

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5164209号  
(P5164209)

(45) 発行日 平成25年3月21日(2013.3.21)

(24) 登録日 平成24年12月28日(2012.12.28)

(51) Int.Cl.		F I			
<b>G06F</b>	<b>17/30</b>	<b>(2006.01)</b>	G06F	17/30	210D
<b>G06N</b>	<b>5/04</b>	<b>(2006.01)</b>	G06N	5/04	580A
			G06N	5/04	550J

請求項の数 11 (全 22 頁)

(21) 出願番号	特願2008-161237 (P2008-161237)	(73) 特許権者	000004226
(22) 出願日	平成20年6月20日 (2008.6.20)		日本電信電話株式会社
(65) 公開番号	特開2010-3106 (P2010-3106A)		東京都千代田区大手町二丁目3番1号
(43) 公開日	平成22年1月7日 (2010.1.7)	(73) 特許権者	504132272
審査請求日	平成23年5月27日 (2011.5.27)		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
		(74) 代理人	100162374
			弁理士 中村 新二
		(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) 【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

10

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、

20

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。

10

#### 【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

20

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、

30

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。

40

#### 【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

50

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

10

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】

20

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】

30

【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。

40

【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

50

## 【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、混合比についての全域的最適解を求め、求めた混合比から重みを決定（推定）することができる。

10

20

30

40

50

## 【 0 0 1 0 】

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。

## 【 0 0 1 1 】

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 ( 1 0 ) を用いてモデルパラメータを推定することができる。

## 【 0 0 1 2 】

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。

10

## 【 0 0 1 3 】

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。

## 【 0 0 1 4 】

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。

20

## 【 0 0 1 5 】

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 】

## 【 0 0 1 6 】

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 】

30

## 【 0 0 1 7 】

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 1 1 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。

## 【 0 0 1 8 】

40

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 2 1 と、モデル構築部 2 2 と、分類部 2 3 と、メモリ 2 4 とを含んで構成される。なお、各部 2 1 ~ 2 3 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 2 1 であるので、重み推定部 2 1 に関して特に詳細に説明する。また、モデル構築部 2 2 と分類部 2 3 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。

## 【 0 0 1 9 】

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 ) 。

50

## 【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、「 $\{x_n, y_n\}_{n=1}^{N_z}$ 」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を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 $x_n = (x_{n1}, \dots, x_{nw})$ 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  $1 \leq n \leq N_z$ ）、 $y_n \in A$ （if  $N_z + 1 \leq n \leq N$ ）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】

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

10

20

30

40

50

## 【0028】

ここで、 $w(z|y)$  はクラス  $y \in Y$  のサンプルがターゲットクラス  $z \in Z$  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  $J(x, z; M) = -\log P(z|x; M)$  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  $\tilde{P}(x|y)$  (本明細書において、経験分布を意味する記号「 $\tilde{\cdot}$ 」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「 $\wedge$ 」についても同様)を推定する(式(2))。ここで、 $\delta(x, x_n)$  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】

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

## 【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  $z \in Z$  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】

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

## 【0032】

なお、混合比  $P_z(y)$ 、および、混合比  $P_z(y)$  の集合  $P$  は、 $P = \{ \{ P_z(y) \}_{y \in Y} \}_{z \in Z}$  ( $0 \leq P_z(y) \leq 1$ ,  $\sum_{y \in Y} P_z(y) = 1$ ) 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】

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

## 【0034】

10

20

30

40

50

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】

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

10

20

【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、 $\mathcal{E}_z[J(x, z; M)]$  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。

30

【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】

$$\begin{aligned}
 L(P) &= \sum_{n=1}^{N_x} \log P(\mathbf{x}_n|y_n) \\
 &= \sum_{n=1}^{N_x} \log \sum_{y \in Y} P_{y_n}(y) \tilde{P}_{-n}(\mathbf{x}_n|y) \quad \dots \text{式(6)}
 \end{aligned}$$

40

【0038】

ここで、 $\tilde{P}_{-n}(\mathbf{x}|y)$  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

50



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。 $P_{-n}(x|y)$ 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  $\tau$  ステップでの推定値を  $P^{(\tau)}$  とする。ここで、 $\tau$  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 ( $\tau = 0, 1, 2, \dots$ ) を指す。なお、 $\tau = 0$  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】

$$Q(P|P^{(\tau)}) = \sum_{n=1}^{N_z} \sum_{y \in Y} P_{y_n}(y|x_n; P^{(\tau)}) \log P_{y_n}(y) \bar{P}_{-n}(x_n|y) \quad \dots \text{式(7)}$$

10

【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】

$$P_{y_n}(y|x_n; P^{(\tau)}) = \frac{P_{y_n}^{(\tau)}(y) \bar{P}_{-n}(x_n|y)}{\sum_{y' \in Y} P_{y_n}^{(\tau)}(y') \bar{P}_{-n}(x_n|y')} \quad \dots \text{式(8)}$$

20

【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】

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

30

【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。

40

【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。

50

## 【 0 0 4 5 】

## &lt;モデル構築&gt;

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。

## 【 0 0 4 6 】

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 $M^{\wedge}$ を推定する。

## 【数10】

$$\hat{M} = \arg \min_M E(M) \quad \dots \text{式(10)}$$

なお、式(10)の左辺においてMに付した記号「 $\hat{\quad}$ (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。

## 【 0 0 4 7 】

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。

## 【 0 0 4 8 】

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。

## 【 0 0 4 9 】

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。

## 【 0 0 5 0 】

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。

## 【 0 0 5 1 】

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。

## 【 0 0 5 2 】

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。

## 【 0 0 5 3 】

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。

10

20

30

40

50

## 【 0 0 5 4 】

まず、図 6 に示すように、重み推定部 2 1 は、入力データ読込部 2 1 1 によって、記憶手段 4 ( 図 1 参照 ) から、入力データ 4 4 を読み込む ( ステップ S 1 )。次に、重み推定部 2 1 は、事後確率推定部 2 1 2 によって、モデル分布の推定を行う ( ステップ S 2 )。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。

## 【 0 0 5 5 】

その後、重み推定部 2 1 は、事後確率推定部 2 1 2 によって、初期化を行う ( ステップ S 3 )。具体的には、事後確率推定部 2 1 2 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。

10

## 【 0 0 5 6 】

次に、重み推定部 2 1 は、事後確率推定部 2 1 2 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 S 4 )。具体的には、事後確率推定部 2 1 2 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 2 1 は、混合比推定部 2 1 3 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 S 5 )。具体的には、混合比推定部 2 1 3 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 2 1 は、混合比推定部 2 1 3 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 S 6 )。具体的には、混合比推定部 2 1 3 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。

20

## 【 0 0 5 7 】

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 S 6 : Yes )、混合比推定部 2 1 3 は、重み  $w ( z | y )$  を計算する ( ステップ S 8 )。具体的には、混合比推定部 2 1 3 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 2 1 は、重み書込部 2 1 4 によって、その重みを、重み 4 5 として、記憶手段 4 ( 図 1 参照 ) に書き込み、処理を終了する。

## 【 0 0 5 8 】

一方、ステップ S 6 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 S 6 : 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に「 1 」を加算し (  $= + 1$  ) ( ステップ S 7 )、ステップ S 4 に戻る。

30

## 【 0 0 5 9 】

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使い、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。

## 【 0 0 6 0 】

また、重み推定部 2 1 が、事後確率推定部 2 1 2 と、混合比推定部 2 1 3 とを備えているので、例えば、事後確率推定部 2 1 2 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 2 1 3 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。

40

## 【 0 0 6 1 】

また、例えば、モデルパラメータ推定部 2 2 3 が、式 ( 1 0 ) を用いてモデルパラメータを推定することができる。

## 【 0 0 6 2 】

また、分類部 2 3 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。

50

## 【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り、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 $\mu_1 = (-1, 0, 0, \dots, 0)$ 、 $\mu_2 = (1, 0, 0, \dots, 0)$ 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、特に軸や目盛りを明示していないが、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

$\mu_3 = (-1, 0, 0, \dots, 0)$ 、 $\mu_4 = (1, 0, 0, \dots, 0)$ 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

$\mu_3 = (-0.5, 0.5, 0, \dots, 0)$ 、

$\mu_4 = (0.5, -0.5, 0, \dots, 0)$ 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

$\mu_3 = (-1, 0, 0, \dots, 0)$ 、 $\mu_4 = (1, 0, 0, \dots, 0)$ 、

$\mu_5 = (0, 1, 0, \dots, 0)$ 、 $\mu_6 = (0, -1, 0, \dots, 0)$ 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

10

20

30

40

50

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。

【 0 0 7 1 】

【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)	<b>65.0 (9.3)</b>	<b>60.4 (7.7)</b>	<b>64.0 (8.7)</b>
4	60.9 (4.9)	<b>75.5 (7.4)</b>	<b>65.0 (8.3)</b>	<b>73.0 (7.7)</b>
8	63.9 (4.3)	<b>80.4 (3.4)</b>	<b>69.8 (6.6)</b>	<b>78.8 (4.0)</b>
16	68.5 (4.1)	<b>81.1 (2.8)</b>	<b>74.2 (4.0)</b>	<b>79.9 (2.7)</b>
32	73.0 (3.7)	<b>81.2 (2.7)</b>	<b>76.7 (3.5)</b>	<b>81.6 (3.1)</b>
64	77.6 (3.5)	<b>82.0 (2.7)</b>	<b>78.3 (3.0)</b>	<b>82.4 (2.4)</b>
128	80.3 (3.0)	<b>82.8 (2.9)</b>	<b>80.9 (2.7)</b>	<b>82.7 (2.7)</b>
256	82.2 (2.6)	<b>83.0 (2.1)</b>	81.7 (3.1)	<b>83.0 (2.7)</b>
平均	70.3 (9.7)	<b>78.9 (7.5)</b>	<b>73.4 (8.9)</b>	<b>78.2 (7.8)</b>

10

【 0 0 7 2 】

《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。

20

【 0 0 7 3 】

<モデル分布>

モデル分布  $P \sim (x | y)$  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 4 4 およびテストデータ 4 7 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  $P \sim (x_n | y)$  (式 (11)) を用いる。

【数 1 1】

$$\tilde{P}(x_n|y) \propto \prod_{j=1}^V \theta_{yj}^{x_{nj}} \quad \dots \text{式(11)}$$

30

ここで、 $V$  は総語彙数、 $\theta_{yj}$  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。

【 0 0 7 4 】

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

【数 1 2】

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

40

【 0 0 7 5 】

ここで、ゼロ確率問題を回避するために、L O O 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 1 3】

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

50

## 【 0 0 7 6 】

ここで、 $\alpha$  はハイパーパラメータである。ハイパーパラメータを手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  $Q(P, \alpha | P^{(\tau)}, \alpha^{(\tau)})$  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  $\alpha$  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】

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

10

## 【 0 0 7 7 】

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】

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

## 【 0 0 7 8 】

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

## 【数16】

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

## 【 0 0 7 9 】

また、式(15)に記載されている式(14)の  $\alpha$  による二階偏微分は式(17)となる。

## 【数17】

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

## 【 0 0 8 0 】

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、 $Q(P, \alpha | P^{(\tau)}, \alpha^{(\tau)})$  は  $\alpha$  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  $\alpha$  をデータから推定した。

30

40

## 【 0 0 8 1 】

<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。

## 【 0 0 8 2 】

(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  $z$  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】

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

【0083】

ここで、 $\phi_{zj}$  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、 $\{\phi_{zj}\}_{j=1}^V$   $z \in \mathcal{Z}$  の事前確率としてディリクレ分布  $P(\phi_{zj})$  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。

10

【数19】

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

【0084】

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

【数20】

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

【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  $\mathbf{x}$  が与えられたとき、クラス  $z$  に属する確率  $P(z|\mathbf{x})$  は式(21)のように表される。

【数21】

$$P(z|\mathbf{x}) = \frac{\exp(\lambda_z^T \mathbf{x})}{\sum_{z' \in \mathcal{Z}} \exp(\lambda_{z'}^T \mathbf{x})} \quad \dots \text{式(21)} \quad 30$$

【0086】

ここで、 $\lambda_z$  はクラス  $z$  に関する未知パラメータベクトル、 $\lambda_z^T$  は  $\lambda_z$  の転置を表す。誤差関数として負の対数尤度を用い、また、 $\lambda_z$  の事前確率として平均0、共分散行列  $\gamma^{-1} I$  ( $I$  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】

$$E(M_{LR}) = - \sum_{n=1}^N \sum_{z \in \mathcal{Z}} w(z|y_n) \left( \lambda_z^T \mathbf{x}_n - \log \sum_{z' \in \mathcal{Z}} \exp(\lambda_{z'}^T \mathbf{x}_n) \right) + \frac{\gamma}{2} \sum_{z \in \mathcal{Z}} \|\lambda_z\|^2 \quad \dots \text{式(22)} \quad 40$$

【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  $\{\lambda_z\}_{z \in \mathcal{Z}}$  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】

&lt;比較手法&gt;

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

50

(C A - L R) 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 (N B)、ロジスティック回帰モデルによる手法 (L R) の 4 手法を比較した。N B の推定値は、推定値である式 (20) の重みを

$w(z|z) = 1$ ,  $w(z|y \neq z) = 0$  としたものである。同様に、L R 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

$w(z|z) = 1$ ,  $w(z|y \neq z) = 0$  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 (もしくは ) を  $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$  の 4 候補から選択した。

10

【0090】

<T o y 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

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  $Z = \{c_1, \dots, c_5\}$ 、

補助クラス集合を  $A = \{c_6, \dots, c_{10}\}$  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。

30

【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)	<b>79.9 (10.0)</b>
4	42.3 (4.1)	45.3 (4.6)	<b>83.6 (3.1)</b>	83.5 (3.9)
8	47.7 (3.4)	53.1 (3.7)	84.8 (2.1)	<b>85.1 (2.0)</b>
16	54.8 (3.8)	61.4 (2.6)	85.5 (1.8)	<b>86.2 (1.5)</b>
32	61.7 (2.9)	67.8 (2.4)	85.7 (1.7)	<b>86.3 (1.5)</b>
64	68.0 (3.1)	73.4 (1.9)	85.7 (1.8)	<b>86.3 (1.6)</b>
128	73.4 (3.0)	78.3 (1.8)	86.2 (1.3)	<b>86.8 (1.5)</b>
256	78.6 (2.2)	81.8 (1.8)	86.2 (1.4)	<b>87.0 (1.5)</b>
平均	58.0 (14.3)	62.5 (14.9)	84.7 (4.4)	<b>85.1 (4.6)</b>

10

## 【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.baseba11, 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。

20

## 【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)	<b>74.3 (7.3)</b>	69.7 (7.1)
4	62.5 (5.3)	64.9 (4.6)	<b>79.0 (4.5)</b>	72.7 (4.8)
8	72.9 (3.5)	73.8 (3.3)	<b>81.5 (3.6)</b>	75.4 (3.7)
16	80.9 (2.7)	81.1 (2.3)	<b>86.1 (2.4)</b>	80.8 (3.0)
32	87.4 (2.1)	86.6 (1.8)	<b>89.7 (1.8)</b>	86.5 (2.4)
64	91.9 (1.3)	90.8 (1.3)	<b>92.9 (1.1)</b>	91.2 (1.3)
128	94.7 (1.2)	93.3 (1.3)	<b>95.2 (1.2)</b>	93.9 (1.2)
256	96.3 (1.0)	95.2 (1.0)	<b>96.4 (1.0)</b>	95.8 (1.0)
平均	79.9 (15.3)	79.9 (14.1)	<b>86.9 (8.3)</b>	83.2 (10.1)

30

40

## 【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）。

50

## 【 0 0 9 9 】

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。

## 【 0 1 0 0 】

## 【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)

10

20

## 【図面の簡単な説明】

## 【 0 1 0 1 】

【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。

30

## 【符号の説明】

## 【 0 1 0 2 】

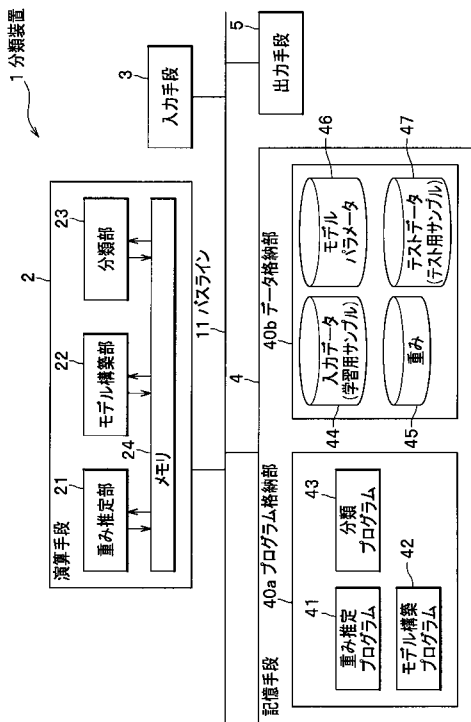
- 1 分類装置
- 2 演算手段
- 3 入力手段
- 4 記憶手段
- 5 出力手段
- 11 バスライン
- 21 重み推定部
- 22 モデル構築部
- 23 分類部
- 24 メモリ
- 40a プログラム格納部
- 41 重み推定プログラム
- 42 モデル構築プログラム
- 43 分類プログラム
- 40b データ格納部
- 44 入力データ

40

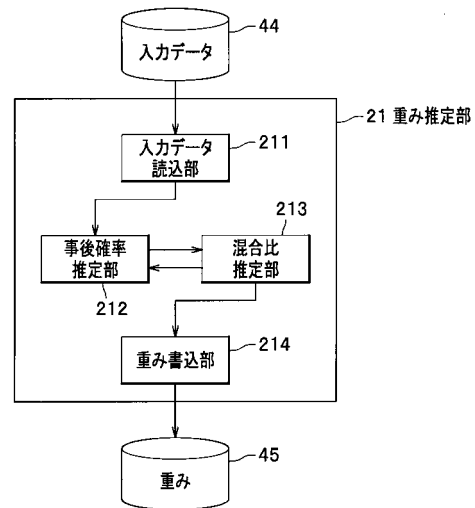
50

- 4 5 重み
- 4 6 モデルパラメータ
- 4 7 テストデータ
- 2 1 1 入力データ読込部
- 2 1 2 事後確率推定部
- 2 1 3 混合比推定部
- 2 1 4 重み書込部
- 2 2 1 入力データ読込部
- 2 2 2 重み読込部
- 2 2 3 モデルパラメータ推定部
- 2 2 4 モデルパラメータ書込部
- 2 3 1 テストデータ読込部
- 2 3 2 モデルパラメータ読込部
- 2 3 3 分類結果出力部

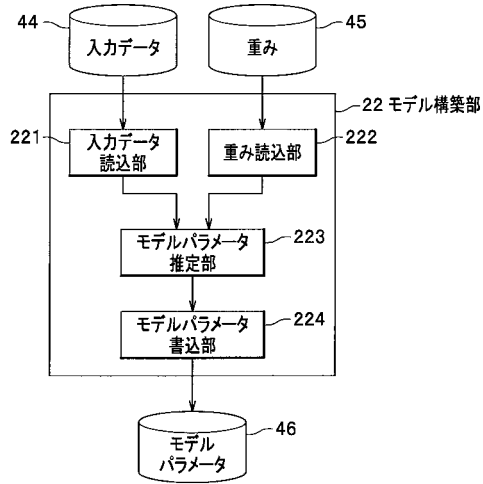
【図1】



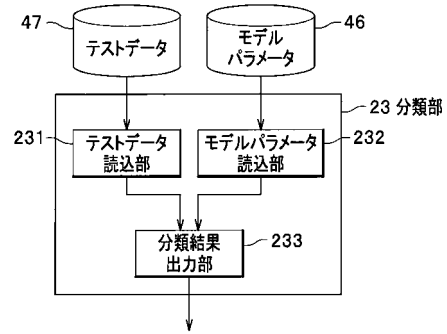
【図2】



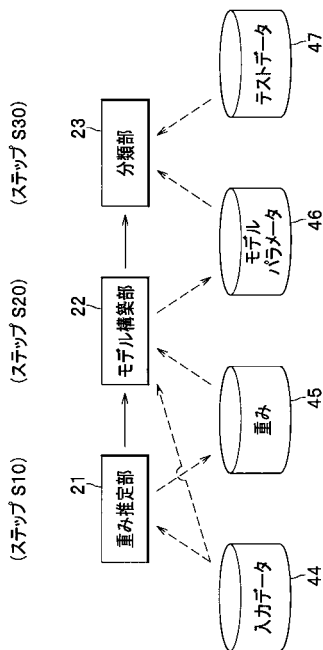
【 図 3 】



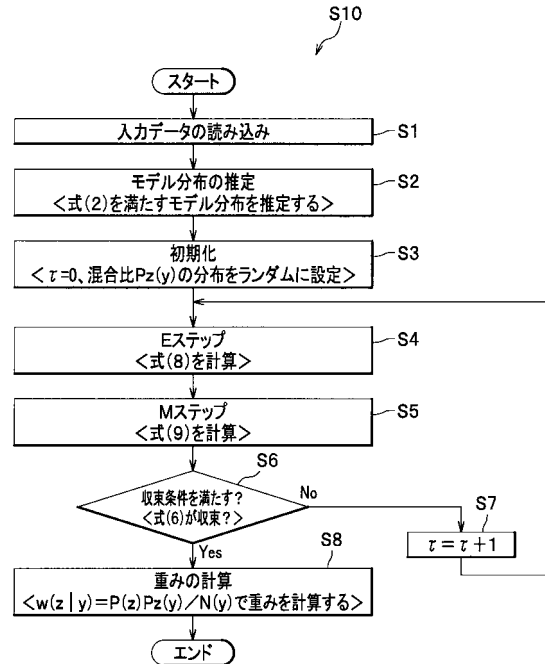
【 図 4 】



【 図 5 】

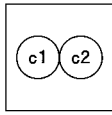


【 図 6 】



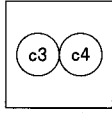
【 図 7 】

(a)



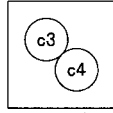
ターゲット

(b)



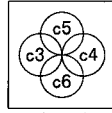
同一補助データ

(c)



相関補助データ

(d)



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

審査官 吉田 誠

(56)参考文献 国際公開第2004/084096(WO, A1)

特開2006-302269(JP, A)

特開2006-4103(JP, A)

特開2006-338263(JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F 17/30

G06N 5/04