自然言語処理 119－7
（1997．5．27）

線形結合モデルを用いたドキュメント分類

李 航　山西 健司

NEC C&C研究所

〒216 川崎市宮前区宮崎4-1-1

{lihang,yamanisi}@sbl.cl.nec.co.jp

本稿では、与えられたドキュメントをカテゴリに自動的に分類するための、新しいドキュメント分類方式を提案する。本方式は、カテゴリを線形結合モデルと呼ばれる確率モデルで表現し、ドキュメント分類問題を線形結合モデル間の統計的検定問題として扱うことを特徴とする。本方式が従来手法より高い分類精度を実現できることを実験的結果を用いて示す。

Document Classification Using a Finite Mixture Model

Hang LI　Kenji YAMANISHI

C&C Res. Labs., NEC

4-1-1 Miyazaki Miyamae-ku Kawasaki, 216 Japan

Abstract

We propose a new method of classifying documents into categories. We define for each category a *finite mixture model* based on *soft clustering* of words. We treat the problem of classifying documents as that of conducting statistical hypothesis testing over finite mixture models. Experimental results indicate that our method outperforms existing methods.

1 はじめに

本稿では、ドキュメント分類の一方式を提案する。ここでいうドキュメント分類とは、人間が予めドキュメントを幾つかのカテゴリ (例えば、「tennis,soccer」) に分類し、計算機がそれらの分類をもとに何らかの知識を獲得し、その知識のもとに新しく入って来たドキュメントをカテゴリの一つ (あるいはいくつか) に自動的に分類することを意味する。特に、本稿では、ドキュメントにおける単語の出現分布をもとにそれを分類する問題を取り上げる。

従来、ドキュメント分類の方式が幾つか提案され、それらの有効性も実証されているが (例： (ADW94; LR94; Lew92; LSCP96; SHP95; YC94))、分類の精度についてはまだまだ改良の余地を残している。最初に思いつく最も簡単な分類方法の一つは、カテゴリを単語ベースの確率分布で表現し、ドキュメント分類をそれらの確率分布間の統計的検定問題として捉える方法である。しかし、この方法では、確率分布を表現するパラメータの数が著しく大きく、それらを推定するための学習データが十分得られないといった、いわゆる「**データ・スパースネス**」(data sparseness) の問題が生じる。また、パラメータの数が大きいため、知識の管理や蓄積にも問題が出てくる。

これらの問題を解決するため、Guthrie らは単語をクラスタにまとめ (クラスタリングし)、カテゴリをクラスタ空間上の確率分布で表現し、ドキュメント分類をそれらの確率分布間の統計的検定問題として捉える方式を提案した (GWG94)。彼らは、単語をクラスタリングする際、一つの単語をたった一つのクラスタにしか割り当てず、同じクラスタに割り当てられた単語を全て一様に扱っていた。本稿では、このような単語のクラスタリング法を**ハード・クラスタリング**と呼ぶ。単語をクラスタにまとめるアプローチはデータ・スパースネスの問題には有効である。しかし、クラスタリングが「ハード」である場合、確率パラメータの数が極端に少く、よって、カテゴリの確率分布の間の差が十分表現できないといった問題がしばしば生ずる。

本稿では、ドキュメント分類に**ソフト・クラスタリング**[1]を用いることを提案する。ソフト・クラスタリングは一つの単語を一つ以上 (複数可) のクラスタに割り当て、しかも各クラスタは固有の単語の確率分布を有するといった構造をもつ。我々はカテゴリを各クラスタが有する単語の確率分布の線形結合によって表現する (これを**線形結合モデル**と呼ぶ)。それによってドキュメント分類の問題を線形結合モデル間の仮説検定問題として捉える。線形結合モデルの重み係数は過去のデータから最尤推定によって推定する。一般にこのような重み係数の最尤推定値は解析的に計算できないので、EM アルゴリズムを用いて効率的に近似する。

線形結合モデルを用いる方法は、単語ベースの方法とハード・クラスタリングによる方法のそれぞれの利点を保ちつつ、同時にそれぞれの問題点を克服している。実際、それは二つの方法を特殊な場合として含んでいる。本研究の実験結果によれば、線形結合モデルを用いる方法は二つの従来方法よりも高い分類精度を実現していることがわかった。

線形結合モデルを自然言語処理に用いることはすでにの他の文脈で提案されているが (例： (JM80; PTL93))、それをドキュメント分類に用いたのは本研究が初めてであると思われる。

[1]「ハード・クラスタリング」及び「ソフト・クラスタリング」という用語は (PTL93) から借用したものである。彼らは異なる文脈でそれらを用いていた。

2 従来研究

2.1 単語ベースの方法

カテゴリを単語ベースの確率分布によって表現することにより、ドキュメント分類問題を単語ベースの確率分布間の仮説検定問題として捉えることができる。本稿ではこの方法を**単語ベースの方法**と呼び、WBM (word-based method) と略記する。

今、W が単語の集合 (ボキャブラリ) を表すとする。WBM では、各カテゴリ c_i $(i = 1, \cdots, n)$ に対して、ヒストグラムで表現される単語ベースの確率分布 $P(w|c_i)$ を対応させる (このような確率分布における自由パラメータの数は $|W|-1$ である)。ただし、w は W に属する任意の単語を値とする確率変数である。WBM では新しく入力されたドキュメント

$$d = w_1, \cdots, w_N$$

を単語の系列とみなし、各単語が未知の確率分布に従って独立に生起していると仮定する。次に、入力されたドキュメントが各カテゴリの確率分布によって生成される確率

$$P(d|c_i) = P(w_1, \cdots, w_N|c_i) = \prod_{t=1}^{N} P(w_t|c_i) \quad (1)$$

を計算し $(i = 1, \cdots, n)$、この確率値の最も大きいカテゴリにドキュメントを分類する。ここで計算された、入力ドキュメントのカテゴリに対する確率は、カテゴリのそのドキュメントに対する尤度と見なすことができる。よって、上で計算される確率値を最大にするカテゴリにドキュメントを分類することは、最大の尤度をもつカテゴリにそれを分類することに等価である。そこで本稿では上記確率値を「尤度」という言葉で統一し、入力ドキュメント d に対するカテゴリ c_i の尤度を $L(d|c_i)$ と表す。

確率分布のパラメータは通常、学習データから推定しなければならない。単語ベースの確率分布を用いる場合には、パラメータの数が著しく大きく、それを推定するためのデータは十分に得られないといった、いわゆるデータ・スパースネスの問題が生じる。また、パラメータ数が大きいことは知識の管理や蓄積にも支障をきたす。

2.2 ハード・クラスタリングによる方法

以上の問題を解決するために、Guthrie らはハード・クラスタリングによる方法を提案した (GWG94)。本稿ではこの方法を HCM (hard-clustering-based method) と略記する。

今、$c_1, \cdots, c_n$ を与えられたカテゴリとする。HCM では、まず単語のハード・クラスタリングを行う。具体的には、(a) 単語集合 W に対して、**クラスタ** $k_1, \cdots, k_m$ を、$\cup_{j=1}^{m} k_j = W$ と $k_i \cap k_j = \emptyset$ $(i \neq j)$ をみたす W の部分集合列として用意する。(注： 一つの単語はたった一つのクラスタにしか割り当てられない。) (b) 同じクラスタに割り当てられた単語は全て一様に扱う。

HCM では、各カテゴリにクラスタの確率分布 $P(k_j|c_i)$ $(j = 1, \cdots, m)$ を対応させる。ドキュメント $d = w_1 \cdots w_N$ が入力されると、HCM はドキュメントを構成する各単語 w_t $(t = 1, \cdots, N)$ をその属するクラスタ k_t によって置き換えることによりドキュメントをクラスタの系列 $k_1 \cdots k_N$ と見なし、このクラスタ系列に関するカテゴリ c_i $(i = 1, \cdots, n)$ の尤度を以下のように計算する。

$$L(d|c_i) = L(k_1, \cdots, k_N|c_i) = \prod_{t=1}^{N} P(k_t|c_i) \quad (2)$$

そこで最大の尤度を達成するカテゴリに入力ドキュメン

表 1: 単語の出現度数

	racket	stroke	shot	goal	kick	ball
c_1	4	1	2	1	0	2
c_2	0	0	0	3	2	2

表 2: クラスタとそれに属する単語 ($L=5, M=5$)

k_1	racket, stroke, shot
k_2	kick
k_3	goal, ball

トを分類する。

HCM による分類の精度を上げるためには如何にクラスタを作成するかが重要な課題となる。以下に Guthrie らが提案した方法を示す。簡単のため、カテゴリが二つしかなく、それぞれが c_1=tennis と c_2=soccer であるとする。また、学習データにおいて、それぞれのカテゴリに分類されたドキュメントにおける単語の出現度数は表1に示す通りであるとする。

HCM は、まず三つのクラスタ k_1, k_2, k_3 を作成する。具体的には、L と M が与えられた正の整数であるとして、カテゴリ c_1 に最もよく現れる L 個の単語に含まれ、かつカテゴリ c_2 に最もよく現れる M 個の単語に含まれない単語を全てクラスタ k_1 に割り当て、カテゴリ c_2 に最もよく現れる L 個の単語に含まれ、かつカテゴリ c_1 に最もよく現れる M 個の単語に含まれない単語を全てクラスタ k_2 に割り当て、それ以外の単語を全て k_3 に割り当てる (表2を参照)。次に、各カテゴリにおける各クラスタの出現度数をカウントし (表3)、各カテゴリにおけるクラスタの確率分布を推定する (表4) [2]。今、図1のドキュメント d が入力されたとする。式 (2) にしたがって、このドキュメントに関する2つのカテゴリの尤度を計算した結果を表5に示す (表5は尤度の対数値を示している)。この場合、$\log_2 L(d|c_2)$ の値が $\log_2 L(d|c_1)$ の値よりも大きいので、HCM はドキュメント d をカテゴリ c_2 に分類する。

HCM は単語をクラスタにまとめることによって、推定しなければならないパラメータの数を削減し、データ・スパースネスの問題に対処することができる。また、パラメータ数を減らすことによって知識の蓄積も効率的になる。しかし、HCM には以下の二つの問題がある。

1. HCM では、**一つの単語を二つ以上のクラスタに割り当てることができない。**例えば、他に c_3=skiing というカテゴリがあって、そこには単語「ball」が全く現れないとする。この場合、「ball」が c_1 と c_2 の対応するクラスタ k_1 と k_2 の両方に属するとしたほうがよい。しかしながら、HCM では、「ball」をクラスタ k_1 と k_2 の両方に割り当てることができない。
2. HCM では、**同じクラスタに割り当てられた単語を一様に扱い、それらの出現頻度の差の情報を有効に利用できない。**例えば、HCM では、単語「racket」と「shot」は同じクラスタに割り当てられ、同様に扱われている。ところが、実際には c_1 の中で前者は後者より多く出現しており、それが存在することによってドキュメントが c_1 に分類される可能性がより強まることを意味している。しかしながら、HCM ではこのような同一クラスタ内の単語間の差異の情報を利用することができない。特に、ハード・クラスタリングにおける L と M の値が大きくなるにつれてこの問題が深刻になる。

さらに、パラメータの数の観点から見ても、HCM が用いる確率モデルは検定に必要な確率分布間の違いを十分表現できるほどパラメータ数が多いとはいえない。

[2] ここでは、拡張された Laplace 推定量 (GC90):

$$P(k_j|c_i) = (f(k_j|c_i) + 0.5)/(f(c_i) + 0.5 \times m)$$

を用いて確率値を推定する。ただし、$f(k_j|c_i)$ はカテゴリ c_i におけるクラスタ k_j の出現度数であり、$f(c_i)$ はカテゴリ c_i におけるクラスタの総出現度数であり、m はクラスタの数である。

表 3: クラスタの出現度数

	k_1	k_2	k_3
c_1	7	0	3
c_2	0	2	5

表 4: クラスタの確率分布

	k_1	k_2	k_3
c_1	0.65	0.04	0.30
c_2	0.06	0.29	0.65

d = kick, goal, goal, ball

図 1: ドキュメントの例

3 線形結合モデル

本節では、単語のソフト・クラスタリングに基づくドキュメント分類方式の数学的モデルを与える。

今、$c_1, \cdots, c_n$ はカテゴリを表すとする。先ず、ソフト・クラスタリングを次のステップ (a) 及び (b) で行う：(a) W を単語の集合 (ボキャブラリ) とし、クラスタ $k_1, \cdots, k_m$ を、$\cup_{j=1}^{m} k_j = W$ をみたす W の部分集合列として用意する。(注：$k_i \cap k_j = \emptyset$ $(i \neq j)$ は必ずしも成り立たない。すなわち、一つの単語は複数のクラスタに属し得る。) (b) 各クラスタ k_j $(j = 1, \cdots, m)$ に対し、このクラスタ上の単語の確率分布 $Q(w|k_j)$ $(\sum_{w \in k_j} Q(w|k_j) = 1)$ を定め、これを用いて W 上の確率分布 $P(w|k_j)$ を次式で定義する。

$$P(w|k_j) = \begin{cases} Q(w|k_j); & w \in k_j \\ 0; & w \notin k_j \end{cases}$$

ここに、w は W に値をとる確率変数である。

次に各カテゴリ c_i $(i = 1, \cdots, n)$ に対してクラスタの確率分布 $P(k_j|c_i)$ を定め、これを用いて単語分布 $P(w|k_j)$ の線形結合を以下のようにして定める。

$$P(w|c_i) = \sum_{j=1}^{m} P(k_j|c_i) \times P(w|k_j) \qquad (3)$$

このようなモデルは**線形結合モデル** (あるいは**有限混合モデル**)(finite mixture model) と呼ばれる (例：(EH81) を参照)。また、$P(k_j|c_i)$ は重み係数と呼ばれる。

表 5: 対数尤度の計算

$\log_2 L(d|c_1) = 1 \times \log_2 .04 + 3 \times \log_2 .30 = -9.85$
$\log_2 L(d|c_2) = 1 \times \log_2 .29 + 3 \times \log_2 .65 = -3.65$

我々はドキュメント分類の問題を線形結合モデル間の尤度比検定の問題として扱う。これを以下に詳しく述べよう。まず、ドキュメント

$$d = w_1, \cdots, w_N$$

を単語の系列とみなし、各単語 $w_t(t = 1, \cdots, N)$ は未知の確率分布に従って独立に生起していると仮定する。我々は線形結合モデル $P(w|c_i)(i = 1, \cdots, n)$ の中のどれが最もこの分布に近いかをドキュメントを観測して判断する。そのために、ドキュメントに関する各カテゴリの尤度を次式で計算する。

$$\begin{aligned} L(d|c_i) &= L(w_1, \cdots, w_N|c_i) = \prod_{t=1}^{N} P(w_t|c_i) \\ &= \prod_{t=1}^{N} \left(\sum_{j=1}^{m} P(k_j|c_i) \times P(w_t|k_j) \right) \end{aligned} \quad (4)$$

そこで最大の尤度を達成するカテゴリにドキュメントを分類する。以下、我々はこの分類方式を FMM と略記する。

FMM は WBM 及び HCM を特殊な場合として含む。実際、一つの単語がたった一つのクラスタのみに割り当てられ、各クラスタにおける単語の分布 $P(w|k_j)$ が

$$P(w|k_j) = \begin{cases} \frac{1}{|k_j|}; & w \in k_j \\ 0; & w \notin k_j \end{cases}$$

($|k_j|$ は k_j に属する元の数を表す) として与えられる特殊な場合には、FMM と HCM のクラスタリングは同じになる。この場合、各カテゴリ c_i の尤度は次式で計算される。

$$\begin{aligned} L(d|c_i) &= \prod_{t=1}^{N} (P(k_t|c_i) \times P(w_t|k_t)) \\ &= \prod_{t=1}^{N} P(k_t|c_i) \times \prod_{t=1}^{N} P(w_t|k_t) \end{aligned}$$

ここに、k_t は w_t の属するクラスタである。ここで $P(w_t|k_t)$ はカテゴリに依存しないので無視することができて、結局、仮説検定の結果は第 1 項の $\prod_{t=1}^{N} P(k_t|c_i)$ のみに依存することになる。よって、この場合の検定結果は HCM のそれと同じになる (式 (2) を参照)。

また、$m = n$ かつ、各 j について $P(w|k_j) = P(w|c_j)$ であり (この場合、確率分布 $P(w|k_j)$ は $|W|$ 個のパラメータを有する)、$P(k_j|c_i)$ が

$$P(k_j|c_i) = \begin{cases} 1; & i = j \\ 0; & i \neq j \end{cases}$$

で与えられる特殊な場合には、尤度の計算は式 (1) と一致する。したがって、この場合は FMM は WBM と同じ検定結果を導く。

4 FMM における推定と検定

本節では FMM の実際のインプリメントにあたっての詳細を述べる。

4.1 クラスタの作成

FMM で分類精度を上げるには、与えられた単語集合上のクラスタを如何に作成するかが鍵となる。以下に我々が用いた、クラスタ作成のためのヒューリスティクスの 1 例を示そう。

表 6: クラスタとそれに属する単語

k_1	racket, stroke, shot, ball
k_2	kick, goal, ball

今、クラスタの数はカテゴリの数と等しいとし ($m = n$)、[3] 各々のカテゴリ c_i には最も関連の深いクラスタが一つあるとし、これを k_i と書く ($i = 1, \cdots, n$)。以下、k_i を c_i に関連したクラスタと呼ぶ。

学習データとして、分類されたドキュメントがすでにあるとする。次に個々の単語を、それらが最も頻繁に現れるカテゴリに関連したクラスタに割り当てる。具体的には以下の方法を実行する。γ ($0 \leq \gamma < 1$) を与えられたしきい値とし、もし、

$$f(w|c_i)/f(w) > \gamma$$

が成立するならば、単語 w を c_i に関連したクラスタ k_i に割り当てる。ここに、$f(w|c_i)$ はカテゴリ c_i に現れる単語 w の (学習データにおける) 出現度数を表し、$f(w)$ は w の総出現度数を表す。

例として、表 1 のデータを用いて 2 つのクラスタ k_1 および k_2 を、それぞれカテゴリ c_1 と c_2 に関連させて作るとしよう。仮に $\gamma = 0.4$ とすると、カテゴリ c_2 における「goal」の相対頻度は 0.75 であり、c_1 におけるそれは 0.25 であるから、「goal」は k_2 のみに割り当てられることになる。この方法によりどのクラスタにも割り当てられない単語は、特定のカテゴリに固有のものではないと見なして無視する。(例えば、$\gamma \geq 0.5$ にすると、「ball」はどのクラスタにも割り当てられない。) これによって分類の効率を上げ、分類精度を高めることが期待できる。表 6にクラスタ作成の結果を示す。

4.2 $P(w|k_j)$ の推定

次に $P(w|k_j)$ を推定するのに、先ずクラスタ中の単語の出現度数を考える。もし、一つの単語が一つのクラスタのみに割り当てられるのならば、その単語の全体における総出現度数をクラスタ内の出現度数と見なすことができる。例えば、「goal」は k_2 のみに割り当てられているので、全カテゴリ上での「goal」の出現度数をそのまま k_2 内での「goal」の出現度数と見なすことができる。ある単語が複数の異なるクラスタに割り当てられている場合には、その単語の総出現度数をそれらのクラスタに、関連したカテゴリにおける度数比に応じて分配する。例えば、「ball」は k_1 と k_2 の 2 つのクラスタに割り当てられているので、その総出現度数を「ball」が c_1 と c_2 に現れた出現頻度に比例させて (1:1)、k_1 と k_2 に分配する。このようにして得られた、各クラスタに分配された単語の度数を表 7に示す。

表 7: 分配された単語の度数

	racket	stroke	shot	goal	kick	ball
k_1	4	1	2	0	0	2
k_2	0	0	0	4	2	2

次に各クラスタ中に現れる単語の確率を最尤推定によって計算する。表 8に計算結果を示す。

表 8: 単語の確率分布

	racket	stroke	shot	goal	kick	ball
k_1	0.44	0.11	0.22	0	0	0.22
k_2	0	0	0	0.50	0.25	0.25

表 9: クラスタの確率分布

	k_1	k_2
c_1	0.86	0.14
c_2	0.04	0.96

[3] 一般には $m \geq n$ が成立すると仮定してよい。

表 10: 対数尤度の計算

$\log_2 L(d\|c_1) = \log_2(.14 \times .25) + 2 \times \log_2(.14 \times .50) + \log_2(.86 \times .22 + .14 \times .25) = -14.67$
$\log_2 L(d\|c_2) = \log_2(.96 \times .25) + 2 \times \log_2(.96 \times .50) + \log_2(.04 \times .22 + .96 \times .25) = -6.18$

4.3 $P(k_j|c_i)$ の推定

次に $P(k_j|c_i)$ の推定方法について述べる。数理統計学でよく用いられる推定方法としては最尤推定法とベイズ推定法があるが、それらを線形結合モデルの重み係数の推定に直接適用しようとすると解析的困難を伴う。しかしながら、EM アルゴリズム (DLR77) を用いると $P(k_j|c_i)$ の最尤推定値の近似値を効率良く求めることができる。そこで、我々は (HSSW95) による EM アルゴリズムの拡張版を用いて $P(k_j|c_i)$ の最尤推定値を近似的に求める。このアルゴリズムを以下に詳しく示そう。(注：マルコフ・チェイン・モンテカルロ法に基づいて $P(k_j|c_i)$ のベイズ推定値の近似値を効率良く計算することもできる (例：(TW87; Yam96) を参照)。)

記法上簡単のために、固定した i に対して、$P(k_j|c_i)$ を θ_j と記し、$P(w|k_j)$ を $P_j(w)$ と記す。この時、$\theta = (\theta_1, \cdots, \theta_m)$ として、式 (3) の線形結合モデルは

$$P(w|\theta) = \sum_{j=1}^{m} \theta_j \times P_j(w)$$

と書くことができる。与えられた学習データ $w_1 \cdots w_N$ に対して、θ の最尤推定値は

$$L(\theta) = \frac{1}{N} \sum_{t=1}^{N} \log \left(\sum_{j=1}^{m} \theta_j P_j(w_t) \right)$$

を最大化する値 $\hat{\theta}$ として定義される。しかしながら、$\hat{\theta}$ を解析的に求めることは困難である。

EM アルゴリズムは先ず θ の初期値 $\theta^{(0)}$ を任意に設定し、θ の値を次の反復過程によって逐次的に更新して行く：

s を与えられた正の整数とせよ。l 番目の反復では、$\theta^{(l)} = (\theta_1^{(l)}, \cdots, \theta_m^{(l)})$ を次式で計算する $(l = 1, \cdots, s)$。

$$\theta_j^{(l)} = \theta_j^{(l-1)} \left(\eta(\nabla L(\theta^{(l-1)})_j - 1) + 1 \right)$$

$(j = 1, \cdots, m)$。ここに $\eta > 0$ であり、特に $\eta = 1$ の場合は標準の EM アルゴリズムに一致する。$\nabla L(\theta)$ は

$$\nabla L(\theta) = (\partial L/\partial\theta_1, \cdots, \partial L/\partial\theta_m)$$

を表す。上の過程を s 回反復した後、EM アルゴリズムは $\theta^{(s)} = (\theta_1^{(s)}, \cdots, \theta_m^{(s)})$ を最尤推定値 $\hat{\theta}$ の近似値として出力する。EM アルゴリズムは与えられた尤度の極大値に収束することが知られている (DLR77)。

表 1の例に対して上記手法により $P(k_j|c_i)$ を推定した結果を表 9に示す。

4.4 検定

表 1の例に対して式 (4) にしたがって図 1のドキュメントに関する 2 つのカテゴリの尤度を計算した結果を表 10に示す (表 10は尤度の対数値を示している)。この場合、$\log_2 L(d|c_2)$ の値が $\log_2 L(d|c_1)$ の値よりも大きいので、FMM はこのドキュメントを c_2 に分類する。

5 FMM の利点

ドキュメント分類に確率的アプローチをとる際に最も重要な問題は、カテゴリの表現形式としてどのような確率モデルを採用するか？ということである。それはさらに、

(1) カテゴリの表現としてモデル自体がふさわしいものかどうか？

(2) パラメータ数という観点から適切な複雑さをもっているか？

という 2 つの問題に帰着される。そして上の意味でのモデルの善し悪しが分類結果に大きく作用するのである。

我々が提案する FMM は先ず (1) の問題をクリアしている。すなわち、線形結合モデルはカテゴリの表現形式として十分適している。これを言語学的な観点から論じよう。クラスタは言語学における「トピック」に対応し、クラスタに割り当てられた単語はそのトピックに関連していると解釈することができる。一般に、ドキュメントはある一つのトピックに集中するが、しばしば別のトピックにしばらく移行する。そしてドキュメントが一つのトピックに集中しているときは、そのトピックに関連の深い単語を多く用いる傾向がある。例えば、「tennis」というカテゴリにあるドキュメントは「tennis」というトピックについて論じることが多く、そのとき「tennis」に関連の深い単語を多く用いるが、たまには「soccer」のトピックに少しだけ移行する、すなわち、「soccer」に関連の深い言葉を用いるようになる。人間はこのようなドキュメントに現れた単語の系列をフォローして、その系列にいくつかのトピックを関連づけ、それらのトピックがどのような分布をなすかという視点からドキュメントを分類することができるのである。したがって、線形結合モデルはまさに以上の過程を確率的に解釈してモデル化したものであると見なすことができる。

表 11: パラメータ数の比較

WBM	$O(n \cdot \|W\|)$
HCM	$O(n \cdot m)$
FMM	$O(\|k\| + n \cdot m)$

次に FMM は (2) の問題をもクリアしていることを示そう。すなわち、FMM はパラメータ数という観点からも適切な複雑さをもっているのである。表 11は FMM、HCM、及び WBM のパラメータ数を比較したものである。ここに $|W|$ は単語集合の大きさを、$|k|$ はクラスタの大きさの総和を (すなわち、$|k| = \sum_{j=1}^{m} |k_j|$)、$n$ はカテゴリの総数を、m はクラスタの総数を表す。FMM のパラメータ数は実際には WBM のそれよりも極めて小さい。なぜなら、例えば FMM で $\gamma \geq 0.5$ として 4 節のクラスタリング法を用いた場合には、$|k|$ の値は $|W|$ の値よりも実際にはずっと小さくなるからである。結果的に FMM は推定すべきパラメータの数が WBM に比べて少くて済むので、データ・スパースネスの問題にうまく対処できる。さらには WBM に比べて知識の管理や蓄積のためのメモリの節約が可能となる。一方、FMM のパラメータ数は HCM のそれに比べて大きく、カテゴリ間の差異を HCM よりも正確に表現できる。よって、2 節で述べた、HCM が抱える問題点を克服することも可能である。

FMM のもう一つの利点はドキュメント分類やドキュメント検索における Latent semantic analysis(DDF+90) と対比することにより明らかになる。Latent semantic

analysis では以下の問題が解決できると主張されている。

同義語問題 「stroke」と「shot」のような同義語を如何にグループ化するかといった問題。個々の単語としてはカテゴリの中には滅多に現れないが、グループとして見なした時はそのカテゴリを特徴づけるといった場合がある。

語義曖昧性問題 単語「ball」がドキュメントに現れた場合、これが「tennis ball」のことなのか「soccer ball」のことなのかをどのようにして判断したらよいかといった問題。

依存性問題 あるカテゴリの中では「kick」と「goal」のように互いに依存した単語をどのように結合して扱うかといった問題。

表6から明らかなように、FMM でも以上の問題は全て解決できる。

6 実験結果

本節では、FMM、HCM、WBM を実際にインプリメントしてドキュメント分類実験を行った結果について述べる。

データ・セット1として Reuters-21578 Distribution1.0 のサブセットを用いた。[4]我々はその中から九つの重なりのあるカテゴリを選択した。(注：一つのドキュメントが複数のカテゴリに属し得る。) さらに、Lewis Split に従い、この九つのカテゴリに属するドキュメントをトレーニング・データとテスト・データに分けた。表12と13にその詳細を示す。実験では、単語の原型への変換は行わず、かつ、不用語辞書 (stop words)[5]は使用しなかった。

次に、データ・セット1に対して、FMM、HCM、WBM を適用し、ドキュメントの二値分類を行った。また、コサイン類似度に基づいた方式 (以下 COS と略す)[6]をインプリメントし、二値分類を行った。具体的には、トレーニング・データを用いて各カテゴリ (例えば、「corn」) における分布を学習し、またそれぞれのカテゴリと排反するカテゴリ (例えば、「corn」でないカテゴリ) における分布を学習し、テスト・データにあるドキュメントがそれぞれのカテゴリに属するかそうでないかを判断した。FMM を適用する際、4節で提案したクラスタの作成法を用いて、しきい値 γ の値を 0,0.4,0.5,0.7 に選んだ。HCM を適用する際、同様にクラスタを作成し、γ の値を 0.5,0.7, 0.9,0.95 にした。(注： HCM に対しては γ の値を 0.5 より小さくすることができない。)

表 12: データ・セット1

トレーニング用ドキュメント数	707
テスト用ドキュメント数	228
単語の種類数	10902
ドキュメントにおける平均単語数	310.6

データ・セット2として Reuters-21578 を用いた。さらに Lewis Split に従い、このコーパスにあるドキュメントをトレーニング・データとテスト・データに分けた。表14はその詳細を示す。我々はさしあたり図15に示すカテゴリだけに対して実験を行った。実験では単語の原型への変換は行わず、かつ、不用語辞書は使用しなかった。データ・セット2に対して FMM、HCM、WBM、COS を適用し、ドキュメントの二値分類を行った。FMM を適用する際、4節で提案したクラスタ作成法を用い、しきい値 γ の値を 0, 0.4, 0.5, 0.7 に選んだ。HCM を適用する際も同様にクラスタを作成し、γ の値を 0.5,0.7, 0.9,0.95 にした。

表 13: データ・セット1におけるカテゴリ

wheat,corn,oilseed,sugar,coffee
soybean,cocoa,rice,cotton

表 14: データ・セット2

トレーニング用ドキュメント数	13625
テスト用ドキュメント数	6188
単語の種類数	50301
ドキュメントにおける平均単語数	181.3

表 15: データ・セット2におけるカテゴリ

earn,acq,crude,money-fx,grain
interest,trade,ship,wheat,corn

二つのデータ・セットに対して「マイクロ平均」による「適合率 (precision)」と「再現率 (recall)」[7]で評価した。

FMM、HCM、WBM を適用した時、標準的な尤度比検定ではなく、以下のヒューリスティクスを用いて検定を行った。簡単のため、カテゴリが c_1, c_2 の二つだけあるとする。ϵ を 0 以上の実数とし、入力ドキュメント d を以下のように分類する:

$$\begin{array}{ll} \frac{1}{N}(\log L(d|c_1) - \log L(d|c_2)) > \epsilon; & d \to c_1 \\ \frac{1}{N}(\log L(d|c_2) - \log L(d|c_1)) \geq \epsilon; & d \to c_2 \\ \text{その他;} & d \text{ を分類しない} \end{array}$$

ここで、N は d に含まれる単語の数である。[8]

図2はデータ・セット1における結果を、図3はデータ・セット2における結果を示す。これらのグラフで、FMM と HCM の後にある数値はクラスタ作成時の γ の値を表す。

各方法を評価する時、グラフ全体だけでなく、一つの値で評価したほうが明確である場合がある。ここでは、我々は評価値として break-even 点を用いた。break-even 点とは適合率と再現率が等しくなる点であり、その値が大きいほど分類の精度が良いといえる。表16はデータ・セット1に対する break-even 点を示し、表17はデータ・セット2に対する break-even 点を示す。データ・セット1に対しては FMM0 が、データ・セット2に対しては FMM0.5 がそれぞれ、break-even 点の最大値を達成していることがわかった。

次に、以下の三つの問題を考察した。

[4]このデータは http://www.research.att.com/lewis から入手できる。

[5]不要語辞書とは、英語の前置詞や冠詞等ドキュメント分類に必要でない単語を登録した辞書のことである。

[6]コサイン類似度に基づいた方式では、カテゴリとドキュメントを単語の頻度ベクトルとみなし、頻度ベクトル間のコサイン値を類似度とし、ドキュメント分類を行う。

[7]「マイクロ平均」(LR94) では、「適合率」とは分類できたドキュメント中の正しく分類できたドキュメントの占める割合で、「再現率」とは全体のドキュメント中の分類できたドキュメントの占める割合である。

[8]クラスタリングの過程で無視された単語はドキュメントの大きさを計る際には計上されない。

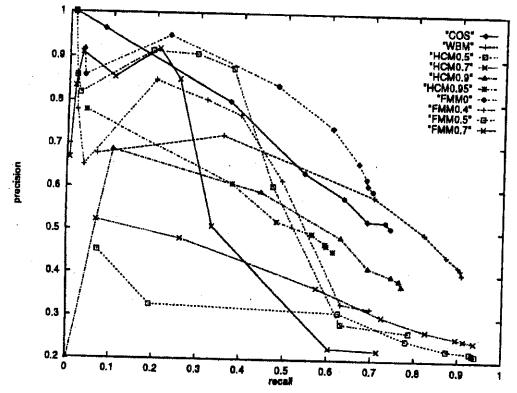

図 2: データ・セット 1 の適合率 – 再現率 曲線

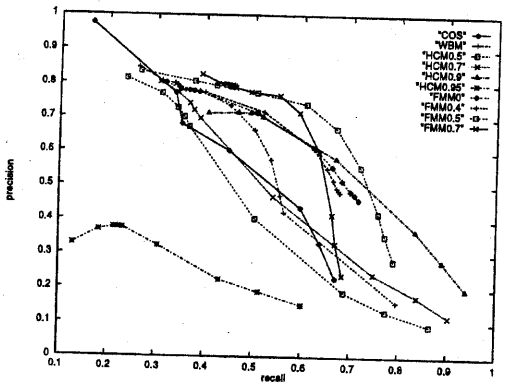

図 3: データ・セット 2 の適合率 – 再現率 曲線

(1) クラスタリングによる方法 (FMM) が単語ベースの方法 (WBM) より良い分類結果をもたらすか？

(2) ソフト・クラスタリングによる方法 (FMM) がハード・クラスタリングによる方法 (HCM) より良い分類結果をもたらすか？

(3) 4.1 節のクラスタ作成方法で、しきい値 γ が 0 に近づくにつれて、FMM が WBM と同じような振舞をし、クラスタリングの効果がほぼなくなる。これは、(a) 単語はすべてのクラスタに分類されること、(b) 各クラスタにおける単語の確率分布がそれぞれのクラスタに関連するカテゴリにおける単語の確率分布に近づくこと、(c) 各カテゴリの尤度が WBM のそれらに近づくこと、に起因する (2 節のケース (2) を参照)。最適なクラスタを作成することは極めて困難であるが、4.1 節のクラスタ作成法を用いれば、γ を 0 に近づけることによって少なくとも WBM と同じ分類精度を達成することができる。そこで γ を 0 に設定した場合、FMM が WBM より良い分類結果をもたらすかという問題がある。

以上の三項目について我々は以下の結論に達した。

(1) $\gamma \gg 0$ の時、つまり、クラスタリングを行う時、データ・セット 1 に対しては FMM は WBM より良い分類結果を出していない。しかし、データ・セット 2 に対しては WBM より良い結果を出している。

分類の結果を各カテゴリベースで評価した場合、データ・セット 1 における九つのカテゴリの内三つのカテゴリに対して FMM0.5 が最も良い分類結果を出している。データ・セット 2 における十個のカテゴリの内二つのカテゴリに対して FMM0.5 が最も良い分類結果を出している (図 4はデータ・セット 1 におけるカテゴリ「corn」の分類結果を示している。ただし、それぞれの方式の最も良い結果のみを示している。図 5はデータ・セット 2 におけるカテゴリ「grain」の分類結果を示す)。これらの結果から、4.1 節のクラスタ作成法に基づく FMM が WBM より良い分類結果を生むことがあることがわかった。しかし、すべての場合について、クラスタリングした方が単語ベースの方法よりも高い分類精度を達成するわけではなかった。よって、より優れたクラスタの作成法を開発することが今後の課題である。

表 16: データ・セット 1 に対する break-even 点

COS	0.60
WBM	0.62
HCM0.5	0.32
HCM0.7	0.42
HCM0.9	0.54
HCM0.95	0.51
FMM0	0.66
FMM0.4	0.54
FMM0.5	0.52
FMM0.7	0.42

表 17: データ・セット 1 に対する break-even 点

COS	0.52
WBM	0.62
HCM0.5	0.47
HCM0.7	0.51
HCM0.9	0.55
HCM0.95	0.31
FMM0	0.62
FMM0.4	0.54
FMM0.5	0.67
FMM0.7	0.62

(2) $\gamma \gg 0$ の時、つまり、クラスタリングを行う時、FMM は常に HCM より良い分類結果を出している。

(3) $\gamma = 0$ の時、FMM はデータ・セット 1 に対して WBM より良い分類結果を出し、データ・セット 2 に対して WBM と同等な分類結果を出している。

以上により、**FMM は HCM より常に高い分類精度を実現でき、WBM より同等以上の分類精度を実現できる**という結論が導かれる。

いずれのデータ・セットに対しても、FMM が COS より良い結果を出している。これは、単語の分布に基づいてドキュメント分類を行う場合、確率的なアプローチはコサイン類似度を用いるアプローチより効果的であることを意味している。

我々はまだ Reuters データの全カテゴリに対する実験を行っていないが、データ・セット 2 に対する FMM の結果が Reuters データの全カテゴリに対する他のアプローチによる結果 (LR94) と同等に良いことがわかった。

7 おわりに

本研究の結論は以下の通りである。

1. 本研究で提案した線形結合モデルを用いたドキュメント分類法は、単語ベースの方法とハード・クラスタリングによる方法の両者の利点を保ちつつ、同時

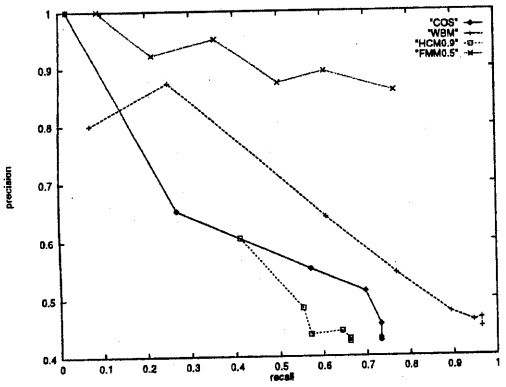

図 4: カテゴリ「corn」の適合率 - 再現率曲線

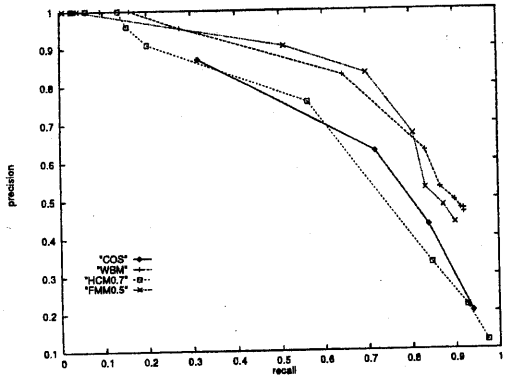

図 5: カテゴリ「grain」の適合率 - 再現率曲線

にそれらの問題点を克服しているものである。

2. 本研究の実験結果によれば、線形結合モデルを用いた方法はハード・クラスタリングによる方法より高い分類精度を実現することがわかった。
3. 本研究の実験結果によれば、線形結合モデルを用いた方法は単語ベースの方法と同等、あるいはそれ以上の分類精度を実現することがわかった。

今後の課題として以下が挙げられる。

1. Reuters 全データ、及び他のコーパス・データを用いて、さらに評価を進めること、
2. クラスタの新しい作成法を開発し、分類精度を向上させること。

本研究が提案した方法はドキュメントの分類だけでなく、その他の自然言語処理の問題にも適用できる。例えば、語義曖昧性解消の問題を考えてみよう。語義の曖昧な単語の周りの文脈をドキュメントと見なし、解消したい語義をカテゴリと見なして、本研究の方法を適用すれば語義曖昧性解消の問題に対処することもできる。

謝辞

本研究を支援して頂いている NEC C&C 研究所 システム基礎研究部長 藤田知之氏と同研究課長 安倍直樹氏に深謝いたします。

参考文献

Chidanand Apte, Fred Damerau, and Sholom M. Weiss. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst.*, 12(3):233–251, 1994.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *J. Am. Socie. Inf. Sci.*, 41(6):391–407, 1990.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Socie., Ser. B*, 39(1):1–38, 1977.

B. Everitt and D. Hand. *Finite Mixture Distributions*. London: Chapman and Hall, 1981.

Williams A. Gale and Kenth W. Church. Poor estimates of context are worse than none. *Proc. DARPA Spee. Nat. Lang. Worksh.*, pages 283–287, 1990.

Louise Guthrie, Elbert Walker, and Joe Guthrie. Document classification by machine: Theory and practice. *Proc. COLING'94*, pages 1059–1063, 1994.

D. Helmbold, R. Schapire, Y. Singer, and M. Warmuth. A comparison of new and old algorithm for a mixture estimation problem. *Proc. COLT'95*, pages 61–68, 1995.

F. Jelinek and R.I. Mercer. Interpolated estimation of markov source parameters from sparse data. *Proc. Worksh. Patt. Recog. in Prac.*, pages 381–402, 1980.

David D. Lewis. Feature selection and feature extraction for text. *Proc. Spee. Nat. Lang. Worksh.*, 1992.

David D. Lewis and Marc Ringuette. A comparison of two learning algorithms for test categorization. *Proc. 3rd Ann. Symp. Doc. Ana. Inf. Retri.*, pages 81–93, 1994.

David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. Training algorithms for linear text classifiers. *Proc. SIGIR'96*, 1996.

Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. *Proc. ACL'93*, pages 183–190, 1993.

Hinrich Schutze, David A. Hull, and Jan O. Pedersen. A comparison of classifiers and document representations for the routing problem. *Proc. SIGIR'95*, 1995.

Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.*, 82(398):528–540, 1987.

Kenji Yamanishi. A randomized approximation of the mdl for stochastic models with hidden variables. *Proc. COLT'96*, pages 99–109, 1996.

Yiming Yang and Christoper G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Trans. Inf. Syst.*, 12(3):252–277, 1994.