの寄与度を示す棒グラフが表示される。これにより、例えば"購入あり"という分類に影響した項目や、"購入なし"

という分類に影響した項目を、項目間で比較することが可能となる。

【０１０２】

図９に示す表示エリア３８には、向上幅αについて説明するテキストが提示される。ここでは、第１の予測モデル５１の学習に用いられたデータ（部分データセット１８）の割合とともに、全データ（全ての学習データセット１７）を用いることで期待される予測精度の向上幅αを提示する説明分が用いられる。ここでは、向上幅αを具体値（Ｘ％）で提示する説明文が用いられているが、向上幅αを複数のレベル（例えば大、中、小等）にわけて提示する説明分が用いられてもよい。

このように、説明文等の文章で推定結果を提示することで、全ての学習データセット１７を用いた学習を行うべきか否かについてのアドバイスを明示的に行うことが可能となる。

【０１０３】

図１０は、向上幅αの表示エリア３８のインターフェースの一例を示す模式図である。

図１０Ａに示す表示エリア３８には、全ての学習データセット１７を用いて行われる第２の予測モデル５２の生成処理を実行する実行ボタン３９が設けられる。そして実行ボタン３９の近くに第２の予測モデル５２の生成処理に要する処理時間と、期待される予測精度の向上幅αとが提示される。

これより、ユーザは、向上幅αと処理時間とを参照して、第２の予測モデル５２の生成処理を行うか否かを判断することが可能となる。また、実行ボタン３９を選択することで、そのまま全ての学習データセット１７を用いた生成処理が開始可能であるため、設定値を再度入力する必要等はない。

【０１０４】

図１０Ｂに示す表示エリア３８には、向上幅αの推定結果がそのまま提示される。ここでは、複数のレベルに分類された向上幅αの推定結果が文字データを用いて提示される。

推定結果を表す方法は限定されず、例えば複数のレベルを表すグラフィックス等を用いて向上幅αのレベルが表されてもよい。あるいは向上幅αの値を表すゲージやグラフ等が用いられてもよい。

これより、ユーザは、向上幅αのレベルや値を容易に把握することが可能となる。

このように、ＵＩ生成部２０は、予測精度の向上幅αを複数のレベルにわけて提示する評価画面３７や、予測精度の向上幅αの値を提示する評価画面３７を生成する。

【０１０５】

図６に戻り、向上幅α（評価画面３７）が提示されると、全ての学習データセット１７（全データセット）での学習処理を開始するか否かが判定される（ステップ１０７）。すなわち、第２の予測モデル５２を生成するか否かが判定される。

例えば、向上幅αが高くユーザが全ての学習データセット１７での学習を選択した場合（ステップ１０７のＹｅｓ）、全ての学習データセット１７での学習や評価が開始される（ステップ１０８）。本実施形態では、例えばサーバ装置３０に学習データセット１７及び設定値等が出力され、サーバ装置３０により第２の予測モデル５２を生成する一連の処理が実行される。あるいは、端末装置１０の予測モデル生成部２１により、第２の予測モデル５２が生成されてもよい。なお、第２の予測モデル５２が生成された後は、その評価画面等が表示される。

また例えば、向上幅αが低くユーザが全ての学習データセット１７での学習を選択しなかった場合（ステップ１０７のＹｅｓ）、予測モデル５０を生成する処理が終了する。

【０１０６】

これにより、例えば大規模なデータセットから学習する際にローカルＰＣ（端末装置１０）を長時間占有することなく、またクラウド上の従量課金制サーバ（サーバ装置３０）等を長時間使用することなく、短時間で予測モデル５０と予測精度及び全データセットで学習した際の精度の目安を知ることが可能となる。これにより、予測モデル５０を効率的に生成することが可能となる。

【０１０７】

本技術に係るモデル生成システム１００（予測分析ツール）の適用例について具体的な

事例を挙げて説明する。
［適用例１］
大規模データで予測モデルを学習させる際に、有用な特徴量の組合せを特定した後に、クラウド上の従量課金制のサーバ装置３０を用いて全データでの学習を実行する事例。
ここでは、保険会社において、顧客がどのような保険商品を好むか予測する予測モデル５０を構築するものとする。
【０１０８】
このケースでは、例えば学習データセット１７として用いる顧客データは、顧客の入金や保険商品に対する手続きのログが含まれる膨大なデータである。このため、全ての学習データセット１７を用いて学習を行うと６時間～１２時間程度の時間がかかってしまい、業務時間内に完了させるといったことが難しい。
さらに、手続きのログの種類は数百種類あり多岐に渡る。このため、学習に用いる特徴量として設定する行動の種類（特徴量の組合せ）を特定することが難しい。
【０１０９】
例えば、ユーザにより、普段の業務の仮説に基づいて、入力データとして用いる特徴量の組合せが１０パターン程度用意される。その上で、モデル生成システム１００を用いて、ユーザが用意した各パターンについて、部分データセット１８での予備的な学習が実行される。これにより、各パターンについて、全ての学習データセット１７で学習した時の予測精度（向上幅α）が推定される。
この予備的な学習（図６のステップ１０３等）は、学習データセット１７からサンプリングされた部分データセット１８に対して行うため、１回あたり３０分程度で完了する。１０パターンの特徴量の組合せを３０分おきに学習することで、業務時間内に特に有用な特徴量の組合せのパターンを３つ程度に絞ることが可能である。
【０１１０】
上記の予備的な学習で絞り込んだ３つ程度の有用な特徴量の組合せのパターンの各々について、全ての学習データセット１７を用いて第２の予測モデル５２が生成される。この処理は、例えば夜間や土日といった時間を利用して、従量課金制のサーバ装置３０を用いて長時間かけて実行される。
例えば、翌朝（あるいは週明け）に出勤したユーザにより、サーバ装置３０で学習させた第２の予測モデル５２の学習結果が確認される。そして、予測精度等が最も優れたモデルが、最終的に使用するモデルとして決定される。
【０１１１】
このように、モデル生成システム１００を用いることで、見込みのあるパラメータの候補等を短時間で絞り込むことが可能である。これにより、業務時間を無駄にすることなく、予測モデル５０を効率的に生成することが可能となる。
【０１１２】
［適用例２］
顧客のログから顧客がサービスに払う金額が予測出来るかどうか試行錯誤する事例。
ここでは、ウェブ上で提供されるサービスにおいて、顧客が１カ月の間にサービスで使用する金額を予測する予測モデルを構築するものとする。このような予測モデルを構築することで、例えば、使用金額の少ないユーザに対してクーポンを発行するといった対策を行うことが可能となり、顧客にサービスの使用を促すことが可能になると期待される。
【０１１３】
このケースでは、例えばアクセス時間等を記録した顧客のログデータが学習データセット１７として用いられる。
なおログデータは、膨大なデータ量であり、またノイズも多く混ざっていると考えられる。このため、ログデータを学習したからと言って、そこから顧客が使用する金額を実際に予測出来るどうかは不明である。
一方で、顧客の使用する金額は、サービスにおいてＫＰＩ（Key Performance Indicator）の１つである。このため、もし顧客の使用する金額が予測可能であるならば、ビジネ

ス的な価値は大きく、可能な限り予測モデルの構築を試みるものとする。
【０１１４】
例えば、数十ギガバイトのログデータが存在し、全てのログデータ（学習データセット１７）を使用して学習を行うと、学習処理に数日程度時間を要する事が判明したとする。
そこで、モデル生成システム１００を用いて、まずは全てのログデータのうち、サンプリングした一部のログデータ（部分データセット１８）で学習処理等が実行される。この学習処理は、例えば６時間程度で行われる。
【０１１５】
部分データセット１８による学習結果を参照すると、第１の予測モデル５１での誤差率中央値が１２０％となり、十分な予測精度が出ていないことが判明した。さらに、全てのデータセットを使用した時の精度も、誤差率中央値が１００％程度になると示唆されており、期待した精度が得られないことが判明した。
このような場合は、データ数を増やしてローカルの端末装置１０やクラウド上のサーバ装置３０で処理を実行したとしても、時間や費用が無駄になってしまう。このため、ログデータから顧客の使用する金額を予測する予測モデルの構築は断念されることになる。
【０１１６】
上記したように、顧客が使用する金額を直接予測できないことがわかった。そこで問題設定を変更し、顧客が１カ月に１０００円以上のお金をサービスで支払うかどうかを分類する二値分類を行う予測モデルについて検討した。
具体的には、モデル生成システム１００を用いて、上記の二値分類を行う予測モデルについて、全てのログデータからサンプリングした一部のデータセットで学習したところ、ＡＵＣが０．６５となった。さらに、全てのログデータで学習することで、ＡＵＣが０．７まで上がることが示唆されたとする。
これは、実用可能な精度であるため、実際にクラウド上のサーバ装置３０を用いて全てのログデータを用いて学習処理を実行し、ＡＵＣが０．７１の予測モデルが得られた。
【０１１７】
このように、問題設定（予測モデルのターゲット）を顧客が１カ月に１０００円以上のお金をサービスで支払うかどうかの二値分類に変更する事で、実用可能な予測が出来る事がわかった。これにより、例えば月に１０００円以下しかサービスにお金を払わない確率の高い顧客に対して、クーポンや割引等を発行し、顧客の消費金額を促す施策を開始することが可能となる。
この適用例では、問題設定を試行錯誤して適切な問題設定を見つける間に、モデル生成システム１００が用いられる。これにより、実際に全てのデータを使った学習を行わなくても予測精度が推定されるため、不要な学習時間や費用を費やすことなくモデルを構築することが可能となっている。
【０１１８】
［適用例３］
大規模データでの学習を行う際に、初めにローカルの端末装置１０を用いて全ての学習データセット１７で学習した時の精度を見積もり、実用可能な見込みが得られた場合に従量課金制のサーバ装置３０で全ての学習データセット１７での学習を実行する事例。
【０１１９】
例えば、端末装置１０を用いて、部分データセット１８での予測モデルの学習が行われる。その後、学習結果等が提示され、全ての学習データセット１７で学習したときに実用に耐えうる予測精度が出るかどうかがユーザにより確認される。例えば、全ての学習データセット１７で学習するとＡＵＣが０．７２であることが予測されたとする。この場合、期待される予測精度は実用に達しているとして、クラウド上のサーバ装置３０を用いて、全ての学習データセット１７を使用した学習処理を行うことが決定される。
【０１２０】
実際に、サーバ装置３０を用いて学習処理を行った結果、ＡＵＣが０．７１となる想定通りの予測モデルが構築されたとする。この場合、予測モデルは実用に耐えうるモデルで

あるとして、本番環境に投入する事が決定される。
　このように、モデル生成システム１００では、大規模データでの学習を行う際に予め全データ使用時の精度を推定可能である。この推定結果を参照することで、ユーザは、サーバ装置３０等の演算リソースを効率的に利用することが可能となる。
【０１２１】
　図１１は、サーバ装置３０での演算を含む学習処理の一例を示すタイムチャートである。図１１には、例えば適用例３で説明した大規模データでの学習を行う事例における、モデル生成システム１００での処理の流れが示されている。
　まず、学習に使用する学習データセット１７が読み込まれた状態で、ユーザにより学習ボタンが押下され、端末装置１０に学習処理を開始する旨の指示が入力される（ステップ２０１）。
　このとき、端末装置１０では、学習データセット１７のデータ容量が算出され、学習時間等が算出される。そして、データ容量や学習時間が閾値を超えて大きい場合等には、データが巨大であるため一部のデータ（部分データセット１８）で学習する旨を伝えるメッセージが表示される（ステップ２０２）。
【０１２２】
　端末装置１０において部分データセット１８での学習処理が実行される（ステップ２０３）。このように、図１１に示す例では、端末装置１０により、学習データセット１７のデータのサイズ等に応じて、部分データセット１８での学習処理が自動的に選択され実行される。なお、部分データセット１８での学習は、ユーザの確認後に実行されてもよい。
　部分データセット１８での学習処理が完了すると、その学習結果（第１の予測モデル５１の評価結果）と、全データで学習した場合に想定される推定予測精度（向上幅α）とが表示される（ステップ２０４）。
【０１２３】
　推定予測精度が高く、ユーザが全ての学習データセット１７を用いた学習処理を実行すると判断したとする。この場合、所定の実行ボタンが押下され、端末装置１０にクラウド（サーバ装置３０）での学習を実行させる旨の指示が入力される（ステップ２０５）。そして端末装置１０により、全ての学習データセット１７と予測モデルの設定値等のデータとがサーバ装置３０にアップロードされる（ステップ２０６）。
【０１２４】
　サーバ装置３０では、全ての学習データセット１７での学習処理が実行される（ステップ２０７）。サーバ装置３０は、一般に高い演算能力を有するため、端末装置１０で行うよりも短時間で学習処理を完了することが可能である。なお、サーバ装置３０で学習処理が実行されている間、端末装置１０には演算負荷がかからない。従って、ユーザはこの時間を利用して端末装置１０に他の処理等を実行させることが可能である。
【０１２５】
　全ての学習データセット１７での学習処理が完了すると、その学習結果（第２の予測モデル５２の評価結果）がサーバ装置３０から端末装置１０に送信される（ステップ２０８）。そして端末装置１０により、全ての学習データセット１７での学習結果を含む評価画面が生成され、表示部に表示される（ステップ２０９）。
　このように、全てのデータを使った本番の学習を行う前に、予測精度の推定結果が提示される。これによりユーザは、本番の学習を行うべきか否かを判断することが可能である。特に大規模なデータでの学習を行う場合等には不要な演算時間や費用を抑制し、必要な演算のみを実行させることが可能となる。これにより、予測モデルの生成処理の効率を大幅に向上することが可能となる。
【０１２６】
　以上、本実施形態に係る制御部１５では、学習データセット１７のうち、部分データセット１８のメタ特徴量Ｆが取得される。このメタ特徴量Ｆに基づいて、学習データセット１７を用いて予測モデル５０（第１の予測モデル５１）を生成した場合の予測精度を表す精度情報（向上幅α）が推定される。これにより、例えば学習データセット１７を用いる

べきか否かを判断することが可能となり、予測モデル５０を効率的に生成することが可能となる。

【０１２７】

機械学習では、一般に学習データ数を増やすほど予測精度が向上することが知られている。一方で、データ数が増えるにつれて学習に必要な時間が増加してしまう。

一例として、パラメータ探索や特徴量探索を行うような場合には、学習時間の増大が問題となる場合が多い。例えば、非専門家向けに提供されている予測分析サービス等では、パラメータや特徴量の探索が必須である。このため、例えば数百メガバイトを超える大きなデータセットを学習する際には、パラメータ探索等の過程で多くの時間を要してしまうことが考えられる。

【０１２８】

データ数を増やすことによる予測精度の向上の度合いを推定する方法として、異なるサイズの複数のデータセットに対して学習を行う方法が挙げられる。この場合、各データセットのデータ数とテストデータに対する予測精度との関係を調べることで、データ数を増やしたときの予測精度の向上幅が推定される。しかしながら、この方法では、複数のデータセットを対象とするため、複数回（例えば５－１０回程度）の学習を行う必要がある。このため、短い学習時間で予測精度を把握するという目的にも関わらず、予測精度を推定すること自体に時間がかかってしまう恐れがある。

【０１２９】

本実施形態では、学習データセット１７の一部である部分データセット１８のメタ特徴量Ｆから、学習データセット１７で学習させた予測モデルの予測精度の向上幅αが推定される。メタ特徴量Ｆは、部分データセット１８を用いた一度の学習から算出される。

【０１３０】

これにより、短い時間で、全ての学習データセット１７で学習させた場合の予測モデルの側精度が推定可能である。従ってユーザは、ローカルの端末装置１０ですぐに予測結果の目安を知ることが可能となり、全データでの学習を実行するか否かを適切に判断することが可能となる。

例えばデータが大規模な場合、パラメータや特徴量を探索する場合、あるいは問題設定を試行錯誤する場合等には、不要な学習を行わずに、短時間で全データセットから学習した際の予測精度を見積もることが可能となる。

【０１３１】

またユーザは、端末装置１０を長時間占有することなく、全データで学習した際の精度の見積もりを知ることが可能である。これにより、例えば業務中は一部のデータ（部分データセット１８）で学習を実行して全データ使用時のおおよその予測精度を把握し、夜間や休日に全データでの学習を実行するなどの使い方が可能となる。

【０１３２】

また、あらかじめ学習が上手くいかない（予測精度が低い、向上が見込めない等）と推定されるデータセットに関してはクラウドで学習を回す必要がなくなる。従って、ユーザは、効果があると推定された時だけ、サーバ装置３０での学習を実行するといったことが可能となる。これにより、従量課金制のサーバ装置３０に無駄な費用を払う必要がなくなり、開発コストを抑えることが可能となる。

【０１３３】

このように、本実施形態では、ローカルの端末装置１０を長時間占有することなく、もしくはクラウド上のサーバ装置３０を長時間占有することなく、予測精度の見積もりを得ることが出来る。

これにより、例えば全ての学習データセット１７で半日～１日の長時間の学習を行ったが、想定した精度が出ずに時間やサーバ代を無駄に使用するといった事態を回避することが可能となる。

また、データセットの精度を改善するにあたりデータ数を増やしたときの予測精度の見積もりが得られれば精度改善の指針を得ることも可能である。すなわち、予測精度の向上

幅α等を参照して、向上幅αが高くなるようなデータセットを開発するといったことも可能である。

【０１３４】

＜その他の実施形態＞

本技術は、以上説明した実施形態に限定されず、他の種々の実施形態を実現することができる。

【０１３５】

上記では、本技術に係る情報処理装置の一実施形態として、単体の制御部１５（端末装置１０）を例に挙げた。しかしながら、制御部１５とは別に構成され、有線又は無線を介して制御部１５に接続される任意のコンピュータにより、本技術に係る情報処理装置が実現されてもよい。例えばクラウドサーバにより、本技術に係る情報処理方法が実行されてもよい。あるいは制御部１５と他のコンピュータとが連動して、本技術に係る情報処理方法が実行されてもよい。

【０１３６】

すなわち本技術に係る情報処理方法、及びプログラムは、単体のコンピュータにより構成されたコンピュータシステムのみならず、複数のコンピュータが連動して動作するコンピュータシステムにおいても実行可能である。なお本開示において、システムとは、複数の構成要素（装置、モジュール（部品）等）の集合を意味し、すべての構成要素が同一筐体中にあるか否かは問わない。したがって、別個の筐体に収納され、ネットワークを介して接続されている複数の装置、及び、１つの筐体の中に複数のモジュールが収納されている１つの装置は、いずれもシステムである。

【０１３７】

コンピュータシステムによる本技術に係る情報処理方法、及びプログラムの実行は、例えば部分データセットの特徴量の取得、精度情報の推定等が、単体のコンピュータにより実行される場合、及び各処理が異なるコンピュータにより実行される場合の両方を含む。また所定のコンピュータによる各処理の実行は、当該処理の一部または全部を他のコンピュータに実行させその結果を取得することを含む。

【０１３８】

すなわち本技術に係る情報処理方法及びプログラムは、１つの機能をネットワークを介して複数の装置で分担、共同して処理するクラウドコンピューティングの構成にも適用することが可能である。

【０１３９】

以上説明した本技術に係る特徴部分のうち、少なくとも２つの特徴部分を組み合わせることも可能である。すなわち各実施形態で説明した種々の特徴部分は、各実施形態の区別なく、任意に組み合わされてもよい。また上記で記載した種々の効果は、あくまで例示であって限定されるものではなく、また他の効果が発揮されてもよい。

【０１４０】

本開示において、「同じ」「等しい」「直交」等は、「実質的に同じ」「実質的に等しい」「実質的に直交」等を含む概念とする。例えば「完全に同じ」「完全に等しい」「完全に直交」等を基準とした所定の範囲（例えば±１０％の範囲）に含まれる状態も含まれる。

【０１４１】

なお、本技術は以下のような構成も採ることができる。

（１）予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得する取得部と、

前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定する推定処理部と

を具備する情報処理装置。

（２）（１）に記載の情報処理装置であって、

前記推定処理部は、前記精度情報として、前記部分データセットを用いて生成される前

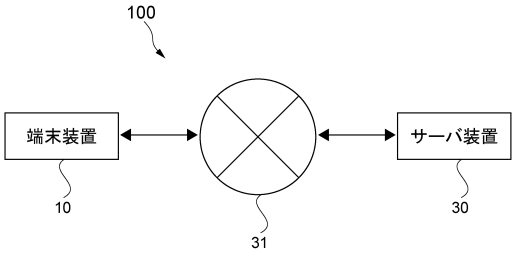
記予測モデルの予測精度に対する前記全データセットを用いて生成される前記予測モデルの予測精度の変化を推定する
情報処理装置。
（３）（２）に記載の情報処理装置であって、
前記推定処理部は、前記予測精度の変化を推定する推定モデルを用いて構成される
情報処理装置。
（４）（３）に記載の情報処理装置であって、
前記推定モデルは、所定のデータセットの一部のデータセットの特徴量と、所定の予測モデルを前記所定のデータセットの全部及び一部を用いて生成した場合に生じる予測精度の変化との関係を学習したモデルである
情報処理装置。
（５）（３）又は（４）に記載の情報処理装置であって、
前記推定モデルは、前記予測精度の変化量を複数のレベルに分類する分類モデルである
情報処理装置。
（６）（３）又は（４）に記載の情報処理装置であって、
前記推定モデルは、前記予測精度の変化量を複数のレベルに分類する分類モデルをルールベースで近似したモデルである
情報処理装置。
（７）（３）又は（４）に記載の情報処理装置であって、
前記推定モデルは、前記予測精度の変化量を推定する回帰モデルである
情報処理装置。
（８）（１）から（７）のうちいずれか１つに記載の情報処理装置であって、
前記部分データセットの特徴量は、前記部分データセットの内容に応じた第１の特徴量を含み、
前記取得部は、前記部分データセットを解析することで前記第１の特徴量を算出する
情報処理装置。
（９）（８）に記載の情報処理装置であって、
前記第１の特徴量は、前記部分データセットに含まれるデータの数、前記データに含まれる特徴量の数、前記データの数と前記データに含まれる特徴量の数との比率の少なくとも１つを含む
情報処理装置。
（１０）（１）から（９）のうちいずれか１つに記載の情報処理装置であって、
前記部分データセットの特徴量は、前記部分データセットを用いて生成される前記予測モデルに応じた第２の特徴量を含み、
前記取得部は、前記部分データセットを用いた前記予測モデルの生成処理を実行することで前記第２の特徴量を算出する
情報処理装置。
（１１）（１０）に記載の情報処理装置であって、
前記部分データセットは、互いに用途の異なる複数のデータグループを含み、
前記第２の特徴量は、前記複数のデータグループの各々に対する前記部分データセットを用いて生成される前記予測モデルの予測値を評価する評価値、又は前記評価値を比較した比較値の少なくとも一方を含む
情報処理装置。
（１２）（１１）に記載の情報処理装置であって、
前記複数のデータグループは、学習データのグループと、検証データのグループと、テストデータのグループとを含む
情報処理装置。
（１３）（１１）又は（１２）に記載の情報処理装置であって、
前記評価値は、前記部分データセットを用いて生成される前記予測モデルの予測値に関する誤差中央値、平均二乗誤差、及び誤差率中央値の少なくとも１つを含む

情報処理装置。
（１４）（１１）から（１３）のうちいずれか１つに記載の情報処理装置であって、
前記比較値は、前記複数のデータグループのうち２つのデータグループについて算出された前記評価値の差分又は比率の少なくとも一方を含む
情報処理装置。
（１５）（１）から（１４）のうちいずれか１つに記載の情報処理装置であって、さらに、
前記精度情報を提示する画面を生成する画面生成部を具備する
情報処理装置。
（１６）（１５）に記載の情報処理装置であって、
前記推定処理部は、前記精度情報として、前記部分データセットを用いて生成される前記予測モデルの予測精度に対する前記全データセットを用いて生成される前記予測モデルの予測精度の変化を推定し、
前記画面生成部は、前記予測精度の変化量を複数のレベルにわけて提示する画面、または前記予測精度の変化量の値を提示する画面の少なくとも一方を生成する
情報処理装置。
（１７）（１５）又は（１６）に記載の情報処理装置であって、
前記画面生成部は、前記部分データセットを用いた前記予測モデルの生成処理の実行を選択するための選択画面を生成し、
前記取得部は、前記生成処理の実行が選択された場合に、前記生成処理を実行して前記部分データセットの特徴量を算出し、
前記推定処理部は、前記部分データセットの特徴量に基づいて前記精度情報を推定する
情報処理装置。
（１８）予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得し、
前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定する
ことをコンピュータシステムが実行する情報処理方法。
（１９）予測モデルの生成に用いる全データセットの一部である部分データセットの特徴量を取得するステップと、
前記部分データセットの特徴量に基づいて、前記全データセットを用いて生成される前記予測モデルの予測精度を表す精度情報を推定するステップと
をコンピュータシステムに実行させるプログラム。
【符号の説明】
【０１４２】
Ｆ…メタ特徴量
１０…端末装置
１４…記憶部
１５…制御部
１６…制御プログラム
１７…学習データセット
１８…部分データセット
２０…ＵＩ生成部
２１…予測モデル生成部
２２…メタ特徴量算出部
２３…精度推定部
３０…サーバ装置
３５…設定画面
３７…評価画面
４０…推定モデル
５０…予測モデル

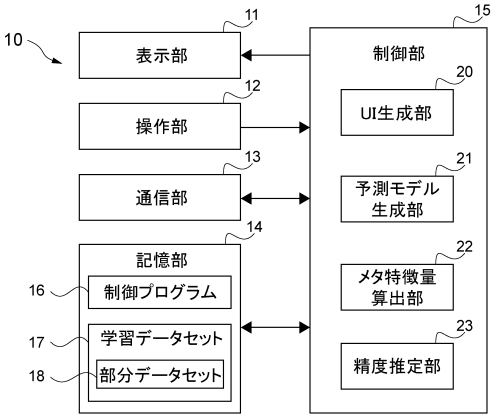
５１…第１の予測モデル
５２…第２の予測モデル
１００…モデル生成システム

【図面】

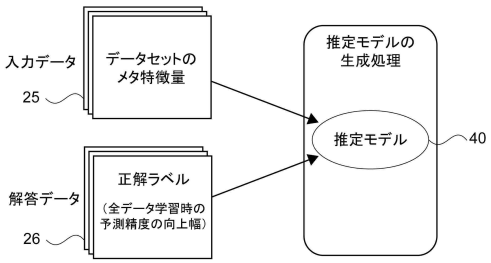
【図１】



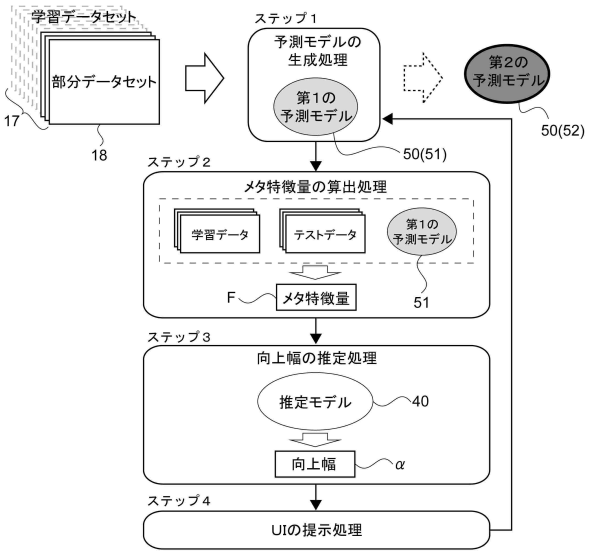
【図２】



【図３】



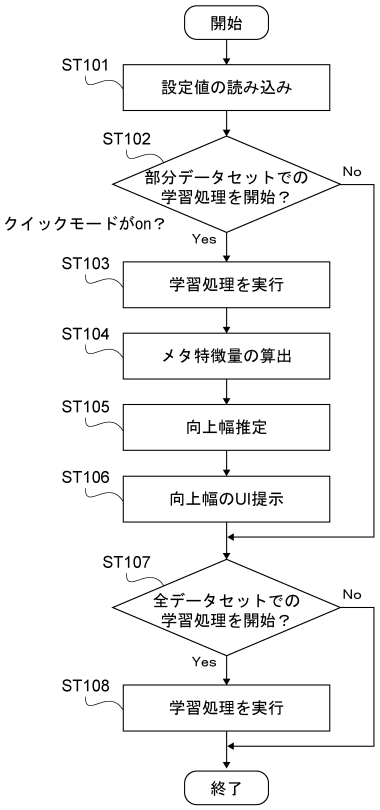
【図４】



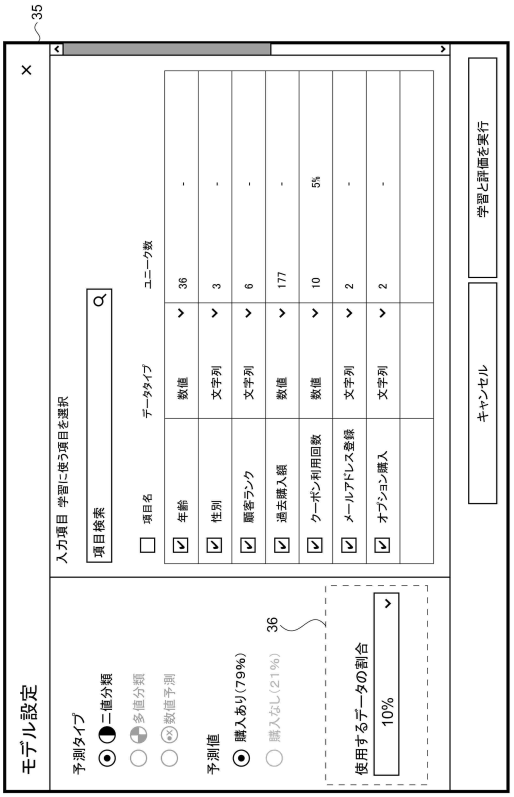
【図５】

No.	項目	内容
F1	データ数	データセットに含まれるデータ数
F2	特徴量数	データセットに含まれる特徴量数
F3	特徴量数/データ数	データ数と特徴量数の比率
F4	展開後の特徴量数	OneHotエンコーディング等の前処理を済ませた後の学習データに使用する特徴量数
F5	Iteration数に応じたテストデータの誤差中央値の変化	テストデータに対するIterationが収束した時の半分時点での誤差中央値と、最終的な誤差中央値の差
F6	学習/検証/テストデータの誤差中央値	学習済みモデルで予測した際の、学習/検証/テストデータの誤差中央値（MAE）の値
F7	学習/検証/テストデータの平均二乗誤差	学習済みモデルで予測した際の、学習/検証/テストデータの平均二乗誤差（RMSE）の値
F8	学習/検証/テストデータの誤差率中央値	学習済みモデルで予測した際の、学習/検証/テストデータの誤差率中央値（MAPE）の値
F9	正解値の分散	予測対象ラベルの値の分散
F10	予測値の分散	予測した値の分散
F11	学習データとテストデータとの誤差中央値の差	学習データに対する誤差中央値とテストデータに対する誤差中央値の差
F12	学習データとテストデータとの誤差中央値の比率	学習データに対する誤差中央値とテストデータに対する誤差中央値の比率
F13	検証データとテストデータとの誤差中央値の差	検証データに対する誤差中央値とテストデータに対する誤差中央値の差
F14	検証データとテストデータとの誤差中央値の比率	検証データに対する誤差中央値とテストデータに対する誤差中央値の比率
F15	学習データと検証データとの誤差中央値の差	学習データに対する誤差中央値と検証データに対する誤差中央値の差
F16	学習データと検証データとの誤差中央値の比率	学習データに対する誤差中央値と検証データに対する誤差中央値の比率

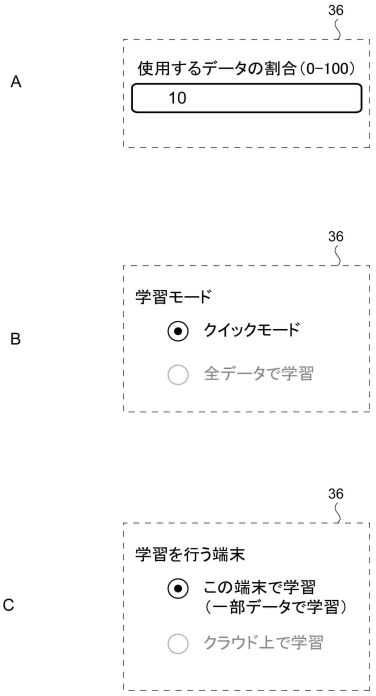
【図６】



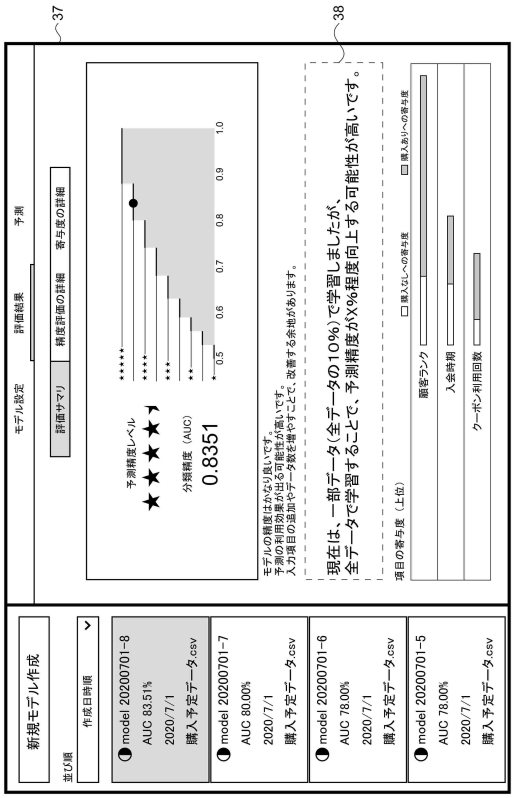
【図７】



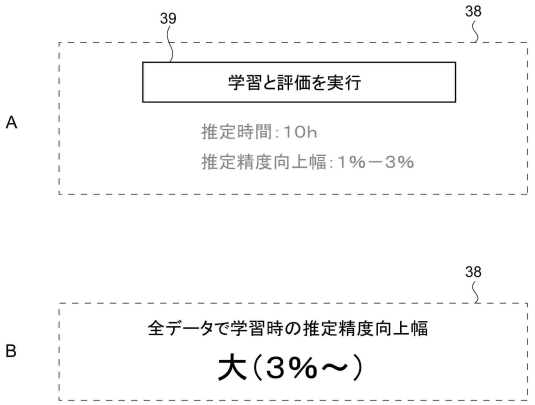
【図８】



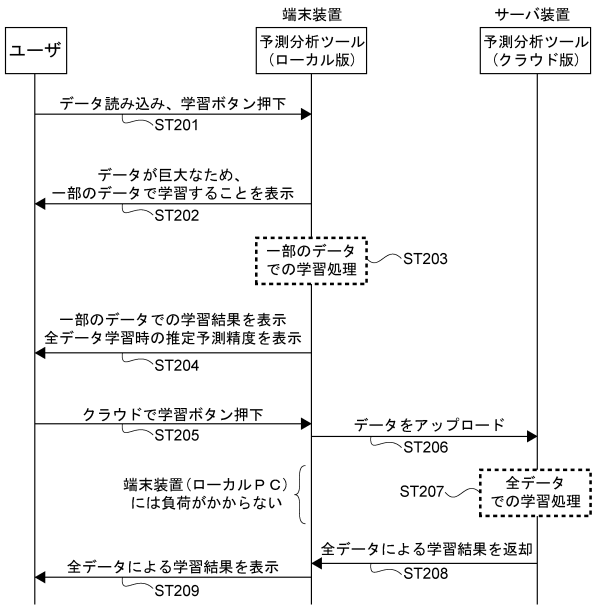
【図９】


37
38
モデル設定
評価結果
予測
評価サマリ
精度評価の詳細
寄与度の詳細
予測精度レベル
分類精度（AUC）
0.8351
モデルの精度はかなり良いです。
予測の利用効果が出る可能性が高いです。
入力項目の追加やデータ数を増やすことで、改善する余地があります。
現在は、一部データ（全データの10%）で学習しましたが、
全データで学習することで、予測精度が○%程度向上する可能性が高いです。
購入あり への寄与度
購入なし への寄与度
項目の寄与度（上位）
顧客ランク
入会時期
クーポン利用回数
新規モデル作成
並び順
作成日時順
model 20200701-8
AUC 83.51%
2020/7/1
購入予定データ.csv
model 20200701-7
AUC 80.00%
2020/7/1
購入予定データ.csv
model 20200701-6
AUC 78.00%
2020/7/1
購入予定データ.csv
model 20200701-5
AUC 78.00%
2020/7/1
購入予定データ.csv

【図１０】


39
38
学習と評価を実行
推定時間：10h
推定精度向上幅：1%－3%
A
38
全データで学習時の推定精度向上幅
大（3%～）
B

【図１１】


端末装置
サーバ装置
ユーザ
予測分析ツール（ローカル版）
予測分析ツール（クラウド版）
データ読み込み、学習ボタン押下
ST201
データが巨大なため、
一部のデータで学習することを表示
ST202
一部のデータでの学習処理
ST203
一部のデータでの学習結果を表示
全データ学習時の推定予測精度を表示
ST204
クラウドで学習ボタン押下
ST205
データをアップロード
ST206
端末装置（ローカルＰＣ）
には負荷がかからない
ST207
全データでの学習処理
全データによる学習結果を返却
ST208
全データによる学習結果を表示
ST209

フロントページの続き

(56)参考文献 特開２０１８－１７３８１３（ＪＰ，Ａ）
国際公開第２０１７／１８３５４８（WO，Ａ１）
FIGUEROA, Rosa L. et al.，"Predicting sample size required for classification performance"，BMC MEDICAL INFORMATICS AND DECISION MAKING [online]，No. 8，2012年，［2024年05月21日検索］，インターネット＜ＵＲＬ：https://doi.org/10.1186/1472-6947-12-8＞，DOI: 10.1186/1472-6947-12-8

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／０２－　３／１０
Ｇ０６Ｎ　２０／００－９９／００