

明 細 書

発明の名称：

予測関数作成装置、予測関数作成方法、及びコンピュータ読み取り可能な記録媒体

技術分野

[0001] 本発明は、説明変数と目的変数とを含むサンプルの集合から、目的変数を導く予測関数を作成するための、予測関数作成装置、予測関数作成方法、及びこれらを実現するためのプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

背景技術

[0002] 種々のビジネスの分野において、入手可能なデータから、将来の予測を行うことは業務改善に有用である。例えば商店において、直近2週間の売り上げデータから将来の売り上げの予測を行うことができれば、在庫管理を適切に行うことができる。また、コールセンターでの対応の業務記録から、お客様からの苦情が来るかを予測し、どのような対応をすると、苦情が起こるのかを予測できれば苦情を減らすことができる。

[0003] ここで、本明細書では、直近2週間の売り上げデータ、コールセンターでの対応の業務記録など、予測の手がかりとするデータを「説明変数」と呼び、将来の売り上げ、苦情が来るかなど予測したい変数を「目的変数」と呼ぶことにする。また、「予測」とは、説明変数の関数を作り、目的変数の予測値を求めることとする。

[0004] また、予測の手がかりとして、過去のデータが使えるとする。過去のデータは、説明変数と目的変数との組からなるサンプルの集合である。以下、このサンプルの集合を「訓練データ」と呼ぶこととする。

[0005] ところで、訓練データを用いて予測を行う方法の一つとして、機械学習を用いる方法が挙げられる。機械学習とは、訓練データに基づいて、説明変数を入力として、目的変数の予測値を出力する関数を作成することである。

[0006] 但し、機械学習の適用には問題がある。それは、訓練データの説明変数の中に欠損値が生じた場合に適用できないことである。例えば、ある時間帯に特定の商品が棚になかった場合、特定の商品の売り上げは欠損してしまうので、機械学習は適用できなくなる。また、対応したオペレータが記録を忘れたことにより、業務記録の一部が欠損した場合も、機械学習は適用できなくなる。即ち、機械学習を用いる方法の多くは、欠損値を含むデータに適用できないという問題がある。

[0007] 一方、欠損値を平均値で補完する方法、他の説明変数の値から欠損値を予測して補完する方法も存在するが、補完に大きな誤差が生じてしまった場合に、これらの方法では、目的変数の予測に余計な誤差を生じさせてしまう。

[0008] このような問題を解決するため、例えば、非特許文献 1 は、入力となる訓練データの一部において説明変数の欠損が含まれている場合であっても、予測を行なうことができる予測システムを開示している。図 11 は、従来からの予測システムの一例を示すブロック図である。

[0009] 図 11 に示すように、予測システム 20 は、データ分割手段 21 と、予測関数学習手段 22 とを備えている。データ分割手段 21 は、訓練データが入力されると、これを分割し、分割した訓練データを出力する。予測関数学習手段 22 は、分割された訓練データが入力されると、それぞれに対して学習を行なって予測関数を生成し、生成した予測関数を出力する。

[0010] また、図 11 に示す従来からの予測システムは、以下のように動作する。まず、データ分割手段 21 は、入力された訓練データの各サンプルに対し、どの説明変数が欠損しているか（以下「欠損の仕方」と呼ぶ）を参照し、欠損の仕方が同じとなっているサンプルに同じラベルを付与する。

[0011] 次に、予測関数学習手段 22 は、データ分割手段 21 によって出力されたラベル付きの訓練データを入力とし、ラベル毎に、同ラベルが付与されているサンプルの集合のみを訓練データとした機械学習を行い、これによって予測関数を出力する。

先行技術文献

非特許文献

- [0012] 非特許文献1: "Handling Missing Values when Applying Classification Models", Maytal Saar-Tsechansky and Foster Provost, Journal of Machine Learning Research 8 (2007) 1625-1657

発明の概要

発明が解決しようとする課題

- [0013] しかしながら、上記非特許文献1に開示された予測システムには問題点が存在する。問題点は、サンプル毎の欠損の仕方のバリエーションが多い場合に予測値の精度が落ちてしまう事である。なぜなら、欠損の仕方のバリエーションが多い場合は、それに合わせて、データ分割手段21が付与するラベルの種類も多くなってしまう。そのため、同ラベルのサンプルの集合におけるサンプルサイズが小さくなってしまう。そして、サンプルサイズが小さい場合、予測関数により作成される予測値の精度は低くなってしまう。
- [0014] このため、説明変数の欠損の仕方のバリエーションが多い場合でも、欠損を含む入力のもとで、高い精度で予測を実行できる、予測装置の実現が求められている。
- [0015] [発明の目的]
- 本発明の目的の一例は、上記問題を解消し、説明変数の欠損の仕方のバリエーションが多い場合でも、高い精度で予測を実行でき得る、予測関数作成装置、予測関数作成方法、及びコンピュータ読み取り可能な記録媒体を提供することにある。

課題を解決するための手段

- [0016] 上記目的を達成するため、本発明の一側面における予測関数作成装置は、説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するための装置であって、
- 各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該

ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合せて重みを設定する、クラスタリング部と、

設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、子モデル作成部と、

ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、混合モデル作成部と、を備えていることを特徴とする。

[0017] また、上記目的を達成するため、本発明の一側面における予測関数作成方法は、説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するための方法であって、

(a) 各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合せて重みを設定する、ステップと、

(b) 設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、ステップと、

(c) ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、ステップと、を備えていることを特徴とする。

[0018] 更に、上記目的を達成するため、本発明の一側面におけるコンピュータ読み取り可能な記録媒体は、コンピュータによって、説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

前記コンピュータに、

(a) 各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合せて重みを設定する、ステップと、

(b) 設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、ステップと、

(c) ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、ステップと、
を実行させる命令を含む、プログラムを記録していることを特徴とする。

発明の効果

[0019] 以上のように、本発明によれば、説明変数の欠損の仕方のバリエーション、即ち、説明変数の欠損パターンが多い場合でも、高い精度で予測を実行することができる。

図面の簡単な説明

[0020] [図1]図1は、本発明の実施の形態における予測関数作成装置の概略構成を示すブロック図である。

[図2]図2は、図1に示す予測関数作成装置の構成を更に具体的に示すブロック図である。

[図3]図3は、本発明の実施の形態における予測関数作成装置の動作を示すフロー図である。

[図4]図4は、本発明の実施の形態における予測関数作成装置を実現するコンピュータの一例を示すブロック図である。

[図5]図5は、実施例で用いられる訓練データの一例を示す図である。

[図6]図6は、実施例において想定される非欠損変数集合の種類毎の計算結果の一例を示す図である。

[図7]図7は、実施例における欠損変数パターンと候補との距離についての計算結果の一例を示す図である。

[図8]図8は、実施例において各ラベルに設定された重みの一例を示す図である。

[図9]図9は、実施例における部分訓練データの一例を示す図である。

[図10]図10は、実施例で得られた最終予測関数の具体例を示す図である。

[図11]図11は、従来からの予測システムの一例を示すブロック図である。

発明を実施するための形態

[0021] (発明の概要)

本発明は、上記非特許文献1に開示されたシステムと同様に、予測が難しい説明変数（欠損値）の補完を行なったときに、誤差の増大を防ぐことを目的としている。本発明でも、上記非特許文献1に開示されたシステムと同様に、説明変数の欠損パターンに応じて、異なる予測関数が作成される。サンプル毎に、最適な予測関数を用いることができるので、予測精度の低下が抑制される。

[0022] 但し、本発明では、上記非特許文献1に開示されたシステムと異なり、説明変数の欠損パターンのクラスタリングが行われる。それにより、各予測関数に用いるサンプルの数を増やすことができるので、欠損パターンの種類が多くなった場合であっても、予測精度の低下が抑制される。

[0023] また、本明細書において、「サンプル」とは、説明変数と目的変数とを含むデータを意味する。通常、サンプルには、複数の説明変数が含まれており、これらは説明変数ベクトルを構成している。「説明変数ベクトル」は、数値、又は欠損していることを示す記号、を要素とするベクトルである。

[0024] また、本明細書において、「訓練データ」とは、サンプルの集合を意味する。「子モデル」とは、本発明の予測関数作成装置が中間出力する予測関数群を意味する。「部分訓練データ」とは、訓練データのサンプルおよび変数の部分集合から作られるデータを意味する。部分訓練データは、子モデルを学習する時の入力となる。「最終予測関数」とは、説明変数の欠損パターン

それぞれについて目的変数の値を求めるための、予測関数を意味する。最終予測関数は、本発明の予測関数作成装置の出力に該当する。

[0025] (実施の形態)

以下、本発明の実施の形態 1 における、予測関数作成装置、予測関数作成方法、及びプログラムについて、図 1 ～図 4 を参照しながら説明する。

[0026] [装置構成]

最初に、図 1 を用いて、本実施の形態における予測関数作成装置 10 の概略構成について説明する。図 1 は、本発明の実施の形態における予測関数作成装置の概略構成を示すブロック図である。

[0027] 図 1 に示す本実施の形態における予測関数作成装置 10 は、説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、目的変数を導く予測関数を作成する装置である。図 1 に示すように、予測関数作成装置 10 は、クラスタリング部 11 と、子モデル作成部 12 と、混合モデル作成部 13 とを備えている。

[0028] クラスタリング部 11 は、各サンプルにおける説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与する。また、クラスタリング部 11 は、各ラベルに、それが付与されているサンプルにおける説明変数の欠損のパターンに合せて重みを設定する。なお、ここでいう「欠損のパターン」には、単独の説明変数が欠損している場合、複数の説明変数が欠損している場合、空集合の場合（いずれの説明変数も欠損していない場合）が含まれる。

[0029] 子モデル作成部 12 は、設定された重みに基づいて、訓練データの一部を部分訓練データとする。また、子モデル作成部 12 は、部分訓練データとなったサンプルにおける説明変数の欠損のパターンに基づいて、ラベル毎に、説明変数の中から、予測関数を構成する説明変数（以下、「予測変数」と表記する。）を決定する。

[0030] 混合モデル作成部 13 は、ラベル毎に決定された説明変数（予測変数）及び部分訓練データを用いて、説明変数の欠損のパターン毎に、目的変数を導

く予測関数（以下「最終予測関数」と表記する。）を作成する。

[0031] このように、本実施の形態では、説明変数の欠損のパターンに応じて、クラスタリングが行なわれ、クラスタリング結果から作成された部分訓練データを用いて、欠損のパターン毎のサンプル数を増加させることができる。このため、本実施の形態によれば、説明変数の欠損のパターンが多い場合でも、高い精度で予測を実行することができる。

[0032] 続いて、図2を用いて、本実施の形態における予測関数作成装置10の具体構成について説明する。図2は、図1に示す予測関数作成装置の構成を更に具体的に示すブロック図である。

[0033] 図2に示すように、本実施の形態では、予測関数作成装置10は、プログラムによってコンピュータ100上に構築されている。また、コンピュータ100には、ユーザの端末110が、ネットワーク等を介して接続されている。訓練データは、この端末110から送信されている。また、図2に示すように、予測関数作成装置10は、上述したクラスタリング部11、子モデル作成部12、及び混合モデル作成部13に加えて、欠損変数補完部14を更に備えている。

[0034] 本実施の形態では、クラスタリング部11は、各サンプルにおける説明変数の欠損のパターンそれぞれにおける欠損していない説明変数の集合の種類毎に、該当するサンプル群、即ち、各種類を構成する説明変数の集合が欠損していないサンプル群を特定する。

[0035] 次に、クラスタリング部11は、説明変数の欠損のパターン毎に、それぞれの欠損のパターンに該当しないサンプルと、特定したサンプル群それぞれと、の共通するサンプルを特定する。

[0036] そして、クラスタリング部11は、特定した共通するサンプルの数が一定以上となるサンプル群それぞれに対して、ラベルを付与する。その後、クラスタリング部11は、各ラベルに、欠損のパターン毎に、特定した共通するサンプルが多いほど値が大きくなるように、重みを設定する。

[0037] また、本実施の形態では、子モデル作成部12は、閾値以上の重みが設定

されたラベルが付与されているサンプル群を、部分訓練データとする。次に、子モデル作成部 12 は、重みが設定されたラベル毎に、そのラベルが付与されている部分訓練データにおける予測変数を決定する。このときの予測変数には、各ラベルにおいて、自身が欠損していないサンプルの数が多くなること、変数の数が多いことが求められる。

[0038] 本実施の形態では、混合モデル作成部 13 は、まず、重みが設定されたラベル毎に、そのラベルが付与されている部分訓練データを用いて、予測変数から目的変数を導く予測関数（以下「中間予測関数」と表記する。）を作成する。そして、混合モデル作成部 13 は、欠損のパターン毎に、そのパターンについて設定された重みと、この重みが設定されているラベルについての中間予測関数とを用いて、最終予測関数を作成する。

[0039] つまり、本実施の形態では、混合モデル作成部 13 は、子モデル作成部 12 によって抽出された部分訓練データから、予測変数を用いて、ラベル毎に、子モデルを学習する。そして、混合モデル作成部 13 は、ラベル毎に学習した子モデルを足し合わせることによって、混合モデルを学習する。

[0040] また、欠損変数補完部 14 は、部分訓練データにおいて、それに付与されているラベルの予測変数（説明変数）が欠損しているサンプルが存在する場合に、そのサンプルにおける欠損している予測変数を補完する。この場合、混合モデル作成部 13 は、補完された部分訓練データを用いて、中間予測関数を作成する。

[0041] 具体的には、欠損変数補完部 14 は、まず、部分訓練データを入力として、予測変数毎に、他の説明変数の値からその予測変数を推定する関数を学習する。そして、欠損変数補完部 14 は、子モデル作成部 12 から、補完すべきサンプルが入力されると、サンプルの欠損値を、学習した関数を用いて推定し、推定した値を含むサンプル（部分訓練データ）を、混合モデル作成部 13 に出力する。

[0042] [装置動作]

次に、本発明の実施の形態における予測関数作成装置 10 の動作について

図3を用いて説明する。図3は、本発明の実施の形態における予測関数作成装置の動作を示すフロー図である。以下の説明においては、適宜図1及び図2を参酌する。また、本実施の形態では、予測関数作成装置10を動作させることによって、予測関数作成方法が実施される。よって、本実施の形態における予測関数作成方法の説明は、以下の予測関数作成装置10の動作説明に代える。

[0043] 図3に示すように、最初に、クラスタリング部11は、入力された訓練データを構成する各サンプルに対し、説明変数の欠損のパターンを特定する。

(ステップA1)。次に、クラスタリング部11は、各サンプルにおける説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なって各ラベルに対する重みを設定する(ステップA2)。

[0044] ステップA2における「クラスタリング」とは、各欠損のパターンに対して、有限個のラベルと重みとを割り当てることである。

[0045] 具体的には、ステップA2では、クラスタリング部11は、説明変数の欠損のパターンそれぞれにおける欠損していない説明変数の集合の種類毎に、各種類に該当するサンプル群(各種類を構成する説明変数の集合が欠損していないサンプル群)を特定する。そして、クラスタリング部11は、説明変数の欠損のパターン毎に、各パターンに該当しないサンプルと、特定したサンプル群それぞれとの共通するサンプルを特定する。また、クラスタリング部11は、特定したサンプルの数が一定以上となる候補を選択し、選択した各候補にラベルを付与する。また、クラスタリング部11は、各ラベルに、欠損のパターン毎に、特定した共通するサンプルが多いほど値が大きくなるように、重みを設定する。

[0046] ステップA2で付与されたラベル毎に、後述する「子モデル」の学習が行なわれる。また、本実施の形態では、ステップA2のクラスタリングによって付与された各ラベルを、以下においては「L」と表記する。

[0047] 子モデル作成部12は、重みが設定されたラベルLが付与されている候補を特定し、特定した候補に含まれるサンプル群を、部分訓練データとして抽

出する（ステップA3）。また、子モデル作成部12は、ラベルLに設定された重みを、部分訓練データを構成する各サンプルの重みとしても紐づける。

[0048] 次に、子モデル作成部12は、重みが設定されたラベル毎に、そのラベルが付与されている部分訓練データにおいて、説明変数の中から予測変数を決定する（ステップA4）。

[0049] 次に、欠損変数補完部14は、ステップA3で抽出された部分訓練データにおいて、それに付与されているラベルの、ステップA4で決定された予測変数が、いずれかのサンプルで欠損しているかどうかを確認する。そして、欠損変数補完部14は、欠損している場合は、そのサンプルにおける欠損している予測変数を補完し、部分訓練データを更新する（ステップA5）。

[0050] 次に、混合モデル作成部13は、重みが設定されたラベル毎に、そのラベルが付与されている部分訓練データを用いて、予測変数から目的変数を導く中間予測関数（子モデル）を学習する（ステップA6）。

[0051] 次に、混合モデル作成部13は、全てのラベルで子モデルを学習すると、欠損のパターン毎に、そのパターンについて設定された重みと、この重みが設定されているラベルについての子モデルとを用いて、最終予測関数を学習する（ステップA7）。

[0052] 以上のように本実施の形態では、説明変数の欠損のパターンに応じて、異なる最終予測関数が作成される。また、欠損パターンのうち、共通の変数を予測変数とするようなパターンはまとめられ、共通のサンプルで学習が実行される。言い換えると、本実施の形態では、最終予測関数の作成においては、欠損パターン毎に、欠損していない説明変数が共通しているサンプルがまとめられて、学習が行なわれている。従って、本実施の形態によれば、説明変数の欠損のパターンが多い場合でも、高い精度で予測を実行することができる。

[0053] [プログラム]

本実施の形態におけるプログラムは、コンピュータに、図3に示すステッ

プA1～A7を実行させるプログラムであれば良い。このプログラムをコンピュータにインストールし、実行することによって、本実施の形態における予測関数作成装置10と予測関数作成方法とを実現することができる。この場合、コンピュータのCPU (Central Processing Unit) は、クラスタリング部11、子モデル作成部12、混合モデル作成部13、及び欠損変数補完部14として機能し、処理を行なう。

[0054] ここで、本実施の形態におけるプログラムを実行することによって、予測関数作成装置10を実現するコンピュータについて図4を用いて説明する。図4は、本発明の実施の形態における予測関数作成装置を実現するコンピュータの一例を示すブロック図である。

[0055] 図4に示すように、コンピュータ100は、CPU111と、メインメモリ112と、記憶装置113と、入力インターフェイス114と、表示コントローラ115と、データリーダ／ライタ116と、通信インターフェイス117とを備える。これらの各部は、バス121を介して、互いにデータ通信可能に接続される。

[0056] CPU111は、記憶装置113に格納された、本実施の形態におけるプログラム（コード）をメインメモリ112に展開し、これらを所定順序で実行することにより、各種の演算を実施する。メインメモリ112は、典型的には、DRAM (Dynamic Random Access Memory) 等の揮発性の記憶装置である。また、本実施の形態におけるプログラムは、コンピュータ読み取り可能な記録媒体120に格納された状態で提供される。なお、本実施の形態におけるプログラムは、通信インターフェイス117を介して接続されたインターネット上で流通するものであっても良い。

[0057] また、記憶装置113の具体例としては、ハードディスクドライブの他、フラッシュメモリ等の半導体記憶装置が挙げられる。入力インターフェイス114は、CPU111と、キーボード及びマウスといった入力機器118との間のデータ伝送を仲介する。表示コントローラ115は、ディスプレイ装置119と接続され、ディスプレイ装置119での表示を制御する。

[0058] データリーダー／ライター 116 は、CPU 111 と記録媒体 120 との間のデータ伝送を仲介し、記録媒体 120 からのプログラムの読み出し、及びコンピュータ 100 における処理結果の記録媒体 120 への書き込みを実行する。通信インターフェイス 117 は、CPU 111 と、他のコンピュータとの間のデータ伝送を仲介する。

[0059] また、記録媒体 120 の具体例としては、CF (Compact Flash (登録商標)) 及びSD (Secure Digital) 等の汎用的な半導体記憶デバイス、フレキシブルディスク (Flexible Disk) 等の磁気記憶媒体、又はCD-ROM (Compact Disk Read Only Memory) などの光学記憶媒体が挙げられる。

実施例

[0060] 上述した実施の形態に対応する実施例について図5～図10を用いて説明する。また、以下の説明では、図3に示した各ステップに沿って説明する。

[0061] 図5は、実施例で用いられる訓練データの一例を示す図である。図5に示すように、本実施例では、訓練データを構成する各サンプルには番号が付与されている。また、各サンプルは、説明変数として、X1、X2、X3の3つを有し、目的変数としてYを有している。また、説明変数における欠損は、記号「NA」で表わされている。

[0062] また、図5において、「*」が示された箇所には、何らかの数値が入っているものとする。図5の例では、サンプル番号1～5で欠損なし、サンプル番号6～10でX1のみが欠損、サンプル番号11～20でX1とX2が欠損、サンプル番号21～40でX3が欠損している。このような訓練データが入力された場合の具体的な処理は、以下の通りである。

[0063] [ステップA1]

まず、クラスタリング部11は、訓練データの各サンプルに対して、想定され得る説明変数の欠損のパターン（以下「欠損変数パターン」と表記する。）を特定する。サンプルの説明変数が、X1、X2、X3であるので、クラスタリング部11は、欠損のパターンとしては、空集合（欠損なし）、X1が欠損、X2が欠損、X3が欠損、X1とX2が欠損、X2とX3が欠損

、X 1 と X 3 が欠損、X 1 ～ X 3 全てが欠損、の 8 通りを特定する。

[0064] なお、図 5 の例では、サンプル番号 1 ～ 40 について、クラスタリング部 11 は、サンプル番号 1 ～ 5 では欠損なし、サンプル番号 6 ～ 10 で X 1 が欠損、サンプル番号 11 ～ 20 では X 1 及び X 2 が欠損、サンプル番号 21 ～ 40 で X 3 が欠損している、と判断する。

[0065] [ステップ A 2]

次に、クラスタリング部 11 は、各サンプルにおける説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルに重みを設定する。

[0066] 具体的には、以下のように、クラスタリング部 11 は、サンプルに対してラベルと重みとを割り当てて、クラスタリングを行なう。まず、クラスタリング部 11 は、説明変数が 3 つとも欠損していない場合、2 つ欠損していない（1 つだけ欠損する）場合、1 つだけ欠損していない（2 つ欠損する）場合、3 つとも欠損している場合等を特定する。本実施例では、説明変数は 3 つであるので、想定される場合は、実際には 8 通りである。そして、想定される各場合における欠損していない説明変数の集合（以下「非欠損変数集合」と表記する。）を A として、非欠損変数集合 A の種類毎に、クラスタリング部 11 は、次の計算を行なう。本実施例では、計算は、[A が欠損していないサンプル数] に [A の要素数] を乗算することによって行なわれる（＝ [A が欠損していないサンプル数] * [A の要素数] ）。

[0067] 図 6 は、実施例において想定される非欠損変数集合の種類毎の計算結果の一例を示す図である。そして、クラスタリング部 11 は、図 6 に示す乗算値が得られると、乗算値が閾値以上となる A を全て候補として、各候補を構成するサンプル群に、候補毎に割り当てられた番号をラベルとして付与する。

[0068] 図 6 の例では、閾値は 20 に設定されているとする。この場合、クラスタリング部 11 は、閾値以上となる非欠損変数集合それぞれ、即ち、X 1、X 2、X 3、(X 1, X 2)、及び (X 2, X 3) を候補とし、これらにラベルを表す番号を付与する（図 9 参照）。また、候補は、以下においては、候

補（X 1）、候補（X 2）、候補（X 3）、候補（X 1、X 2）、候補（X 2、X 3）のように示す。更に、候補（X 1）、候補（X 2）、候補（X 3）、候補（X 1、X 2）、候補（X 2、X 3）には、後述の図 8 に示すように、順に、ラベルとして、番号 5、1、2、3、4 が付与されている。

[0069] 次に、クラスタリング部 11 は、想定され得る欠損変数パターン毎に、各候補との距離 $d(I, L)$ を計算する。具体的には、クラスタリング部 11 は、両者間で共通して欠損していない説明変数を特定し、特定した変数が欠損していないサンプルの数が多い程、距離 d の値が大きくなるように計算する。例えば、欠損変数パターン X 1 と候補（X 2）とであれば、共通して欠損していない説明変数は X 2 であるので、説明変数 X 2 が欠損していないサンプルの数に基づいて距離 d が計算される。

[0070] 例えば、欠損変数パターンにおける、欠損していない説明変数のサンプル集合を I_1 、ラベル L の候補の欠損していない変数のサンプル集合を I_2 として、クラスタリング部 11 は、距離 d を下記数 1 を用いて計算することができる。

[0071] [数 1]

$$d(I, L) := |I_1 \setminus I_2| + \alpha |I_2 \setminus I_1|$$

[0072] また、上記数 1 において、 α は任意の整数を示し、 I は任意の欠損変数パターンを示し、 L は任意のラベル番号を示している。また、 $|\cdot|$ は、集合「 $\cdot$ 」の要素数を表し、「 $A \setminus B$ 」は、A の要素のうち、B の要素でないものの集合を表すとする。また、上記数 1 において、 $\alpha = 2$ とした場合の計算結果は、図 7 に示す通りとなる。図 7 は、実施例における欠損変数パターンと候補との距離についての計算結果の一例を示す図である。

[0073] 次に、クラスタリング部 11 は、欠損変数パターン I それぞれについて、各候補との距離 d が閾値未満である場合は、距離 d が閾値未満となった候補との類似度を、下記の数 2 を用いて計算する。

[0074]

[数2]

$$\text{類似度} = e^{-\lambda d(I, L)}$$

[0075] 次に、クラスタリング部 11 は、欠損変数パターン I それぞれにおいて、距離 d が閾値未満となった候補について計算された類似度の和が 1 となるように、下記の数 3 を用いて各類似度を正規化する。そして、クラスタリング部 11 は、得られた値を、欠損変数パターン I に対するラベル L の重み $w(I, L)$ として出力する。

[0076] [数3]

$$w(I, L) := \frac{e^{-\lambda d(I, L)}}{\sum_L e^{-\lambda d(I, L)}}$$

[0077] 上記数 3 の算出結果の具体例は、図 8 に示す通りである。図 8 は、実施例において各ラベルに設定された重みの一例を示す図である。なお、図 8 の例は、距離 d の閾値を 2 として算出されている。

[0078] また、本実施例では、上述の想定し得る非欠損変数集合（候補）を用いた方法とは別の方法によってクラスタリングを実行することもできる。例えば、説明変数の次元の $\{0, 1\}$ 値ベクトルを、各サンプルで、その成分の説明変数が欠損していれば 1、欠損していなければ 0 として定義し、既存の k -means 法、デンドログラム法などを用いてクラスタリングを実行することもできる。

[0079] [ステップ A 3]

次に、子モデル作成部 12 は、ラベルに設定された重みが大きい欠損変数パターンとなっているサンプルを、訓練データから抽出し、ラベルへの重みが大きいほど、大きいサンプル重みを紐づけて、部分訓練データとして出力する。

[0080] 例えば、子モデル作成部 12 は、欠損変数パターン毎に、ラベル L に正の重みが付いたサンプル集合を特定し、特定した各サンプルに、ラベル L に設

定された重みを、サンプル重みとして紐づける。そして、子モデル作成部 12 は、サンプル重みが紐付けられたサンプルの集合を部分訓練データとする。

[0081] [ステップ A 4]

次に、子モデル作成部 12 は、ラベル毎に、説明変数の中から、予測に用いる説明変数（予測変数）を選択する。その際、子モデル作成部 12 は、部分訓練データ中において、予測変数として選択される説明変数の数が可能な限り多くなり、且つ、予測変数として選択される説明変数が欠損していないサンプルの数が、他の説明変数が欠損していないサンプルの数よりも多くなるようにする。

[0082] 具体的には、本実施例では、子モデル作成部 12 は、クラスタリング時に候補と決定した非欠損変数集合を構成する説明変数（以下「候補変数」と表記する。）を、各ラベルの予測変数とする。よって、子モデル作成部 12 は、ラベル 1 について X 2、ラベル 2 について X 3、ラベル 3 について X 1 及び X 2、ラベル 4 について X 2 及び X 3、ラベル 5 について X 1 と、予測変数を選択する。また、子モデル作成部 12 は、この時点で、訓練データ中で候補変数がひとつも欠損していないサンプルを特定し、部分訓練データに追加する。

[0083] また、本実施例では、ラベルが付与された部分訓練データにおいて欠損していない説明変数から、予測変数が選択されているが、これに限定される趣旨ではない。例えば、部分訓練データにおいて欠損している説明変数から、予測変数が選択され、その後に補完されても良い。

[0084] 図 9 は、実施例における部分訓練データの一例を示す図である。図 9 に示すように、本実施例では、ラベル毎に部分訓練データが特定されている。また、図 9 に示されたサンプル番号は、図 3 に示した各サンプルの番号と一致している。

[0085] [ステップ A 5]

次に、欠損変数補完部 14 は、部分訓練データにおいて、予測変数が欠損

しているサンプルがあった場合は、そのサンプルに対して補完を行う。例えば、欠損変数補完部 14 は、予測変数が欠損しているサンプルにおける、欠損している予測変数の値として、他の予測変数の平均値を代入することによって、補完を行なうことができる。また、欠損変数補完部 14 は、他の変数からの回帰を行う単一代入法、多重代入法などを用いて補完を行なうこともできる。なお、本実施例では、部分訓練データは、予測変数が欠損しているサンプルが発生しないように特定されている。

[0086] [ステップ A 6]

次に、混合モデル作成部 13 は、部分訓練データを用い、例えば、線型回帰、Support Vector Regression等の方法を利用することによって、ラベル L 毎に、予測変数から目的変数への予測関数を作成する。また、この場合、混合モデル作成部 13 は、サンプルの重みに応じて、損失関数または尤度関数を重みづけして計算する。

[0087] 具体的には、混合モデル作成部 13 は、まず、サンプル i の重みを w_i 、モデルの損失関数を $l(x_i, y_i)$ として、下記数 4 を用いてパラメータ β を推定し、ラベル L 毎に、 β を含む予測関数 f_L を作成する。

[0088] [数 4]

$$\hat{\beta} = \arg \min_{\beta} \sum_i w_i l(x_i, y_i)$$

[0089] また、混合モデル作成部 13 は、例えば、二乗誤差を最小化するように線形回帰を行う場合は、下記数 5 を用いてパラメータ β を推定し、ラベル L 毎に、 β を含む予測関数 f_L を作成する。

[0090] [数 5]

$$\hat{\beta} = \arg \min_{\beta} \sum_i w_i |y_i - (\beta^T x_i + \beta_0)|^2$$

[0091] [ステップ A 7]

最後に、混合モデル作成部 13 は、欠損変数パターン l 毎に、ラベル L の

重み $w(I, L)$ が大きいほど、該当するラベル L において作成された予測関数が重視されるように、最終予測関数 $f_I(x)$ を生成する。最終予測関数 $f_I(x)$ の具体例は、下記数 6 の通りとなる。

[0092] [数6]

$$f_I(x) := \sum_L w(I, L) f_L(x)$$

[0093] 図 10 は、実施例で得られた最終予測関数の具体例を示す図である。なお、図 10 に示した最終予測関数は、図 3 の訓練データの例において示されていない欠損変数パターンについても $w(I, L)$ を計算することによって得られている。

[0094] 以上のように本実施例では、図 10 に示すように、欠損変数パターンに応じて、異なる最終予測関数が作成される。従って、本実施例から、説明変数の欠損パターンが多い場合でも、高い精度で予測を実行することができることが分かる。

[0095] 上述した実施の形態及び実施例の一部又は全部は、以下に記載する（付記 1）～（付記 15）によって表現することができるが、以下の記載に限定されるものではない。

[0096] （付記 1）

説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するための装置であって、

各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合せて重みを設定する、クラスタリング部と、

設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を

構成する説明変数を決定する、子モデル作成部と、

ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、混合モデル作成部と、を備えていることを特徴とする予測関数作成装置。

[0097] (付記2)

前記クラスタリング部が、
前記欠損のパターンそれぞれにおける欠損していない説明変数の集合の種類毎に、該当するサンプル群を特定し、
更に、前記欠損のパターン毎に、当該パターンに該当しないサンプルと、特定した前記サンプル群それぞれとの共通するサンプルを特定し、
そして、特定した前記共通するサンプルの数が一定以上となるサンプル群それぞれに対して、ラベルを付与し、
その後、各ラベルに、前記欠損のパターン毎に、特定した前記共通するサンプルが多いほど値が大きくなるように、前記重みを設定する、
付記1に記載の予測関数作成装置。

[0098] (付記3)

前記子モデル作成部が、
前記重みが設定されたラベルが付与されているサンプル群を、前記部分訓練データとし、
前記重みが設定されたラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を選択する、
付記2に記載の予測関数作成装置。

[0099] (付記4)

前記混合モデル作成部が、
前記重みが設定されたラベル毎に、当該ラベルが付与されている前記部分訓練データを用いて、決定された前記説明変数から前記目的変数を導く第2の予測関数を作成し、
そして、前記欠損のパターン毎に、当該パターンについて設定された前記重

みと、当該重みが設定されているラベルについての前記第 2 の予測関数とを用いて、前記予測関数を作成する、
付記 3 に記載の予測関数作成装置。

[0100] (付記 5)

前記部分訓練データにおいて、それに付与されているラベルについて決定された、前記予測関数を構成する説明変数が欠損しているサンプルが存在する場合に、当該サンプルにおける欠損している説明変数を補完する、欠損変数補完部を更に備え、

前記混合モデル作成部が、補完された前記部分訓練データを用いて、前記第 2 の予測関数を作成する、
付記 4 に記載の予測関数作成装置。

[0101] (付記 6)

説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するための方法であって、

(a) 各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合せて重みを設定する、ステップと、

(b) 設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、ステップと、

(c) ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、ステップと、
を備えていることを特徴とする予測関数作成方法。

[0102] (付記 7)

前記 (a) のステップにおいて、
前記欠損のパターンそれぞれにおける欠損していない説明変数の集合の種類

毎に、該当するサンプル群を特定し、
更に、前記欠損のパターン毎に、当該パターンに該当しないサンプルと、特定した前記サンプル群それぞれとの共通するサンプルを特定し、
そして、特定した前記共通するサンプルの数が一定以上となるサンプル群それぞれに対して、ラベルを付与し、
その後、各ラベルに、前記欠損のパターン毎に、特定した前記共通するサンプルが多いほど値が大きくなるように、前記重みを設定する、
付記 6 に記載の予測関数作成方法。

[0103] (付記 8)

前記 (b) のステップにおいて、
前記重みが設定されたラベルが付与されているサンプル群を、前記部分訓練データとし、
前記重みが設定されたラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を選択する、
付記 7 に記載の予測関数作成方法。

[0104] (付記 9)

前記 (c) のステップにおいて、
前記重みが設定されたラベル毎に、当該ラベルが付与されている前記部分訓練データを用いて、決定された前記説明変数から前記目的変数を導く第 2 の予測関数を作成し、
そして、前記欠損のパターン毎に、当該パターンについて設定された前記重みと、当該重みが設定されているラベルについての前記第 2 の予測関数とを用いて、前記予測関数を作成する、
付記 8 に記載の予測関数作成方法。

[0105] (付記 10)

(d) 前記部分訓練データにおいて、それに付与されているラベルについて決定された、前記予測関数を構成する説明変数が欠損しているサンプルが存在する場合に、当該サンプルにおける欠損している説明変数を補完する、ス

テップを更に有し、

前記（c）のステップにおいて、補完された前記部分訓練データを用いて、前記第2の予測関数を作成する、
付記9に記載の予測関数作成方法。

[0106] （付記11）

コンピュータによって、説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、
前記コンピュータに、

（a）各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合せて重みを設定する、ステップと、

（b）設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、ステップと、

（c）ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、ステップと、
を実行させる命令を含む、プログラムを記録しているコンピュータ読み取り可能な記録媒体。

[0107] （付記12）

前記（a）のステップにおいて、
前記欠損のパターンそれぞれにおける欠損していない説明変数の集合の種類毎に、該当するサンプル群を特定し、
更に、前記欠損のパターン毎に、当該パターンに該当しないサンプルと、特定した前記サンプル群それぞれとの共通するサンプルを特定し、
そして、特定した前記共通するサンプルの数が一定以上となるサンプル群そ

れぞれに対して、ラベルを付与し、
その後、各ラベルに、前記欠損のパターン毎に、特定した前記共通するサンプルが多いほど値が大きくなるように、前記重みを設定する、
付記 1 1 に記載のコンピュータ読み取り可能な記録媒体。

[0108] (付記 1 3)

前記 (b) のステップにおいて、
前記重みが設定されたラベルが付与されているサンプル群を、前記部分訓練データとし、
前記重みが設定されたラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を選択する、
付記 1 2 に記載のコンピュータ読み取り可能な記録媒体。

[0109] (付記 1 4)

前記 (c) のステップにおいて、
前記重みが設定されたラベル毎に、当該ラベルが付与されている前記部分訓練データを用いて、決定された前記説明変数から前記目的変数を導く第 2 の予測関数を作成し、
そして、前記欠損のパターン毎に、当該パターンについて設定された前記重みと、当該重みが設定されているラベルについての前記第 2 の予測関数とを用いて、前記予測関数を作成する、
付記 1 3 に記載のコンピュータ読み取り可能な記録媒体。

[0110] (付記 1 5)

(d) 前記部分訓練データにおいて、それに付与されているラベルについて決定された、前記予測関数を構成する説明変数が欠損しているサンプルが存在する場合に、当該サンプルにおける欠損している説明変数を補完する、ステップを、更に前記コンピュータに実行させ、

前記 (c) のステップにおいて、補完された前記部分訓練データを用いて、前記第 2 の予測関数を作成する、
付記 1 4 に記載のコンピュータ読み取り可能な記録媒体。

[0111] 以上、実施の形態を参照して本願発明を説明したが、本願発明は上記実施の形態に限定されるものではない。本願発明の構成や詳細には、本願発明の範囲内で当業者が理解し得る様々な変更をすることができる。

[0112] この出願は、2013年6月12日に提出された日本出願特願2013-123482を基礎とする優先権を主張し、その開示の全てをここに取り込む。

産業上の利用可能性

[0113] 以上のように、本発明によれば、説明変数の欠損の仕方のバリエーション、即ち、説明変数の欠損パターンが多い場合でも、高い精度で予測を実行することができる。本発明は、業務ログの分析、購買履歴からの将来購買予測といった、予測分析の分野に有用である。

符号の説明

- [0114]
- 11 クラスタリング部
 - 12 子モデル作成部
 - 13 混合モデル作成部
 - 14 欠損変数補完部
 - 100 コンピュータ
 - 110 端末
 - 111 CPU
 - 112 メインメモリ
 - 113 記憶装置
 - 114 入力インターフェイス
 - 115 表示コントローラ
 - 116 データリーダー／ライター
 - 117 通信インターフェイス
 - 118 入力機器
 - 119 ディスプレイ装置
 - 120 記録媒体

1 2 1 バス

請求の範囲

- [請求項1] 説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するための装置であって、
- 各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合せて重みを設定する、クラスタリング部と、
- 設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、子モデル作成部と、
- ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、混合モデル作成部と、
- を備えていることを特徴とする予測関数作成装置。
- [請求項2] 前記クラスタリング部が、
- 前記欠損のパターンそれぞれにおける欠損していない説明変数の集合の種類毎に、該当するサンプル群を特定し、
- 更に、前記欠損のパターン毎に、当該パターンに該当しないサンプルと、特定した前記サンプル群それぞれとの共通するサンプルを特定し、
- そして、特定した前記共通するサンプルの数が一定以上となるサンプル群それぞれに対して、ラベルを付与し、
- その後、各ラベルに、前記欠損のパターン毎に、特定した前記共通するサンプルが多いほど値が大きくなるように、前記重みを設定する、
- 請求項1に記載の予測関数作成装置。
- [請求項3] 前記子モデル作成部が、

前記重みが設定されたラベルが付与されているサンプル群を、前記部分訓練データとし、

前記重みが設定されたラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を選択する、
請求項 2 に記載の予測関数作成装置。

[請求項4]

前記混合モデル作成部が、
前記重みが設定されたラベル毎に、当該ラベルが付与されている前記部分訓練データを用いて、決定された前記説明変数から前記目的変数を導く第 2 の予測関数を作成し、
そして、前記欠損のパターン毎に、当該パターンについて設定された前記重みと、当該重みが設定されているラベルについての前記第 2 の予測関数とを用いて、前記予測関数を作成する、
請求項 3 に記載の予測関数作成装置。

[請求項5]

前記部分訓練データにおいて、それに付与されているラベルについて決定された、前記予測関数を構成する説明変数が欠損しているサンプルが存在する場合に、当該サンプルにおける欠損している説明変数を補完する、欠損変数補完部を更に備え、

前記混合モデル作成部が、補完された前記部分訓練データを用いて、前記第 2 の予測関数を作成する、
請求項 4 に記載の予測関数作成装置。

[請求項6]

説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するための方法であって、
(a) 各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合わせて重みを設定する、ステップと、
(b) 設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記

説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、ステップと、
(c) ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、ステップと、
を備えていることを特徴とする予測関数作成方法。

[請求項7]

コンピュータによって、説明変数と目的変数とを含むサンプルの集合を訓練データとして用いて、前記目的変数を導く予測関数を作成するためのプログラムを記録したコンピュータ読み取り可能な記録媒体であって、

前記コンピュータに、

(a) 各サンプルにおける前記説明変数の欠損のパターンに応じて、各サンプルに対して、クラスタリングを行なってラベルを付与し、各ラベルには、当該ラベルが付与されているサンプルにおける前記説明変数の欠損のパターンに合わせて重みを設定する、ステップと、

(b) 設定された前記重みに基づいて、訓練データの一部を部分訓練データとし、前記部分訓練データとなった前記サンプルにおける前記説明変数の欠損のパターンに基づいて、ラベル毎に、前記説明変数の中から、前記予測関数を構成する説明変数を決定する、ステップと、

(c) ラベル毎に決定された前記説明変数及び前記部分訓練データを用いて、前記欠損のパターン毎に、前記予測関数を作成する、ステップと、

を実行させる命令を含む、プログラムを記録しているコンピュータ読み取り可能な記録媒体。