(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2010-3106
(P2010-3106A)
(43)公開日　平成22年1月7日(2010.1.7)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	210D	5B075
G06N 5/04	(2006.01)	G06N	5/04	580A	
		G06N	5/04	550J	

審査請求 未請求　請求項の数 11　ＯＬ　（全 22 頁）

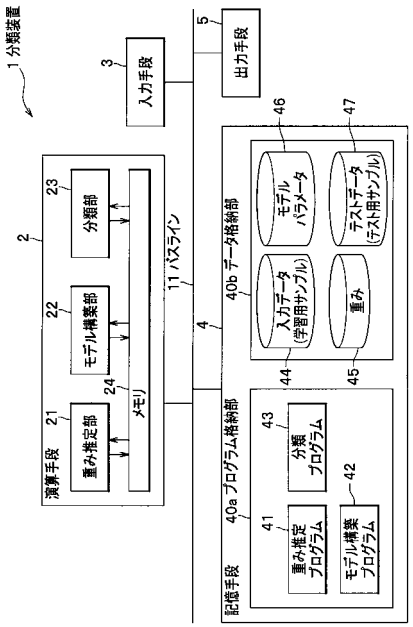
(21)出願番号　特願2008-161237 (P2008-161237)
(22)出願日　平成20年6月20日 (2008.6.20)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(71)出願人　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(74)代理人　100064414
弁理士　磯野　道造
(74)代理人　100127720
弁理士　大石　恵
(72)発明者　岩田　具治
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】分類モデル生成装置、分類装置、分類モデル生成方法、分類方法、分類モデル生成プログラム、分類プログラムおよび記録媒体

(57)【要約】
【課題】ターゲット分類体系のデータだけでなく、補助分類体系のデータも有効に利用することで、ターゲット分類体系に関する高精度な分類モデルを生成する。
【解決手段】分類装置１は、ターゲット分類体系における既分類データだけでなく、補助分類体系における既分類データも用い、誤差関数と重みとの積の総和である期待誤差を最小化させるように重みを推定し、その推定された重みと２種類の既分類データとを用いて分類モデルを生成することで、ターゲット分類体系のデータだけでなく、補助分類体系のデータも有効に利用し、ターゲット分類体系に関する高精度な分類モデルを生成することができる。
【選択図】図１



【特許請求の範囲】
【請求項１】
分類対象データを分類する分類体系であるターゲット分類体系においてすでに分類されている１つ以上の既分類データと、前記ターゲット分類体系とは異なる分類体系である補助分類体系においてすでに分類されている１つ以上の既分類データと、を用いて学習を行うことで、前記分類対象データを前記ターゲット分類体系における複数のクラスのいずれかに分類するための分類モデルを生成する分類モデル生成装置であって、
情報を記憶する記憶手段と、
前記した２種類の既分類データにおける個別の各既分類データを前記ターゲット分類体系のいずれかのクラスに分類したと予測したときの前記分類モデルの誤差関数と、当該予測をしたときの前記した２種類の既分類データにおける個別の各既分類データの前記分類モデルへの影響度を示す各重みと、を用いて、前記した２種類の既分類データにおける個別の既分類データごとの前記誤差関数の値と前記重みとの積の総和である期待誤差を最小化させるように、前記重みを推定して、当該重みを前記記憶手段に格納する重み推定部と、
前記記憶手段に格納された重みと、前記した２種類の既分類データと、を用いて、前記分類モデルを生成するモデル構築部と、
を備えることを特徴とする分類モデル生成装置。
【請求項２】
前記重み推定部は、
前記ターゲット分類体系と前記補助分類体系とを統合した場合の確率分布モデルを、前記ターゲット分類体系の確率分布モデルに近似させるための、前記ターゲット分類体系と前記補助分類体系とのクラスごとの前記分類モデルへの影響度の比率を示す混合比を用いて、前記した２種類の既分類データに関する事後確率を推定して、当該事後確率を前記記憶手段に格納する事後確率推定部と、
前記記憶手段に格納された事後確率を用いて、前記ターゲット分類体系の既分類データに対する尤度を最大化するように、前記混合比を推定し、前記尤度が最大化されたときの前記混合比から前記重みを推定して、当該重みを前記記憶手段に格納する混合比推定部と、
を備えることを特徴とする請求項１に記載の分類モデル生成装置。
【請求項３】
前記モデル構築部は、
前記記憶手段に格納された重みと、前記した２種類の既分類データと、を用いて、前記分類モデルにおいて前記分類対象データを前記ターゲット分類体系に分類するためのモデルパラメータを推定して、当該モデルパラメータを前記記憶手段に格納するモデルパラメータ推定部
を備えることを特徴とする請求項１に記載の分類モデル生成装置。
【請求項４】
請求項３に記載の分類モデル生成装置の前記記憶手段に格納されたモデルパラメータを用いて、前記分類対象データを前記ターゲット分類体系における複数のクラスのいずれかに分類する分類部
を備えることを特徴とする分類装置。
【請求項５】
分類対象データを分類する分類体系であるターゲット分類体系においてすでに分類されている１つ以上の既分類データと、前記ターゲット分類体系とは異なる分類体系である補助分類体系においてすでに分類されている１つ以上の既分類データと、を用いて学習を行うことで、前記分類対象データを前記ターゲット分類体系における複数のクラスのいずれかに分類するための分類モデルを生成する分類モデル生成装置による分類モデル生成方法であって、
前記分類モデル生成装置は、情報を記憶する記憶手段と、重み推定部と、モデル構築部

と、を備えており、
前記重み推定部は、前記した２種類の既分類データにおける個別の各既分類データを前記ターゲット分類体系のいずれかのクラスに分類したと予測したときの前記分類モデルの誤差関数と、当該予測をしたときの前記した２種類の既分類データにおける個別の各既分類データの前記分類モデルへの影響度を示す各重みと、を用いて、前記した２種類の既分類データにおける個別の既分類データごとの前記誤差関数の値と前記重みとの積の総和である期待誤差を最小化させるように、前記重みを推定して、当該重みを前記記憶手段に格納する重み推定ステップを実行し、
前記モデル構築部は、前記記憶手段に格納された重みと、前記した２種類の既分類データと、を用いて、前記分類モデルを生成するモデル構築ステップを実行する
ことを特徴とする分類モデル生成方法。
【請求項６】
前記重み推定部は、事後確率推定部と、混合比推定部と、を備えており、
前記重み推定ステップにおいて、
前記事後確率推定部は、前記ターゲット分類体系と前記補助分類体系とを統合した場合の確率分布モデルを、前記ターゲット分類体系の確率分布モデルに近似させるための、前記ターゲット分類体系と前記補助分類体系とのクラスごとの前記分類モデルへの影響度の比率を示す混合比を用いて、前記した２種類の既分類データに関する事後確率を推定して、当該事後確率を前記記憶手段に格納し、
前記混合比推定部は、前記記憶手段に格納された事後確率を用いて、前記ターゲット分類体系の既分類データに対する尤度を最大化するように、前記混合比を推定し、前記尤度が最大化されたときの前記混合比から前記重みを推定して、当該重みを前記記憶手段に格納する
ことを特徴とする請求項５に記載の分類モデル生成方法。
【請求項７】
前記モデル構築部は、モデルパラメータ推定部を備えており、
前記モデル構築ステップにおいて、
前記モデルパラメータ推定部は、前記記憶手段に格納された重みと、前記した２種類の既分類データと、を用いて、前記分類モデルにおいて前記分類対象データを前記ターゲット分類体系に分類するためのモデルパラメータを推定して、当該モデルパラメータを前記記憶手段に格納する
ことを特徴とする請求項５に記載の分類モデル生成方法。
【請求項８】
請求項７に記載の分類モデル生成方法によって前記記憶手段に格納されたモデルパラメータを用いて、
前記分類対象データを分類する分類装置における分類部は、
前記分類対象データを前記ターゲット分類体系における複数のクラスのいずれかに分類するステップを実行する
ことを特徴とする分類方法。
【請求項９】
コンピュータを請求項１から請求項３のいずれか一項に記載の分類モデル生成装置の各部として機能させるための分類モデル生成プログラム。
【請求項１０】
コンピュータを請求項４に記載の分類装置の分類部として機能させるための分類プログラム。
【請求項１１】
請求項９に記載の分類モデル生成プログラム、または、請求項１０に記載の分類プログラムが記録されたことを特徴とするコンピュータに読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】

【０００１】
本発明は、分類対象データを分類する分類体系（以下、「ターゲット分類体系」という。）のデータだけでなく、別の分類体系（以下、「補助分類体系」という。）のデータも用いて、分類モデルを学習し、また、その学習した分類モデルを用いて分類対象データをターゲット分類体系において分類する技術に関する。
【背景技術】
【０００２】
学習データ（学習用のデータ）の数が少ない場合、一般に、分類モデルの性能は低くなる。そこで、補助分類体系におけるクラスのラベル（以下、「クラスラベル」または単に「ラベル」という。）が付与されたデータを用いることにより、分類モデルの性能を向上させることができれば好ましい。その場合、例えば、あるＷｅｂページを、あるターゲット分類体系のクラス（以下、「ターゲットクラス」ともいう。）に分類したいとする。そして、ディレクトリ型検索エンジンやソーシャルブックマークサイトにおける多数のユーザによって、ターゲット分類体系とは異なる補助分類体系に、多くのＷｅｂページがすでに分類されており、そのような情報を活用できれば望ましい。
【０００３】
また、例えば、オンラインショッピングなどの商品について購買順序を考慮した予測（分類）に関する技術が知られている（非特許文献１参照）。
【非特許文献１】岩田具治、山田武士、上田修功、“購買順序を考慮した協調フィルタリング”、人工知能と知識処理研究会、AI2007-3,13-18,2007
【発明の開示】
【発明が解決しようとする課題】
【０００４】
しかし、補助分類体系とターゲット分類体系とでは、一般にクラスラベルが異なり、また、同じラベルがあったとしても意味が異なる可能性もある。そのため、従来の教師あり学習の技術（非特許文献１など）を用いて、補助分類体系のクラス（以下、「補助クラス」ともいう。）のデータを利用することはできないという問題がある。
【０００５】
そこで、本発明は、前記問題に鑑みてなされたものであり、補助分類体系のデータを有効に利用することで、ターゲット分類体系に関する高精度な分類モデルを生成することを課題とする。また、その生成した分類モデルを用いて、分類対象データをターゲット分類体系において高精度に分類することを、他の課題とする。
【課題を解決するための手段】
【０００６】
前記課題を解決するために、本発明は、分類対象データを分類する分類体系であるターゲット分類体系においてすでに分類されている１つ以上の既分類データと、前記ターゲット分類体系とは異なる分類体系である補助分類体系においてすでに分類されている１つ以上の既分類データと、を用いて学習を行うことで、前記分類対象データを前記ターゲット分類体系における複数のクラスのいずれかに分類するための分類モデルを生成する分類モデル生成装置であって、情報を記憶する記憶手段と、前記した２種類の既分類データにおける個別の各既分類データを前記ターゲット分類体系のいずれかのクラスに分類したと予測したときの前記分類モデルの誤差関数と、当該予測をしたときの前記した２種類の既分類データにおける個別の各既分類データの前記分類モデルへの影響度を示す各重みと、を用いて、前記した２種類の既分類データにおける個別の既分類データごとの前記誤差関数の値と前記重みとの積の総和である期待誤差を最小化させるように、前記重みを推定して、当該重みを前記記憶手段に格納する重み推定部と、前記記憶手段に格納された重みと、前記した２種類の既分類データと、を用いて、前記分類モデルを生成するモデル構築部と、を備えることを特徴とする。
【０００７】
かかる発明によれば、ターゲット分類体系における既分類データだけでなく、補助分類

体系における既分類データも用い、誤差関数と重みとの積の総和である期待誤差を最小化させるように重みを推定し、その推定された重みと２種類の既分類データとを用いて分類モデルを生成することで、補助分類体系のデータも有効に利用し、ターゲット分類体系に関する高精度な分類モデルを生成することができる。

【０００８】

また、本発明は、前記重み推定部が、前記ターゲット分類体系と前記補助分類体系とを統合した場合の確率分布モデルを、前記ターゲット分類体系の確率分布モデルに近似させるための、前記ターゲット分類体系と前記補助分類体系とのクラスごとの前記分類モデルへの影響度の比率を示す混合比を用いて、前記した２種類の既分類データに関する事後確率を推定して、当該事後確率を前記記憶手段に格納する事後確率推定部と、前記記憶手段に格納された事後確率を用いて、前記ターゲット分類体系の既分類データに対する尤度を最大化するように、前記混合比を推定し、前記尤度が最大化されたときの前記混合比から前記重みを推定して、当該重みを前記記憶手段に格納する混合比推定部と、を備えることを特徴とする。

【０００９】

かかる発明によれば、重み推定部が、事後確率推定部と、混合比推定部とを備えているので、例えば、事後確率推定部が、ＥＭ（Expectation-Maximization）アルゴリズムにおけるＥ(Expectation)ステップを行い、かつ、混合比推定部がＭ(Maximization)ステップを行うことで、混合比についての大域的最適解を求め、求めた混合比から重みを決定（推定）することができる。

【００１０】

また、本発明は、前記モデル構築部が、前記記憶手段に格納された重みと、前記した２種類の既分類データと、を用いて、前記分類モデルにおいて前記分類対象データを前記ターゲット分類体系に分類するためのモデルパラメータを推定して、当該モデルパラメータを前記記憶手段に格納するモデルパラメータ推定部を備えることを特徴とする。

【００１１】

かかる発明によれば、モデルパラメータ推定部が、例えば、後記する式（１０）を用いてモデルパラメータを推定することができる。

【００１２】

また、本発明は、分類装置が、分類モデル生成装置の前記記憶手段に格納されたモデルパラメータを用いて、前記分類対象データを前記ターゲット分類体系における複数のクラスのいずれかに分類する分類部を備えることを特徴とする。

【００１３】

かかる発明によれば、分類部が、推定したモデルパラメータを用いて分類対象データをターゲット分類体系における複数のクラスのいずれかに分類する、つまり、高精度な分類モデルを用いることで高精度な分類を実現することができる。

【００１４】

また、本発明は、コンピュータを、分類モデル生成装置または分類装置の各部として機能させるためのプログラムである。これにより、このプログラムをインストールされたコンピュータは、このプログラムに基づいた各機能を実現することができる。

【００１５】

また、本発明は、前記プログラムが記録されたことを特徴とするコンピュータに読み取り可能な記録媒体である。これにより、この記録媒体を装着されたコンピュータは、この記録媒体に記録されたプログラムに基づいた各機能を実現することができる。

【発明の効果】

【００１６】

本発明によれば、補助分類体系のデータも有効に利用することで、ターゲット分類体系に関する高精度な分類モデルを生成することができる。また、その生成した分類モデルを用いて、分類対象データをターゲット分類体系において高精度に分類することができる。

【発明を実施するための最良の形態】

【００１７】
以下、本発明を実施するための最良の形態（以下、「実施形態」という。）について、詳細に説明する。図１は、本実施形態に係る分類装置の構成を示すブロック図である。図１に示すように、分類装置１は、演算手段２と、入力手段３と、記憶手段４と、出力手段５とを備えている。各手段２～５はバスライン１１に接続されている。なお、分類装置１は、分類モデル（以下、単に「モデル」ともいう。）を生成する分類モデル生成装置としての機能と、その生成した分類モデルによって分類対象データを分類する分類装置としての機能とを兼ね備えるものであるが、いずれか一方の機能のみを有するものとして実現されてもよい。

【００１８】
演算手段２は、例えば、ＣＰＵ（Central Processing Unit）およびＲＡＭ（Random Access Memory）から構成される主制御装置である。この演算手段２は、図１に示すように、重み推定部２１と、モデル構築部２２と、分類部２３と、メモリ２４とを含んで構成される。なお、各部２１～２３の説明は後記するが、従来手法と比較した場合の本実施形態における主な特徴は重み推定部２１であるので、重み推定部２１に関して特に詳細に説明する。また、モデル構築部２２と分類部２３に関しては、従来手法を大きく変更せずに適用できるので、詳細な説明を省略する。

【００１９】
入力手段３は、例えば、キーボード、マウス、ディスクドライブ装置等から構成される。この入力手段３は、各種データを入力し、記憶手段４に格納する（詳細は後記）。

【００２０】
記憶手段４は、例えば、一般的なハードディスク装置等から構成され、演算手段２で用いられる各種プログラムや各種データ等を記憶する。この記憶手段４は、プログラムとして、重み推定プログラム４１と、モデル構築プログラム４２と、分類プログラム４３とをプログラム格納部４０ａに記憶する。そして、演算手段２は、これらのプログラム４１～４３を記憶手段４から読み込んでメモリ２４に展開して実行することで、前記した重み推定部２１、モデル構築部２２、分類部２３の各機能を実現する。

【００２１】
また、記憶手段４は、入力データ４４と、重み４５と、モデルパラメータ４６と、テストデータ４７とをデータ格納部４０ｂに記憶する。ここで、入力データ４４は、入力手段３から入力されるデータであり、学習用サンプルである。重み４５は、演算手段２の重み推定部２１の演算処理によって推定された重みに関するデータである（詳細は後記）。モデルパラメータ４６は、演算手段２のモデル構築部２２の演算処理によって算出されたデータである（詳細は後記）。テストデータ４７は、テスト用サンプルである（詳細は後記）。なお、入力データ４４、重み４５、モデルパラメータ４６およびテストデータ４７に関しては、以下、符号を適宜省略する。

【００２２】
出力手段５は、例えば、グラフィックボード（出力インタフェース）およびそれに接続されたモニタである。このモニタは、例えば、液晶ディスプレイ等から構成され、演算処理結果（分類対象データの分類結果等）を表示する。

【００２３】
本実施形態では、ターゲット分類体系のデータ（既分類データ。以下、「ターゲットデータ」ともいう。）だけでなく、補助分類体系のデータ（既分類データ以下、「補助データ」ともいう。）も用いて、分類器（分類モデル）を学習する。ターゲットクラス集合をＺ、補助クラス集合をＡ、全クラス集合をＹ＝｛Ｚ，Ａ｝とする。

【００２４】
学習データとして、ターゲットデータである$D_z=\{x_n, y_n\}_{n=1}^{N_z}$（本明細書において、「$N_z{}_{n=1}$」は「ｎ」に「１」から「$N_z$」までを代入することを意味する。他の文字についても同様）と、
補助データである$D_a=\{x_n, y_n\}_{n=N_z+1}^{N}$とが与えられたとき、クラスが未

知のサンプルｘ（分類対象データ。後記するテストデータ４７）のクラスｙ∈Ｚを予測する分類モデルを学習する。

【００２５】

ここで、Ｗｅｂページデータの場合、サンプルは例えば単語出現頻度ベクトルx_n＝（x_{n1}，・・・，x_{nw}）で表される（x_{nw}は第ｎサンプルに単語ｗが出現した回数を表す）。

【００２６】

また、$y_n \in Z$（ｉｆ　１≦ｎ≦N_z）、$y_n \in A$（ｉｆ　N_z＋１≦ｎ≦Ｎ）であり、Ｗｅｂページの場合、y_nは第ｎサンプルが分類されているカテゴリを表す。なお、ｙは離散値である。また、ｘを離散変数として扱うが、連続変数の場合へも容易に拡張可能である。

【００２７】

本実施形態では、ターゲットデータに補助データ（補助分類体系のデータ）も含めた全データに関する重み付き経験誤差Ｅ（Ｍ）（式（１））を最小化することにより、モデルＭを学習する。

【数１】

$$E(M) = \sum_{n=1}^{N} \sum_{z \in Z} w(z|y_n) J(x_n, z; M) \qquad \cdots \text{式(1)}$$

【００２８】

ここで、ｗ（ｚ|ｙ）はクラスｙ∈Ｙのサンプルがターゲットクラスｚ∈Ｚのモデル学習にどのくらい参考になるかをあらわす重みを表す。なお、式（１）において、太字の文字（ここではx_nとＺ）は、複数の成分を有していることを示し、以下の他の式についても同様である。また、文章中の文字については、いずれも太字で示していないが、各式と整合をとったものであるものとする。

【００２９】

また、Ｊ（x_n，ｚ；Ｍ）はサンプルｘのクラスをｚと予測したときのモデルＭの誤差関数を表す。誤差関数の例として、
負の対数尤度Ｊ（ｘ，ｚ；Ｍ）＝－ｌｏｇＰ（ｚ｜ｘ；Ｍ）や、
０－１損失関数Ｊ（ｘ，ｚ；Ｍ）＝０（ｉｆ　ｆ（ｘ）＝ｙ），
Ｊ（ｘ，ｚ；Ｍ）＝１（otherwise）、などが考えられる。なお、本明細書では、対数は自然対数、すなわち、対数ｌｏｇの底は「ｅ」であるものとする。

【００３０】

重みを、以下のように決定する（動作主体については後記。以下同様）。まず、クラスｙにおける経験分布を近似するモデル分布Ｐ~（ｘ｜ｙ）（本明細書において、経験分布を意味する記号「~」はその直前の文字の上に付される記号であるものとする。後記する「^」についても同様）を推定する（式（２））。ここで、δ（ｘ，x_n）はクロネッカーのデルタを表し、Ｎ（ｙ）はクラスがｙであるサンプルの数を表す。

【数２】

$$\tilde{P}(x|y) \approx \frac{1}{N(y)} \sum_{n:y_n=y} \delta(x, x_n) \qquad \cdots \text{式(2)}$$

【００３１】

次に、モデル分布の全クラスの混合がターゲットクラスｚ∈Ｚの真の分布Ｐ（ｘ｜ｚ）を近似するように、混合比P_z（ｙ）を推定する（式（３））。ここで、混合比とは、ターゲット分類体系と補助分類体系とを統合した場合の確率分布モデルを、ターゲット分類

体系の確率分布モデルに近似させるための、ターゲット分類体系と補助分類体系とのクラスごとの、分類モデルに対する影響度の比率を示すものである。

【数３】

$$P(\boldsymbol{x}|z) \approx \sum_{y \in \boldsymbol{Y}} P_z(y)\tilde{P}(\boldsymbol{x}|y) \qquad \cdots \text{式(3)}$$

【００３２】

なお、混合比Ｐ$_{z}$（ｙ）、および、混合比Ｐ$_{z}$（ｙ）の集合Ｐは、
Ｐ＝｛｛Ｐ$_{z}$（ｙ）｝$_{y \in Y}$｝$_{z \in Z}$（０≦Ｐ$_{z}$（ｙ）≦１，Σ$_{y \in Y}$Ｐ$_{z}$（ｙ）＝１）
を満たすものとする。

【００３３】

そして、重みｗ（ｚ|ｙ）を設定（算出）する（式（４））。なお、Ｐ（ｚ）は、あるサンプルに関してクラスｚが選ばれる確率である。

【数４】

$$w(z|y) = \frac{P(z)P_z(y)}{N(y)} \qquad \cdots \text{式(4)}$$

【００３４】

このとき、重み付き誤差Ｅ（Ｍ）は期待誤差の近似となる（式（５））。

【数５】

$$\begin{aligned}
E(M) &= \sum_{n=1}^{N} \sum_{z \in \boldsymbol{Z}} w(z|y_n) J(\boldsymbol{x}_n, z; M) \\
&= \sum_{\boldsymbol{x}} \sum_{z \in \boldsymbol{Z}} J(\boldsymbol{x}, z; M) \sum_{y \in \boldsymbol{Y}} w(z|y) \sum_{n:y_n=y} \delta(\boldsymbol{x}, \boldsymbol{x}_n) \\
&= \sum_{\boldsymbol{x}} \sum_{z \in \boldsymbol{Z}} J(\boldsymbol{x}, z; M) \sum_{y \in \boldsymbol{Y}} \frac{P(z)P_z(y)}{N(y)} \sum_{n:y_n=y} \delta(\boldsymbol{x}, \boldsymbol{x}_n) \\
&\approx \sum_{\boldsymbol{x}} \sum_{z \in \boldsymbol{Z}} J(\boldsymbol{x}, z; M) P(z) \sum_{y \in \boldsymbol{Y}} P_z(y)\tilde{P}(\boldsymbol{x}|y) \\
&\approx \sum_{\boldsymbol{x}} \sum_{z \in \boldsymbol{Z}} J(\boldsymbol{x}, z; M) P(z) P(\boldsymbol{x}|z) \\
&= \sum_{\boldsymbol{x}} \sum_{z \in \boldsymbol{Z}} P(\boldsymbol{x}, z) J(\boldsymbol{x}, z; M) \\
&= \mathcal{E}_z[J(\boldsymbol{x}, z; M)] \qquad \cdots \text{式(5)}
\end{aligned}$$

【００３５】

式（５）において、右辺の１行目から２行目への式変形は、ｎについての総和の式をｘとｙについての総和の式に変えたものである。右辺の２行目から３行目への式変形は、式（４）を使ったものである。右辺の３行目から４行目への式変形は、式（２）を使ったものである。右辺の４行目から５行目への式変形は、式（３）を使ったものである。右辺の５行目から６行目への式変形は、条件付確率の公式（定義）を使ったものであり、Ｐ（ｘ，ｚ）はｘとｚが同時に発生する確率を示す。

【００３６】

右辺の６行目から７行目への式変形は、期待値の公式（定義）を使ったものであり、ε

$_z$[J（x，z；M）]はターゲットクラスzに関する誤差の期待値を示す。このため、補助データも利用した重み付き誤差E（M）を最小化することにより、頑健な（高精度な）モデルが推定できると期待できる。

【００３７】

式（３）の近似を満たす集合Pは、ターゲットデータに対する対数尤度L（P）をEM（Expectation-Maximization）アルゴリズムを用いて最大化することにより推定する（式（６））。EMアルゴリズムとは、E(Expectation)ステップとM(Maximization)ステップとの２つの手順を収束条件が満たされるまで繰り返すことで、パラメータ（ここでは集合P）の最尤推定を行うアルゴリズムである。

【数６】

$$L(\boldsymbol{P}) = \sum_{n=1}^{N_z} \log P(\boldsymbol{x}_n|y_n)$$
$$= \sum_{n=1}^{N_z} \log \sum_{y \in \boldsymbol{Y}} P_{y_n}(y)\tilde{P}_{-n}(\boldsymbol{x}_n|y) \quad \cdots \text{式(6)}$$

【００３８】

ここで、$\tilde{P}_{-n}$（x｜y）は、n番目のサンプルを除いたデータを用いて推定したモデル分布を表す。モデル分布の推定に用いたサンプルを用いて混合比を推定する場合、過学習を起こし、P_z（z）＝１、P_z（y≠z）＝０という自明な解が得られてしまうため、式（６）のようにleave-one-out（LOO）法を用いる。$\tilde{P}_{-n}$（x｜y）をクラスyのデータを用いて推定し固定した場合、L（P）はPに関して上に凸であるため、解の大域的最適性が保証される。EMアルゴリズムにおける第τステップでの推定値を$P^{(\tau)}$とする。ここで、τは、EステップとMステップとの２つの手順を繰り返した回数（τ＝０，１，２，…）を指す。なお、τ＝０のときには推定値の予め定められた初期値を示す。このとき、最大化すべき完全データ対数尤度の条件付き期待値Q（P｜$P^{(\tau)}$）は、式（７）のように表すことができる。

【数７】

$$Q(\boldsymbol{P}|\boldsymbol{P}^{(\tau)}) = \sum_{n=1}^{N_z} \sum_{y \in \boldsymbol{Y}} P_{y_n}(y|\boldsymbol{x}_n; \boldsymbol{P}^{(\tau)}) \log P_{y_n}(y)\tilde{P}_{-n}(\boldsymbol{x}_n|y) \quad \cdots \text{式(7)}$$

【００３９】

Eステップにおける計算は式（８）のように表すことができる。なお、式（８）の右辺の分母におけるy’は、式（８）の他の箇所におけるyと区別するために便宜上記号を変えたもので、yと同じ意味である。

【数８】

$$P_{y_n}(y|\boldsymbol{x}_n; \boldsymbol{P}^{(\tau)}) = \frac{P_{y_n}^{(\tau)}(y)\tilde{P}_{-n}(\boldsymbol{x}_n|y)}{\sum_{y' \in \boldsymbol{Y}} P_{y_n}^{(\tau)}(y')\tilde{P}_{-n}(\boldsymbol{x}_n|y')} \quad \cdots \text{式(8)}$$

【００４０】

Mステップにおける計算は式（９）のように表すことができる。

【数９】

$$P_z^{(\tau+1)}(y) = \frac{1}{N(z)} \sum_{n:y_n=z} P_{y_n}(y|\boldsymbol{x}_n; \boldsymbol{P}^{(\tau)}) \quad \cdots \text{式(9)}$$

【００４１】
このＥステップにおける計算とＭステップにおける計算を、収束条件が満たされるまで繰り返すことにより、集合Ｐの推定値が得られる。
【００４２】
なお、ＥＭアルゴリズムではなく、準ニュートン法など他の最適化手法を用いて式（６）を最大化することによっても、集合Ｐを推定できる。
【００４３】
＜重み推定＞
図２を参照しながら、重み推定部２１の構成について説明する。図２は、本実施形態に係る重み推定部のブロック図を含む図である。図２に示すように、重み推定部２１は、入力データ読込部２１１と、事後確率推定部２１２と、混合比推定部２１３と、重み書込部２１４とを備えている。
【００４４】
まず、入力データ読込部２１１により、入力データ４４を読み込む。そして、事後確率推定部２１２によって式（８）を用いて全学習用サンプルの全時刻に対する事後確率を推定し、また、混合比推定部２１３によって式（９）を用いて混合比を推定する。この事後確率推定と混合比推定を式（６）が収束するまで交互に繰り返し、重み書込部２１４において、
重みをｗ（ｚ|ｙ）＝Ｐ（ｚ）Ｐ$_{z}$（ｙ）／Ｎ（ｙ）と設定（算出）し、重み４５に格納する。なお、格納された重み４５は、モデル構築部２２で利用される。
【００４５】
＜モデル構築＞
図３を参照しながら、モデル構築部２２の構成について説明する。図３は、本実施形態に係るモデル構築部のブロック図を含む図である。図３に示すように、モデル構築部２２は、入力データ読込部２２１と、重み読込部２２２と、モデルパラメータ推定部２２３と、モデルパラメータ書込部２２４とを備えている。
【００４６】
まず、入力データ読込部２２１により、入力データ４４を読み込む。また、重み読込部２２２により、重み４５を読み込む。そして、モデルパラメータ推定部２２３によって式（１０）を用いてモデルパラメータＭ＾を推定する。
【数１０】

$$\hat{M} = \arg\min_{M} E(M) \qquad \cdots \text{式(10)}$$

なお、式（１０）の左辺においてＭに付した記号「＾（ハット）」は、そのＭがargmin関数の引数を最小化させることを示すものである。
【００４７】
モデルパラメータ書込部２２４は、モデルパラメータ推定部２２３が推定したモデルパラメータをモデルパラメータ４６に格納する。なお、格納されたモデルパラメータ４６は、分類部２３で利用される。
【００４８】
図４を参照しながら、分類部２３の構成について説明する。図４は、本実施形態に係る分類部のブロック図を含む図である。図４に示すように、分類部２３は、テストデータ読込部２３１と、モデルパラメータ読込部２３２と、分類結果出力部２３３とを備えている。
【００４９】
まず、テストデータ読込部２３１により、未分類のテストデータ４７を読み込む。また、モデルパラメータ読込部２３２により、モデルパラメータ４６を読み込む。そして、分類結果出力部２３３において、テストデータとモデルパラメータを使って分類結果を計算し、分類結果を出力する。

【００５０】
図１に示した分類装置１の動作について図５を参照（適宜図１参照）して説明する。図５は、本実施形態に係る分類装置の処理の流れを示す説明図である。
【００５１】
まず、分類装置１は、重み推定部２１によって、記憶手段４（図１参照）に予め格納された入力データ４４に基づいて重みを推定する（ステップＳ１０：重み推定ステップ）。推定された重みは、重み４５として記憶手段４に格納される。次に、分類装置１は、モデル構築部２２によって、記憶手段４（図１参照）に予め格納された入力データ４４および重み４５に基づいてモデルを構築する（ステップＳ２０：モデル構築ステップ）。構築されたモデルは、モデルパラメータ４６として記憶手段４に格納される。このステップＳ１０とステップＳ２０はモデルの学習に関する処理である。
【００５２】
続いて、分類装置１は、分類部２３によって、記憶手段４（図１参照）に予め格納された未分類であるテストデータ４７（分類対象データ）を、モデルパラメータ４６に基づいて分類する（ステップＳ３０：分類ステップ）。このステップＳ３０は分類対象データの分類に関する処理である。
【００５３】
次に、前記したステップＳ１０の重み推定ステップについて図６を参照（適宜図１ないし図５参照）して説明する。図６は、重み推定ステップの処理を示すフローチャートである。
【００５４】
まず、図６に示すように、重み推定部２１は、入力データ読込部２１１によって、記憶手段４（図１参照）から、入力データ４４を読み込む（ステップＳ１）。次に、重み推定部２１は、事後確率推定部２１２によって、モデル分布の推定を行う（ステップＳ２）。具体的には、前記した式（２）を満たすモデル分布を推定する。
【００５５】
その後、重み推定部２１は、事後確率推定部２１２によって、初期化を行う（ステップＳ３）。具体的には、事後確率推定部２１２は、ＥＭアルゴリズムのＥステップとＭステップとの２つの手順の繰り返し回数τを０に設定し、混合比P_z（ｙ）の分布をランダムに設定する。
【００５６】
次に、重み推定部２１は、事後確率推定部２１２によって、ＥＭアルゴリズムのＥステップを実行する（ステップＳ４）。具体的には、事後確率推定部２１２は、前記した式（８）により、前記事後確率を推定する。続いて、重み推定部２１は、混合比推定部２１３によって、ＥＭアルゴリズムのＭステップを実行する（ステップＳ５）。具体的には、混合比推定部２１３は、前記した式（９）により、前記混合比を推定する。次に、重み推定部２１は、混合比推定部２１３によって、収束条件が満たされたか否かを判別する（ステップＳ６）。具体的には、混合比推定部２１３は、前記した式（６）に示す尤度Ｌ（Ｐ）が収束したか否かを判別する。この収束の判別は、閾値や変化率などを使用することにより行うことができる。
【００５７】
収束条件が満たされた場合、すなわち前記した式（６）に示す尤度Ｌ（Ｐ）が収束した場合（ステップＳ６：Ｙｅｓ）、混合比推定部２１３は、重みｗ（ｚ|ｙ）を計算する（ステップＳ８）。具体的には、混合比推定部２１３は、
ｗ（ｚ|ｙ）＝Ｐ（ｚ）P_z（ｙ）／Ｎ（ｙ）の式を用いて重みを計算する。そして、重み推定部２１は、重み書込部２１４によって、その重みを、重み４５として、記憶手段４（図１参照）に書き込み、処理を終了する。
【００５８】
一方、ステップＳ６において、収束条件が満たされていない場合、すなわち前記した式（６）に示す尤度Ｌ（Ｐ）が収束していない場合（ステップＳ６：Ｎｏ）、重み推定部２

１は、ＥステップおよびＭステップの繰り返し回数τに「１」を加算し（τ＝τ＋１）（ステップＳ７）、ステップＳ４に戻る。

【００５９】

本実施形態によれば、分類装置１は、ターゲット分類体系における既分類データだけでなく、補助分類体系における既分類データも用い、誤差関数と重みとの積の総和である期待誤差（式（５）参照）を最小化させるように重みを推定し、その推定された重みと２種類の既分類データとを用いて分類モデルを生成することで、補助分類体系のデータも有効に利用し、ターゲット分類体系に関する高精度な分類モデルを生成することができる。

【００６０】

また、重み推定部２１が、事後確率推定部２１２と、混合比推定部２１３とを備えているので、例えば、事後確率推定部２１２が、ＥＭアルゴリズムにおけるＥステップを行い、かつ、混合比推定部２１３がＭステップを行うことで、混合比についての大域的最適解を求め、求めた混合比から重みを決定（推定）することができる。

【００６１】

また、例えば、モデルパラメータ推定部２２３が、式（１０）を用いてモデルパラメータを推定することができる。

【００６２】

また、分類部２３が、推定したモデルパラメータを用いて分類対象データをターゲット分類体系における複数のクラスのいずれかに分類する、つまり、高精度な分類モデルを用いることで高精度な分類を実現することができる。

【００６３】

また、分類装置１は、一般的なコンピュータに、前記した各処理のプログラムを実行させることで実現することもできる。このプログラムは、通信回線を介して配布することも可能であるし、ＣＤ－ＲＯＭ（Compact Disc-Read Only Memory）等の記録媒体に書き込んで配布することも可能である。

【００６４】

以上で本実施形態の説明を終えるが、本発明の態様はこれらに限定されるものではない。例えば、本発明は、任意の誤差関数およびモデルを用いることが可能である。その他、ハードウェアやフローチャート等の具体的な構成について、本発明の趣旨を逸脱しない範囲で適宜変更が可能である。

【実施例】

【００６５】

《人工データにおける実施例》

本実施形態の分類装置１を評価するため、人工データを用いた２クラス分類実験を行った。この２クラス分類実験とは、ターゲットデータと補助データから生成した分類モデルに基づき、テストデータを２つのクラスのいずれかに分類する実験である。

【００６６】

ターゲットデータは平均の異なる２つの１００次元正規分布からデータが生成されるものとする。ここで、クラスc_1、c_2の平均はそれぞれ$\mu_1=(-1, 0, 0, \cdots, 0)$，$\mu_2=(1, 0, 0, \cdots, 0)$であり、共分散行列はともに単位行列であるものとする。そして、補助データとして，以下の３パターンを考える。なお、第３次元以降の平均はターゲットデータと同じく全て０、共分散行列は全て単位行列とする。図７（ａ）にターゲットデータ，図７（ｂ）～（ｄ）に各補助データの生成モデルの第１，第２次元を示す。図７（ａ）～（ｄ）は、特に軸や目盛りを図示していないが、２次元の座標平面を表しており、中央部分が原点である。また、各円は標準偏差のラインを表す。

【００６７】

図７（ｂ）に示す同一補助データは、ターゲットデータと同一の生成モデルから生成され、クラスc_3、c_4の平均はそれぞれ$\mu_3=(-1, 0, 0, \cdots, 0)$，$\mu_4=(1, 0, 0, \cdots, 0)$である。

【００６８】
図７（ｃ）に示す相関補助データは、ターゲットデータとクラス間関係に相関がある生成モデルから生成され、クラスc_3、c_4の平均はそれぞれ
μ_3＝（－√０．５，√０．５，０，・・・，０），
μ_4＝（√０．５，－√０．５，０，・・・，０）である。
【００６９】
図７（ｄ）に示す混合補助データは、同一補助データ、および、ターゲットデータとクラス間関係が直交する補助データの組合せ（混合）であり、クラスc_3、c_4、c_5、c_6の平均はそれぞれ
μ_3＝（－１，０，０，・・・，０），μ_4＝（１，０，０，・・・，０），
μ_5＝（０，１，０，・・・，０），μ_6＝（０，－１，０，・・・，０）である。なお、補助データのうち、この混合補助データのみ４補助クラスであり、それ以外は２補助クラスである。
【００７０】
ターゲットデータとして各クラス２，４，８，１６，３２，６４，１２８，２５６サンプル（入力データ４４）、補助データとして各クラス２５６サンプル（入力データ４４）、テストデータとして各クラス１００サンプル（テストデータ４７）を生成した。これらに基づき、分類モデルを生成し、補助データを使わない場合（ターゲットデータのみ）と各補助データを使った場合の、テストデータの分類に関する正答率を計算した。その結果、表１のようになった。表１において、右４列の数字は平均正答率の百分率を示し、それぞれの括弧内の数字は標準偏差を示している。本実施形態の分類装置１の分類方法に基づいて補助データを使うことによって、補助データを使わない場合よりも正答率が向上していることがわかる。
【００７１】
【表１】

ターゲットデータ数	ターゲットデータのみ	同一補助データ	相関補助データ	混合補助データ
2	56.2 (4.4)	**65.0** (9.3)	**60.4** (7.7)	**64.0** (8.7)
4	60.9 (4.9)	**75.5** (7.4)	**65.0** (8.3)	**73.0** (7.7)
8	63.9 (4.3)	**80.4** (3.4)	**69.8** (6.6)	**78.8** (4.0)
16	68.5 (4.1)	**81.1** (2.8)	**74.2** (4.0)	**79.9** (2.7)
32	73.0 (3.7)	**81.2** (2.7)	**76.7** (3.5)	**81.6** (3.1)
64	77.6 (3.5)	**82.0** (2.7)	**78.3** (3.0)	**82.4** (2.4)
128	80.3 (3.0)	**82.8** (2.9)	**80.9** (2.7)	**82.7** (2.7)
256	82.2 (2.6)	**83.0** (2.1)	81.7 (3.1)	**83.0** (2.7)
平均	70.3 (9.7)	**78.9** (7.5)	**73.4** (8.9)	**78.2** (7.8)

【００７２】
《テキストデータにおける実施例》
本実施形態の分類装置１を評価するため、テキストデータを用いて分類実験を行った。
【００７３】
＜モデル分布＞
モデル分布$\tilde{P}(x|y)$として、正規分布、多項分布など任意の分布を仮定することができる。ここでは、入力データ４４およびテストデータ４７としてテキストデータを想定し、ｘを単語出現頻度ベクトルと考え、モデル分布として多項分布$\tilde{P}(x_n|y)$（式（１１））を用いる。

【数１１】

$$\tilde{P}(\boldsymbol{x}_n|y) \propto \prod_{j=1}^{V} \theta_{yj}^{x_{nj}} \qquad \cdots\text{式(11)}$$

ここで、Ｖは総語彙数、θ$_{yj}$はクラスｙのときｊ番目の単語が出現する確率、ｘ$_{nj}$はｎ番目のサンプルにおけるｊ番目の単語の出現頻度を表す。

【００７４】

多項分布のパラメータθ$_{yj}$のｎ番目のサンプルを除いたときのＬＯＯ最尤推定値θ^$_{-n,yj}$は式（１２）で得られる。

【数１２】

$$\hat{\theta}_{-n,yj} = \frac{\sum_{m:y_m=y} x_{mj} - x_{nj}}{\sum_{k=1}^{V} \sum_{m:y_m=y} x_{mk} - x_{nk}} \qquad \cdots\text{式(12)}$$

【００７５】

ここで、ゼロ確率問題を回避するために、ＬＯＯ最尤推定値と一様分布の線形和を用いてスムージングする（式（１３））。

【数１３】

$$\tilde{\theta}_{-n,yj} = \alpha\hat{\theta}_{-n,yj} + (1-\alpha)\frac{1}{V} \qquad \cdots\text{式(13)}$$

【００７６】

ここで、０≦α≦１はハイパーパラメータである。ハイパーパラメータを人手で設定してもよいが、一般化ＥＭアルゴリズムを用いることにより、以下のＱ（Ｐ，α｜Ｐ$^{(\tau)}$，α$^{(\tau)}$）を最大化するように、混合比の集合Ｐとハイパーパラメータαを同時にデータから推定することも可能である（式（１４））。

【数１４】

$$Q(\boldsymbol{P},\alpha|\boldsymbol{P}^{(\tau)},\alpha^{(\tau)}) = \sum_{n=1}^{N_x} \sum_{y\in\boldsymbol{Y}} P_{y_n}(y|\boldsymbol{x}_n;\boldsymbol{P}^{(\tau)},\alpha^{(\tau)}) \left(\log P_{y_n}(y) + \sum_{j=1}^{V} x_{nj} \log\left(\alpha\hat{\theta}_{-n,yj} + (1-\alpha)\frac{1}{V}\right)\right) \qquad \cdots\text{式(14)}$$

【００７７】

Ｅステップは式（８）、Ｍステップにおける混合比の更新は式（９）で、通常のＥＭアルゴリズムと同様に実現できる。Ｍステップにおけるハイパーパラメータの更新はニュートン法を用いて行う（式（１５））。

【数１５】

$$\alpha^{(\tau+1)} = \alpha^{(\tau)} - \frac{\partial Q(\boldsymbol{P},\alpha|\boldsymbol{P}^{(\tau)},\alpha^{(\tau)})}{\partial\alpha} \left(\frac{\partial^2 Q(\boldsymbol{P},\alpha|\boldsymbol{P}^{(\tau)},\alpha^{(\tau)})}{\partial\alpha^2} \right)^{-1} \qquad \cdots\text{式(15)}$$

【００７８】

ここで、式（１５）に記載されている式（１４）のαによる一階偏微分は式（１６）となる。

【数１６】

$$\frac{\partial Q(\boldsymbol{P},\alpha|\boldsymbol{P}^{(\tau)},\alpha^{(\tau)})}{\partial\alpha} = \sum_{n=1}^{N_x} \sum_{y\in\boldsymbol{Y}} P_{y_n}(y|\boldsymbol{x}_n;\boldsymbol{P}^{(\tau)},\alpha^{(\tau)}) \sum_{j=1}^{V} x_{nj} \left(\frac{\hat{\theta}_{-n,yj} - \frac{1}{V}}{\alpha\hat{\theta}_{-n,yj} + (1-\alpha)\frac{1}{V}} \right) \qquad \cdots\text{式(16)}$$

【００７９】
また、式（１５）に記載されている式（１４）のαによる二階偏微分は式（１７）となる。
【数１７】

$$\frac{\partial^2 Q(P,\alpha|P^{(\tau)},\alpha^{(\tau)})}{\partial\alpha^2} = -\sum_{n=1}^{N_z}\sum_{y\in Y} P_{y_n}(y|x_n;P^{(\tau)},\alpha^{(\tau)})\sum_{j=1}^{V} x_{nj}\left(\frac{\hat{\theta}_{-n,yj}-\frac{1}{V}}{\alpha\hat{\theta}_{-n,yj}+(1-\alpha)\frac{1}{V}}\right)^2 \quad \cdots式(17)$$

【００８０】
式（１７）から明らかなように、二階偏微分は常に負になるため、
Ｑ（Ｐ，α｜Ｐ（τ），α（τ））はαに関して上に凸である。この実験では、一般化ＥＭアルゴリズムを用いて混合比の集合Ｐおよびハイパーパラメータαをデータから推定した。
【００８１】
＜分類モデル＞
代表的なテキスト分類モデルであるナイーブベイズモデルとロジスティック回帰モデルをモデルＭとして用いた場合について説明する。
【００８２】
（ナイーブベイズモデル）
ナイーブベイズモデルではクラスが与えられたとき、文書中の各単語は独立に生成されると仮定され、クラスｚにおける単語出現頻度ベクトルｘの分布Ｐ（ｘ｜ｚ）が多項分布で表される（式（１８））。
【数１８】

$$P(\boldsymbol{x}|z) \propto \prod_{j=1}^{V} \phi_{zj}^{x_j} \quad \cdots式(18)$$

【００８３】
ここで、ϕ_{zj}はクラスｚの文書におけるj番目の単語が出現する確率を表す。誤差関数として負の対数尤度を用い、また、$\phi = \{\{\phi_{zj}\}_{j=1}^{V}\}_{z\in Z}$の事前確率としてディリクレ分布$P(\phi) \propto \Pi_{z\in Z}\Pi_{j=1}^{V}\phi_{zj}^{\beta}$を用いたとき、重み付き誤差関数$E(M_{NB})$は、式（１９）のように表される。
【数１９】

$$E(M_{NB}) = -\sum_{n=1}^{N}\sum_{z\in Z} w(z|y_n)\sum_{j=1}^{V} x_{nj}\log\phi_{zj} + \beta\sum_{z\in Z}\sum_{j=1}^{V}\log\phi_{zj} \quad \cdots式(19)$$

【００８４】
式（１９）を最小化するϕ_{zj}の推定値$\hat{\phi}_{zj}$は、式（２０）によって得られる。
【数２０】

$$\hat{\phi}_{zj} = \frac{\sum_{n=1}^{N} w(z|y_n)x_{nj} + \beta}{\sum_{k=1}^{V}\sum_{n=1}^{N} w(z|y_n)x_{nk} + \beta V} \quad \cdots式(20)$$

【００８５】
（ロジスティック回帰モデル）
ロジスティック回帰モデルでは、単語出現頻度ベクトルｘが与えられたとき、クラスｚ

に属する確率Ｐ（ｚ｜ｘ）は式（２１）のように表される。
【数２１】

$$P(z|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\lambda}_z^T \boldsymbol{x})}{\sum_{z' \in Z} \exp(\boldsymbol{\lambda}_{z'}^T \boldsymbol{x})} \quad \cdots \text{式(21)}$$

【００８６】
ここで、λ_zはクラスｚに関する未知パラメータベクトル、λ_z^Tはλ_zの転置を表す。誤差関数として負の対数尤度を用い、また、λ_zの事前確率として平均０、共分散行列γ^{-1}Ｉ（Ｉは単位行列）の正規分布を用いたとき、重み付き誤差（期待誤差）Ｅ（M_{LR}）は、式（２２）のように表される。
【数２２】

$$E(M_{LR}) = -\sum_{n=1}^{N} \sum_{z \in Z} w(z|y_n) \left(\boldsymbol{\lambda}_z^T \boldsymbol{x}_n - \log \sum_{z' \in Z} \exp(\boldsymbol{\lambda}_{z'}^T \boldsymbol{x}_n) \right) + \frac{\gamma}{2} \sum_{z \in Z} \| \boldsymbol{\lambda}_z \|^2 \quad \cdots \text{式(22)}$$

【００８７】
準ニュートン法などを用いて式（２２）の値を最小化することにより、未知パラメータベクトル$\{\lambda_z\}_{z \in Z}$を推定できる。ロジスティック回帰モデルを用いた場合、各サンプルの誤差関数を付加するのみであるため、これまで提案されている多くの分類モデルを若干修正するのみで適用することができる。
【００８８】
＜比較手法＞
分類モデルとしてナイーブベイズモデルを用いた本手法（本実施形態の分類装置１による手法）（ＣＡ－ＮＢ）と、分類モデルとしてロジスティック回帰モデルを用いた本手法（ＣＡ－ＬＲ）と、補助データを用いないナイーブベイズモデルによる手法（ＮＢ）、ロジスティック回帰モデルによる手法（ＬＲ）の４手法を比較した。ＮＢの推定値は、推定値である式（２０）の重みを
ｗ（ｚ｜ｚ）＝１，ｗ（ｚ｜ｙ≠ｚ）＝０としたものである。同様に、ＬＲの推定値は、本手法における重み付き誤差である式（２２）の重みを
ｗ（ｚ｜ｚ）＝１，ｗ（ｚ｜ｙ≠ｚ）＝０として最小化することにより得られる。
【００８９】
それぞれの実験において評価用データセットを１００作成し、その平均正答率を用いて評価した。また、評価用データセットとは別に１つの開発用データセットを作成し、各手法において開発用データセットの正答率を最も高くする分類モデルのハイパーパラメータ（βもしくはγ）を｛10^{-3}，10^{-2}，10^{-1}，１｝の４候補から選択した。
【００９０】
＜Ｔｏｙデータ＞
20Newsgroups（20news）から作成したデータセットを用い、各補助クラスの分布が、あるターゲットクラスと同じ分布である場合の、本手法の効果を評価する。20newsは、２０のディスカッショングループに投稿された約２万の英語文書から成る。各文書の特徴量として単語出現頻度を用いた。このとき、停止語（文書に含まれる意味的な内容を持たない前置詞や冠詞などの一般的に機能語と呼ばれ検索に役立たない単語）および出現頻度が１以下の単語は省き、総語彙数は52,647であった。
【００９１】
２０のグループのうち、コンピュータ（comp）を親ディレクトリにもつ５つのグループ（graphics, os.ms-windows.misc, sys.ibm.pc.hardware, sys.mac.hardware, windows.x）に分類する問題について、
ターゲットクラス集合をＺ＝｛c_1，・・・，c_5｝、

補助クラス集合をA＝｛c_6，・・・，c_{10}｝とする。

【００９２】

そして、graphicsの記事をターゲットクラスc_1もしくは補助クラスc_6に、os.ms-windows.miscの記事をターゲットクラスc_2もしくは補助クラスc_7に、sys.ibm.pc.hardwareの記事をターゲットクラスc_3もしくは補助クラスc_8に、sys.mac.hardwareの記事をターゲットクラスc_4もしくは補助クラスc_9に、windows.xの記事をターゲットクラスc_5もしくは補助クラスc_{10}に、ランダムに割り当て、ターゲットデータおよび補助データを作成した。

【００９３】

このとき、テストデータとして各クラス１００サンプル、ターゲットデータとして各クラス２，４，８，１６，３２，６４，１２８，２５６サンプル、補助データとして残り全サンプル用いた。総学習サンプル数は4,363であった。このときの正答率を表２に示す。表２において、右４列の数字は平均正答率の百分率を示し、それぞれの括弧内の数字は標準偏差を示している。

【００９４】

【表２】

ターゲットデータ数	NB	LR	CA-NB	CA-LR
2	37.8 (4.0)	39.0 (4.6)	79.8 (9.6)	**79.9** (10.0)
4	42.3 (4.1)	45.3 (4.6)	**83.6** (3.1)	83.5 (3.9)
8	47.7 (3.4)	53.1 (3.7)	84.8 (2.1)	**85.1** (2.0)
16	54.8 (3.8)	61.4 (2.6)	85.5 (1.8)	**86.2** (1.5)
32	61.7 (2.9)	67.8 (2.4)	85.7 (1.7)	**86.3** (1.5)
64	68.0 (3.1)	73.4 (1.9)	85.7 (1.8)	**86.3** (1.6)
128	73.4 (3.0)	78.3 (1.8)	86.2 (1.3)	**86.8** (1.5)
256	78.6 (2.2)	81.8 (1.8)	86.2 (1.4)	**87.0** (1.5)
平均	58.0 (14.3)	62.5 (14.9)	84.7 (4.4)	**85.1** (4.6)

【００９５】

本手法であるＣＡ－ＮＢ、ＣＡ－ＬＲの正答率は学習サンプル数が少ない場合でも極めて高く、補助データを適切に利用することにより、頑健な（高精度な）モデル推定ができていると言える。

【００９６】

＜20Newsgroupsデータ＞

20newsの２０グループのうち、comp.graphics, rec.sport.baseball, sci.electronics, talk.religion.miscの４グループをターゲットクラスとし、他の１６グループを補助クラスとしてデータを作成し、本手法を評価した。テストデータ４７として各クラス１００サンプル、ターゲットデータ（入力データ４４）として各クラス２，４，８，１６，３２，６４，１２８，２５６サンプル、補助データ（入力データ４４）として全サンプル用いた、総補助サンプル数は15,211であった。このときの正答率を表３に示す。表３において、右４列の数字は平均正答率の百分率を示し、それぞれの括弧内の数字は標準偏差を示している。本手法であるＣＡ－ＮＢの正答率が最も高くなっている。

【００９７】

【表３】

ターゲットデータ数	NB	LR	CA-NB	CA-LR
2	52.4 (4.9)	53.7 (5.1)	**74.3** (7.3)	69.7 (7.1)
4	62.5 (5.3)	64.9 (4.6)	**79.0** (4.5)	72.7 (4.8)
8	72.9 (3.5)	73.8 (3.3)	**81.5** (3.6)	75.4 (3.7)
16	80.9 (2.7)	81.1 (2.3)	**86.1** (2.4)	80.8 (3.0)
32	87.4 (2.1)	86.6 (1.8)	**89.7** (1.8)	86.5 (2.4)
64	91.9 (1.3)	90.8 (1.3)	**92.9** (1.1)	91.2 (1.3)
128	94.7 (1.2)	93.3 (1.3)	**95.2** (1.2)	93.9 (1.2)
256	96.3 (1.0)	95.2 (1.0)	**96.4** (1.0)	95.8 (1.0)
平均	79.9 (15.3)	79.9 (14.1)	**86.9** (8.3)	83.2 (10.1)

【００９８】

＜Webページデータ＞

日本語のディレクトリ型検索エンジンgoo（登録商標）カテゴリ検索（２００３年９月取得）とyahoo（登録商標）カテゴリ（２００３年３月取得）のデータを用いて本手法を評価した。形態素解析により単語を抽出し、両カテゴリで出現数が１０以上の単語を特徴量として用いた。このとき、総語彙数は43,200であった。goo（登録商標）とyahoo（登録商標）でクラスラベルが同一のクラスや、関連していると思われるクラスもあるが、明確な対応付けが難しいクラスもあり、また、クラス数も異なる（goo（登録商標）：１３クラス、yahoo（登録商標）：１４クラス）。

【００９９】

goo（登録商標）ディレクトリのクラスをターゲットクラスとし、テストデータ４７として各クラス１００サンプル、ターゲットデータ（入力データ４４）として各クラス２，４，８，１６，３２，６４，１２８，２５６サンプル、補助データ（入力データ４４）としてyahoo（登録商標）ディレクトリに含まれる全サンプル用いた。総補助サンプル数は51,728であった。このときの正答率を表４に示す。表４において、右４列の数字は平均正答率の百分率を示し、それぞれの括弧内の数字は標準偏差を示している。本手法であるＣＡ－ＮＲ、ＣＡ－ＬＲの正答率が総じて高くなっている。

【０１００】

【表４】

ターゲットデータ数	NB	LR	CA-NB	CA-LR
2	20.0 (2.5)	18.5 (2.5)	**27.8** (4.6)	26.7 (4.6)
4	25.8 (2.4)	23.5 (2.0)	**33.4** (3.3)	31.0 (3.1)
8	32.7 (2.2)	29.9 (2.0)	**40.0** (2.8)	37.2 (2.7)
16	39.8 (1.6)	37.4 (1.7)	**45.5** (2.4)	43.4 (2.6)
32	46.3 (1.5)	45.0 (1.5)	**50.5** (1.6)	49.5 (1.7)
64	51.8 (1.6)	51.3 (1.4)	**53.7** (1.7)	**53.7** (1.7)
128	56.1 (1.3)	56.7 (1.4)	57.1 (1.5)	**58.5** (1.5)
256	59.1 (1.3)	61.1 (1.2)	59.2 (1.3)	**62.4** (1.3)
平均	41.5 (13.5)	40.4 (14.8)	**45.9** (10.9)	45.3 (12.4)

【図面の簡単な説明】

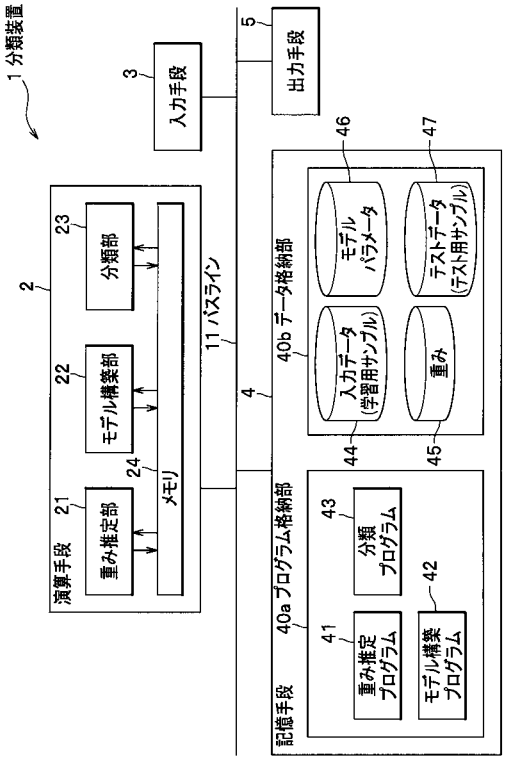
【０１０１】

【図１】本実施形態に係る分類装置の構成を示すブロック図である。

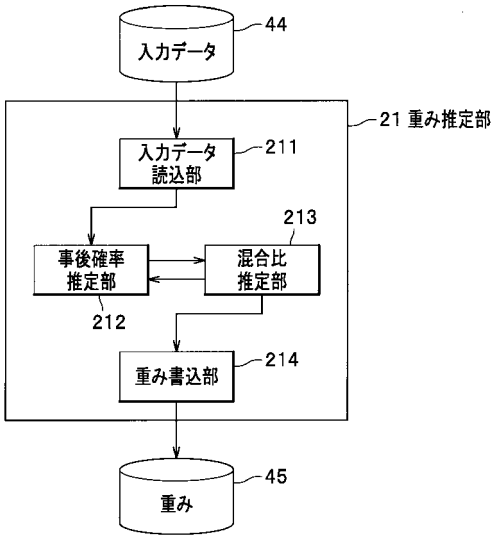
【図２】本実施形態に係る重み推定部のブロック図を含む図である。

【図３】本実施形態に係るモデル構築部のブロック図を含む図である。
【図４】本実施形態に係る分類部のブロック図を含む図である。
【図５】本実施形態に係る分類装置の処理の流れを示す説明図である。
【図６】重み推定ステップの処理を示すフローチャートである。
【図７】（ａ）はターゲットデータ、（ｂ）～（ｄ）は各補助データの生成モデルの第１、第２次元を示す図である。
【符号の説明】
【０１０２】
１　分類装置
２　演算手段
３　入力手段
４　記憶手段
５　出力手段
１１　バスライン
２１　重み推定部
２２　モデル構築部
２３　分類部
２４　メモリ
４０ａ　プログラム格納部
４１　重み推定プログラム
４２　モデル構築プログラム
４３　分類プログラム
４０ｂ　データ格納部
４４　入力データ
４５　重み
４６　モデルパラメータ
４７　テストデータ
２１１　入力データ読込部
２１２　事後確率推定部
２１３　混合比推定部
２１４　重み書込部
２２１　入力データ読込部
２２２　重み読込部
２２３　モデルパラメータ推定部
２２４　モデルパラメータ書込部
２３１　テストデータ読込部
２３２　モデルパラメータ読込部
２３３　分類結果出力部

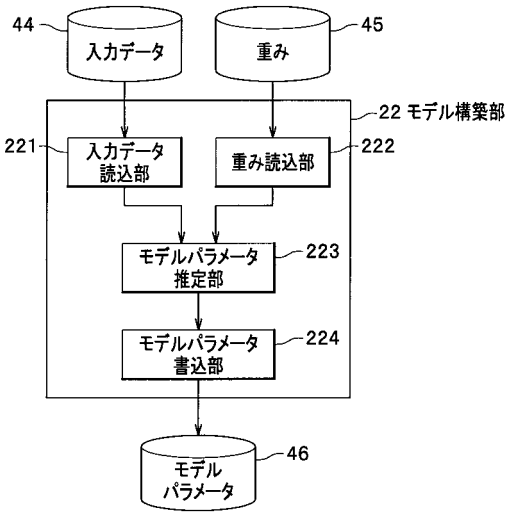
【図１】


1 分類装置
3 入力手段
5 出力手段
2
23 分類部
22 モデル構築部
21 重み推定部
演算手段
24 メモリ
11 バスライン
4
40b データ格納部
46 モデルパラメータ
47 テストデータ（テスト用サンプル）
入力データ（学習用サンプル）
重み
44
45
記憶手段
40a プログラム格納部
43 分類プログラム
42 モデル構築プログラム
41 重み推定プログラム

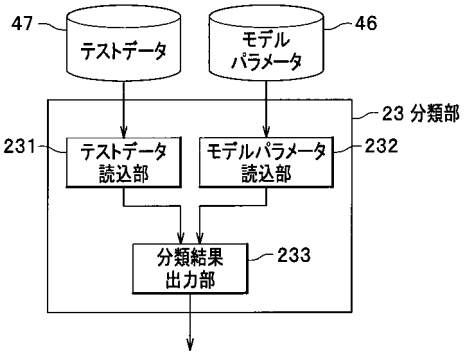
【図２】


44 入力データ
21 重み推定部
入力データ読込部 211
213
事後確率推定部
混合比推定部
212
重み書込部 214
45 重み

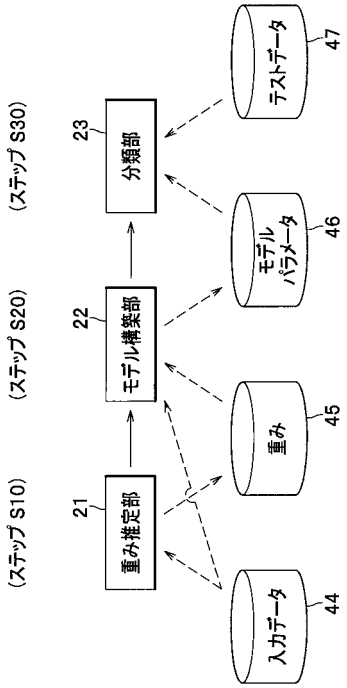
【図３】


44 入力データ
45 重み
22 モデル構築部
221 入力データ読込部
重み読込部 222
モデルパラメータ推定部 223
モデルパラメータ書込部 224
46 モデルパラメータ

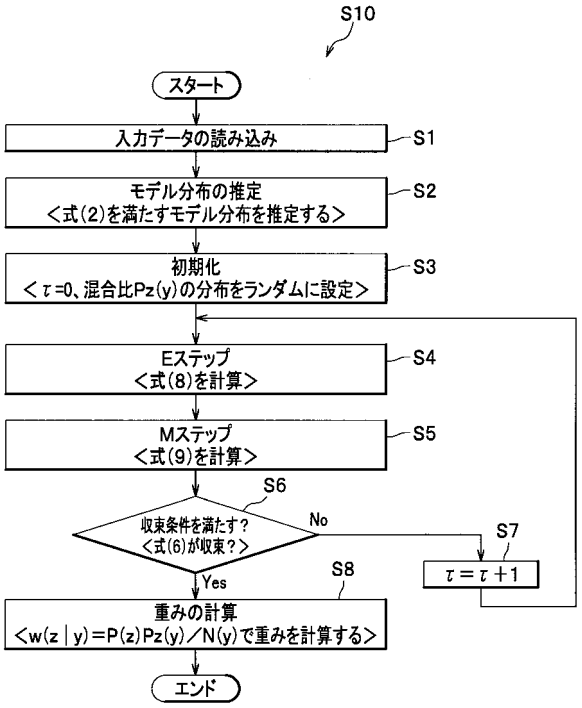
【図４】


47 テストデータ
46 モデルパラメータ
23 分類部
231 テストデータ読込部
モデルパラメータ読込部 232
分類結果出力部 233

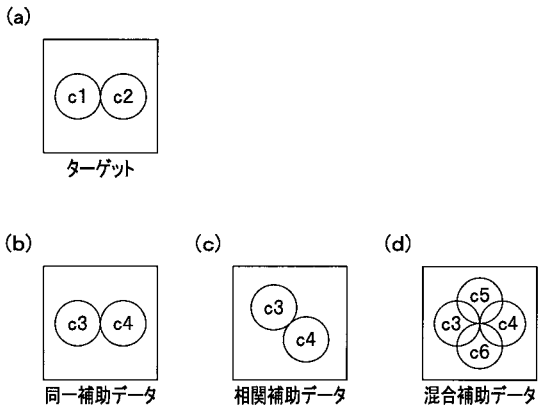
【図５】


テストデータ
47
分類部
23
(ステップ S30)
モデルパラメータ
46
モデル構築部
22
(ステップ S20)
重み
45
重み推定部
21
(ステップ S10)
入力データ
44

【図６】


S10
スタート
入力データの読み込み
S1
モデル分布の推定
<式(2)を満たすモデル分布を推定する>
S2
初期化
<τ=0、混合比Pz(y)の分布をランダムに設定>
S3
Eステップ
<式(8)を計算>
S4
Mステップ
<式(9)を計算>
S5
S6
収束条件を満たす？
<式(6)が収束？>
No
S7
τ=τ+1
Yes
S8
重みの計算
<w(z|y)=P(z)Pz(y)／N(y)で重みを計算する>
エンド

【図７】


(a)
c1
c2
ターゲット
(b)
c3
c4
同一補助データ
(c)
c3
c4
相関補助データ
(d)
c5
c3
c4
c6
混合補助データ

フロントページの続き

(72)発明者　田中　利幸
京都府京都市左京区吉田本町　国立大学法人京都大学大学院情報学研究科内

Ｆターム(参考) 5B075 NR02