社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－NL－158 (8)
2003／11／6

カーネル関数によるカテゴリ構造のモデル化

高村 大也[†] 松本 裕治[††] 山田寛泰[‡]

カテゴリの事後確率を基にして構築される TOP (Tangent vector Of the Posterior log-odds) カーネルの一つの特殊形を提案し, それを文書分類に適用する. 自然言語処理の, 文書分類を含むいくつかの二値分類タスクでは, 一般に負例 (正例のカテゴリに属さないもの) は正例に比べて多く, またその性質も多様であることが多い. この事実を鑑みて, 我々は負例集合の確率モデルを, 正例と異なるカテゴリの混合モデルとみなすことにより, TOP カーネルを構成する. 我々のモデル化では, さらに各カテゴリの事後確率が一次元ガウス分布を用いて構成され, これによりカーネル関数の計算が効率的に行われる. 我々の実験では, 提案したカーネルは, 線形カーネルや PLSI (Probabilistic Latent Semantic Indexing) に基づいたフィッシャー・カーネルより高い精度を示した.

キーワード: 文書分類, TOP カーネル

Modeling Category Structures with a Kernel Function

Hiroya Takamura Yuji Matsumoto Hiroyasu Yamada

We propose one type of TOP (Tangent vector Of the Posterior log-odds) kernels and apply it to text categorization. In a number of categorization tasks including text categorization, negative examples are usually more common than positive examples and there may be several different types of negative examples. Therefore, we regard the probabilistic model of negative examples as a mixture of several models respectively corresponding to given categories. Since each model of ours is expressed using a one-dimensional Gaussian-type function, our kernel has an advantage in computational time. In our experiments, our kernel, combined with SVMs, outperformed the linear kernel and the Fisher kernel based on the Probabilistic Latent Semantic Indexing model.

Keywords: text categorization, TOP kernel

1 序論

近年, サポートベクターマシン (SVM) (Vapnik, 1998) が, その高い分類性能のため, 盛んに研究されている. SVM は, 二つの事例の類似度を与えるカーネル関数と組み合わせて用いられる. カーネル関数として使用されるのは, 事例に対応するベクトルの普通の内積が最も一般的である. また, 高次多項式カーネルや RBF カーネルなどもよく使用されるが, これらのカーネル関数においては事例の分布は考慮されていない.

しかし, 事例の確率分布を基にした新たなカーネルが提案されてきている: 一つはフィッシャー・カーネル (Jaakkola and Hassler, 1998) であり, もう一つは TOP (Tangent vector Of the Posterior log-odds) カーネル (Tsuda, 2002) である. フィッシャー・カーネルは事例の生成モデルを基にしており, TOP カーネルはカテゴリの事後確率, つまり事例が与えられたときのカテゴリの分布を基にしている. TOP カーネルは, 蛋白質の分類タスクなどにおいてフィッシャー・カーネルより高い分類精度を示しており (Tsuda, 2002), TOP カーネルの有効的な活用方法を探ることは重要な研究課題である. カテゴリの事後確率を決定すれば TOP カーネルも決定する. よって問題は, 高い分類性能を示し, か

[†]東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology, takamura@pi.titech.ac.jp
[††]奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology, matsu@is.aist-nara.ac.jp
[‡]北陸先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Japan Institute of Science and Technology, h-yamada@jaist.co.jp

つ効率的な計算が可能であるようなカーネル関数を決定する事後確率を選ぶことである．本稿ではそのような確率モデル，すなわちTOPカーネルを提案する．

我々の提案するカーネル関数を簡単に説明する．我々は，二値分類の際の負例集合に着目する．負例は一般的に正例よりも数が多く，その性質は多様であると考えられる．さらに，負例のカテゴリが与えられている場合も考えられる（例えば，“スポーツ”，“政治”，“経済”の3つのカテゴリから，“政治”に属する文書を抽出する場合を考えてみるとよい）．このような状況では，負例集合の確率モデルは，各カテゴリに対応する確率モデルの混合モデルとして表現できる．この事実を有効に利用する．多種の混合モデルが存在するが，我々は各カテゴリの分離平面を用いて計算されるモデルを提案する．もう少し詳しくいうと，各分離平面の垂線方向に一次元ガウス型関数を考え，それらを用いて事後確率モデルを構築する．このような確率モデルは，計算量の面で効率的なカーネル関数を導くことを後で示す．本稿では，このカーネルをHyperplane-based TOP (HP-TOP) カーネルと呼ぶことにする．

SVMを分類器として用いた文書分類の実験で，提案したカーネル関数は線形カーネルやProbabilistic Latent Semantic Indexing （PLSI）ベースのフィッシャー・カーネル (Hofmann, 2000) よりも良い精度を示した．

2 SVMとカーネル法

本節では，SVMとカーネル法について簡単に説明する．

事例とラベルのペアの集合 $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)\}$, $(\forall i,\ \mathbf{x}_i \in \mathbf{R}^d, y_i \in \{-1, 1\})$ が与えられているとする．SVMでは，マージン（分離平面とその平面に最も近い訓練事例との距離）が最大となるような分離平面 $(f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b)$ が構築される．

詳細は割愛するが，ラグランジュ法による変形の結果，最適分離平面は，実数 α_i^* $(\forall i)$ と b^* を用いて，

$$f(\mathbf{x}) = \sum_i \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} - b^* \tag{1}$$

と表される．事例は，内積の形でのみ現れることに注意されたい．

SVMは線形分類器であるが，非線形の分類を可能にするために，カーネル法 (Vapnik, 1998) と組み合わせて用いられることが多い．

カーネル法においては，式(1)の内積はより一般的な内積であるカーネル関数 $K(\mathbf{x}_i, \mathbf{x})$ に置き換えられる．この置き換えにより，非線形の分離が可能になる．多項式カーネル $(\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$ $(d \in \mathbf{N}_+)$ やRBFカーネル $\exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\}$ などがよく用いられる．

3 確率分布を基にしたカーネル関数

前節で紹介した多項式カーネルなどにおいては，事例の分布は考慮されていない．しかし，事例の分布は，分類においても非常に重要な情報源であると予想される．そのような予想の下に，データの生成モデルとSVMのような分類器を繋ぐ役割を果たす，新しいタイプのカーネル関数が提案されてきている：フィッシャー・カーネル (Jaakkola and Hassler, 1998) とTOPカーネル (Tsuda, 2002) である．ここではこれらのカーネル関数の説明をする．ただし，これらのカーネル関数は，確率モデルを与えたときの一般的な枠組となっており，確率モデルによって実際のカーネルの形は異なる．

3.1 フィッシャー・カーネル

データの確率的生成モデル $p(\mathbf{d}|\theta)$ が与えられているとしよう．ここで $\mathbf{d}$ は事例，θ はパラメータである．$\mathbf{d}$ のフィッシャー・スコアは，$\nabla_\theta \log p(\mathbf{d}|\theta)$ と定義される．モデル空間の幾何的構造を決めるフィッシャー情報量行列を $I(\theta)$ で表すと，推定値 $\hat{\theta}$ におけるフィッシャー・カーネルは次のようになる：

$$K(\mathbf{d}^1, \mathbf{d}^2) \quad = \quad (\nabla_\theta \log p(\mathbf{d}^1|\hat{\theta}))^t I^{-1}(\hat{\theta})(\nabla_\theta \log p(\mathbf{d}^2|\hat{\theta})). \tag{2}$$

フィッシャー・スコアは，大まかに述べると，その事例が訓練データとして加えられたときに推定されたモデルがどのように変化するかを定量的に示している．つまり，二つの事例に対するフィッシャー・カーネルは，その二事例のモデルに対する影響が類似しておりかつ大きい時に，その値が大きくなる (Tsuda, 2002).

3.2 TOP カーネル

データの確率モデルを基にして, TOP カーネルは線形分離平面（$\mathbf{w}\cdot\mathbf{f}_{\hat{\theta}}-b=0$）での分類に有用と考えられる素性から成るベクトル $\mathbf{f}_{\hat{\theta}}$ の内積として表される.

汎化誤差 $R(\mathbf{f}_{\hat{\theta}})$ と推定された事後確率 $\hat{P}(y=+1|\mathbf{x})$ のエラーの期待値との間には, $R(\mathbf{f}_{\hat{\theta}})-L^*\leq 2E_{\mathbf{X}}|\hat{P}(y=+1|\mathbf{x})-P(y=+1|\mathbf{x},\theta^*)|$ なる関係が成立することが知られている. ここで, L^* はベイズ誤差である. この不等式は, 事後確率のエラーの期待値の最小化が $R(\mathbf{f}_{\hat{\theta}})$ の最小化に繋がることを示している. ロジスティック関数 $F(t)=1/(1+\exp(-t))$ を用いて分離平面からの距離を確率値に写像することにして, $D(\mathbf{f}_{\hat{\theta}})$ を,

$$D(\mathbf{f}_{\hat{\theta}}) \;=\; \min_{\mathbf{w},b} E_{\mathbf{X}}|F(\mathbf{w}\cdot\mathbf{f}_{\hat{\theta}}-b)-P(y=+1|\mathbf{x},\theta^*)|, \tag{3}$$

と定義する. ここで θ^* はモデルの真のパラメータである.

TOP カーネルにおいては, $D(\mathbf{f}_{\hat{\theta}})$ を最小化しうる素性が使用される. 言い替えると, 我々はある $\mathbf{w}$ と b に対して次式を満たすような素性ベクトル $\mathbf{f}_{\hat{\theta}}$ を求めたい：

$$\forall\mathbf{x},\quad \mathbf{w}\cdot\mathbf{f}_{\hat{\theta}}(\mathbf{x})-b=F^{-1}(P(y=+1|\mathbf{x},\theta^*)). \tag{4}$$

そのような素性ベクトルを導出するため, まず関数 $v(\mathbf{x},\theta)$ を定義する：

$$\begin{aligned} v(\mathbf{x},\theta) &\equiv F^{-1}(P(y=+1|\mathbf{x},\theta)) \\ &= \log P(y=+1|\mathbf{x},\theta)-\log P(y=-1|\mathbf{x},\theta). \end{aligned} \tag{5}$$

$v(\mathbf{x},\theta^*)$ の $\hat{\theta}=(\hat{\theta}_1,\cdots,\hat{\theta}_p)$ の周りでの一次テイラー展開は,

$$v(\mathbf{x},\theta^*)\approx v(\mathbf{x},\hat{\theta})+\sum_i(\theta_i^*-\hat{\theta}_i)\partial v(\mathbf{x},\hat{\theta})/\partial\theta_i \tag{6}$$

と表される. もし, $\mathbf{f}_{\hat{\theta}}$ が, 次の形をしており, しかも $\mathbf{w}$ と b が次のように適切に選ばれるならば,

$$\mathbf{f}_{\hat{\theta}}(\mathbf{x})=(v(\mathbf{x},\hat{\theta}),\partial v(\mathbf{x},\hat{\theta})/\partial\theta_1,\cdots,\partial v(\mathbf{x},\hat{\theta})/\partial\theta_p), \tag{7}$$

$$\mathbf{w}=(1,\theta_1^*-\hat{\theta}_1,\cdots,\theta_p^*-\hat{\theta}_p),\quad b=0, \tag{8}$$

式 (4) は近似的に成立する.

このようにして, TOP カーネルは,

$$K(\mathbf{x}_1,\mathbf{x}_2)=\mathbf{f}_{\hat{\theta}}(\mathbf{x}_1)\cdot\mathbf{f}_{\hat{\theta}}(\mathbf{x}_2) \tag{9}$$

と定義される.

4 関連研究

ここでは, 確率分布を基にしたカーネル関数を文書分類に応用した例を挙げる. ホフマン (Hofmann, 2000) は, Probabilistic Latent Semantic Indexing (PLSI) モデル (Hofmann, 1999) の上に, フィッシャー・カーネルを構築し, それを文書分類に適用した.

PLSI では, 文書 $\mathbf{d}$ と単語 w の同時確率は,

$$P(\mathbf{d},w)=\sum_k P(z_k)P(\mathbf{d}|z_k)P(w|z_k), \tag{10}$$

と表される. ここで z_k は隠れ変数である.

EM アルゴリズムによるモデル推定の後, このモデルのフィッシャー・カーネルが計算される.

正規化された文書 $\mathbf{d}$ の平均対数尤度は,

$$l(\mathbf{d}) = \sum_j \hat{P}(w|\mathbf{d}) \log \sum_k P(w_j|z_k)P(z_k|\mathbf{d}), \quad \hat{P}(w|\mathbf{d}) = \frac{freq(w_j, \mathbf{d})}{\sum_m freq(w_m, \mathbf{d})}, \tag{11}$$

と計算される.

カーネル構築の際のパラメータとして, モデルの元々のパラメータの代わりに, *spherical parameterization* (Kass, 1997) が使用されている. $\rho_{jk} = 2\sqrt{P(w_j|z_k)}$, $\rho_k = 2\sqrt{P(z_k)}$ と定義すると,

$$\frac{\partial l(\mathbf{d})}{\partial \rho_{jk}} = \frac{\hat{P}(w_j|\mathbf{d})P(z_k|\mathbf{d}, w_j)}{\sqrt{P(w_j|z_k)}}, \quad \frac{\partial l(\mathbf{d})}{\partial \rho_k} \approx \frac{P(z_k|\mathbf{d})}{\sqrt{P(z_k)}}, \tag{12}$$

となる.

このようにして, PLSI モデルのフィッシャー・カーネルは,

$$\begin{aligned} K(\mathbf{d}^1, \mathbf{d}^2) &= \sum_k \frac{P(z_k|\mathbf{d}^1)P(z_k|\mathbf{d}^2)}{P(z_k)} \\ &\quad + \sum_j \hat{P}(w_j|\mathbf{d}^1)\hat{P}(w_j|\mathbf{d}^2) \sum_k \frac{P(z_k|\mathbf{d}^1, w_j)P(z_k|\mathbf{d}^2, w_j)}{P(w_j|z_k)}, \end{aligned} \tag{13}$$

と求められる. ここで,

$$P(z_k|\mathbf{d}, w_j) = \frac{P(z_k)P(\mathbf{d}|z_k)P(w_j|z_k)}{\sum_l P(z_l)P(\mathbf{d}|z_l)P(w_j|z_l)}$$

である.

式 (13) の第一項は, 隠れ変数の空間における類似度を表しており, 第二項は, 各単語の確率分布からの類似度への寄与を表している. 隠れ変数 z_k の数は, カーネル関数の値に影響を与えると思われる. ホフマンら (Hofmann, 2000) は, 異なる隠れ変数の数 (1 から 64) 全てについてカーネル関数を計算し, それらを足し合わせることで, 隠れ変数の数を決定するという問題を回避している.

ホフマンらは, PLSI のモデル推定に多量の未知データが使用できるとき, PLSI ベースのフィッシャー・カーネルは効果的であると結論している.

5 Hyperplane-based TOP カーネル

本節では, 我々の TOP カーネルについて説明する.

元々与えられた素性空間 ($\mathbf{d}$ が存在する空間) で SVM などの分類器を適用することにより, その素性空間における各カテゴリ c の分離平面を表すパラメータ $\mathbf{w}_c$ と b_c を得ることができる. これらのパラメータを使って, 提案する TOP カーネルを計算していく. これ以後, c は正例のカテゴリを, e は c でないカテゴリを, c' は c もしくは e を表すとする.

正例カテゴリの事後確率 $P_c(+1|\mathbf{d})$ と, 負例カテゴリの事後確率 $P_c(-1|\mathbf{d})$ は, 以下のように表されると仮定する:

$$P_c(+1|\mathbf{d}) = \frac{P(c)q(\mathbf{d}|c)}{\sum_{c'} P(c')q(\mathbf{d}|c')}, \quad P_c(-1|\mathbf{d}) = \frac{\sum_{e \neq c} P(e)q(\mathbf{d}|e)}{\sum_{c'} P(c')q(\mathbf{d}|c')}. \tag{14}$$

ここで, $q(\mathbf{d}|x) = (2\pi\sigma_x^2)^{-1/2} \exp\{-|(\mathbf{w}_x \cdot \mathbf{d} - b_x) - \mu_x|^2 / 2\sigma_x^2\}$ であり[1], μ_x, σ_x は確率変数 $(\mathbf{w}_x \cdot \mathbf{d} - b_x)$ の, それぞれ平均と分散である.

[1] $q(\mathbf{d}|x)$ はクラス x が与えられたときの事例 $\mathbf{d}$ の生成確率のような働きをしているが, その言明は正確ではない. なぜなら, $q(\mathbf{d}|x)$ は一次元関数であり, 一般的に多次元である元の素性空間では確率密度関数として適切でないからである (全体での積分は 1 にならない).

ここで, ガウス分布の自然パラメータに習って, $\theta_{x1}=\mu_x/\sigma_x^2, \theta_{x2}=-1/2\sigma_x^2$ とパラメータをセットする. これらのパラメータはこのモデルに関しては自然パラメータではないが, 簡単のため, θ_{x1}, θ_{x2}, $\mathbf{w}_x$, b_x, $P(x)$ という四種類のパラメータを使ってモデルを表すことにする.

この確率モデル (14) を基にして, 以下のように関数 $v(\mathbf{d}, \{\mathbf{w}_x, b_x, \theta_{x1}, \theta_{x2}\}_x)$ が計算できる (θ は $\{\mathbf{w}_x, b_x, \theta_{x1}, \theta_{x2}\}_x$) を表す):

$$
\begin{aligned}
& v(\mathbf{d}, \theta) \\
= \; & \log \frac{P(c)q(\mathbf{d}|c)}{\sum_{c'} P(c')q(\mathbf{d}|c')} - \log \frac{\sum_{e\neq c} P(e)q(\mathbf{d}|e)}{\sum_{c'} P(c')q(\mathbf{d}|c')} \\
= \; & \log P(c)q(\mathbf{d}|c) - \log \sum_{e\neq c} P(e)q(\mathbf{d}|e) \\
= \; & \log P(c) \exp\{\theta_{c1}(\mathbf{w}_c\cdot\mathbf{d}-b_c) + \theta_{c2}(\mathbf{w}_c\cdot\mathbf{d}-b_c)^2 + \frac{\theta_{c1}^2}{4\theta_{c2}} - \frac{1}{2}\log\frac{-\pi}{\theta_{c2}}\} \\
& - \log \sum_{e\neq c} P(e) \exp\{\theta_{e1}(\mathbf{w}_e\cdot\mathbf{d}-b_e) + \theta_{e2}(\mathbf{w}_e\cdot\mathbf{d}-b_e)^2 + \frac{\theta_{e1}^2}{4\theta_{e2}} - \frac{1}{2}\log\frac{-\pi}{\theta_{e2}}\}.
\end{aligned} \tag{15}
$$

この関数 $v(\mathbf{d}, \theta)$ の各パラメータに関する偏微分は次のように計算できる:

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial \theta_{c1}} = \mathbf{w}_c\cdot\mathbf{d} - b_c - \mu_c, \tag{16}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial \theta_{e1}} = -\frac{P(e)q(\mathbf{d}|e)}{\sum_{c'\neq c} P(c')q(\mathbf{d}|c')}(\mathbf{w}_e\cdot\mathbf{d} - b_e - \mu_e), \tag{17}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial \theta_{c2}} = (\mathbf{w}_c\cdot\mathbf{d} - b_c)^2 - \mu_c^2 - \sigma_c^2, \tag{18}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial \theta_{e2}} = -\frac{P(e)q(\mathbf{d}|e)}{\sum_{c'\neq c} P(c')q(\mathbf{d}|c')}\{(\mathbf{w}_e\cdot\mathbf{d} - b_e)^2 - \mu_e^2 - \sigma_e^2\}, \tag{19}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial w_{ci}} = \frac{\mu_c - (\mathbf{w}_c\cdot\mathbf{d} - b_c)}{\sigma_c^2} d_i, \tag{20}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial w_{ei}} = -\frac{P(e)q(\mathbf{d}|e)}{\sum_{c'\neq c} P(c')q(\mathbf{d}|c')}\frac{\mu_e - (\mathbf{w}_e\cdot\mathbf{d} - b_e)}{\sigma_e^2} d_i, \tag{21}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial b_c} = \frac{\mathbf{w}_c\cdot\mathbf{d} - b_c - \mu_c}{\sigma_c^2}, \tag{22}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{\partial b_e} = -\frac{P(e)q(\mathbf{d}|e)}{\sum_{c'\neq c} P(c')q(\mathbf{d}|c')}\frac{\mathbf{w}_e\cdot\mathbf{d} - b_e - \mu_e}{\sigma_e^2}, \tag{23}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{P(c)} = \frac{1}{P(c)}, \tag{24}
$$

$$
\frac{\partial v(\mathbf{d},\theta)}{P(e)} = -\frac{q(\mathbf{d}|e)}{\sum_{c'\neq c} P(c')q(\mathbf{d}|c')}. \tag{25}
$$

次に, 式 (9) の定義に従うことにより, 我々の TOP カーネルが計算できる. 本稿ではこのカーネルを *hyperplane-based TOP (HP-TOP)* カーネルと呼ぶことにする.

最も多くの素性を提供している式 (21) の形の偏導関数では, 元の素性 d_i が確率分布から計算された値で重み付けされている. このような式の形から, 二つの事例は, 類似したカテゴリ事後確率を持つ場合に, より似ていると見なされることがわかる. 導関数 (16) と (17) は, それぞれ正例集合と負例集合の一次の差異の影響を表している. 同様に, 導関数 (18) と (19) は二次の差異の影響を表している. さらに, 導関数 (22) と (23) は分散の自乗で正規化された一次の差異を表している.

式 (21) の形の素性数は, (元の素性空間の次元) × (カテゴリ数) オーダーであり, カーネル関数をこのままの形で計算するのはコストが高い. そこで, 我々は以下のように工夫することで高コストの計算を避ける. 任意の二つのベクトル $\mathbf{d}^1$ と $\mathbf{d}^2$ に関する HP-TOP カーネルのうち, 導関数 (21) の

表 1: The categories and their sizes of Reuters-21578

Category	tr texts	test texts
earn	2725	1051
acq	1490	644
money-fx	464	141
grain	399	135
crude	353	164

Category	tr texts	test texts
trade	339	133
interest	291	100
ship	197	87
wheat	199	66
corn	161	48

寄与部分を計算してみる：

$$\begin{aligned} & \sum_{e\neq c}\sum_i \frac{\partial v(\mathbf{d}^1,\theta)}{\partial w_{ei}}\frac{\partial v(\mathbf{d}^2,\theta)}{\partial w_{ei}} \\ = & \sum_{e\neq c}\sum_i \frac{P(e)^2 q(\mathbf{d}^1|e)q(\mathbf{d}^2|e)}{P_{-c}(\mathbf{d}^1)P_{-c}(\mathbf{d}^2)}\frac{\mu_e-(\mathbf{w}_e\cdot\mathbf{d}-b_e)}{\sigma_e^2}\frac{\mu_e-(\mathbf{w}_e\cdot\mathbf{d}-b_e)}{\sigma_e^2}d_i^1 d_i^2 \\ = & \left(\sum_{e\neq c}\frac{P(e)^2 q(\mathbf{d}^1|e)q(\mathbf{d}^2|e)}{P_{-c}(\mathbf{d}^1)P_{-c}(\mathbf{d}^2)}\frac{\mu_e-(\mathbf{w}_e\cdot\mathbf{d}-b_e)}{\sigma_e^2}\frac{\mu_e-(\mathbf{w}_e\cdot\mathbf{d}-b_e)}{\sigma_e^2}\right)\mathbf{d}^1\cdot\mathbf{d}^2, \end{aligned} \tag{26}$$

ここで, $P_{-c}(\mathbf{d})$ は $\sum_{c'\neq c}P(c')q(\mathbf{d}|c')$ を表すとする.

(26) の最後の式は, 二つの内積値の積とみなすことができる. よって, ベクトル $\mathbf{d}$ と

$$\left(\frac{P(e)q(\mathbf{d}|e)}{P_{-c}(\mathbf{d})}\frac{\mu_e-(\mathbf{w}_e\cdot\mathbf{d}-b_e)}{\sigma_e^2}\right)_{e\neq c}, \tag{27}$$

を素性 (21) の代わりに持っておくことにより, この部分の内積を効率的に計算できる. この部分の内積は, 全体の内積の計算量について支配的であるので, 結局全体の内積計算量のオーダーは元の素性空間の次元になる (元の次元はカテゴリ数より大きいとする). 以上のように, (元の次元) × (クラスタ数) オーダーの計算量を必要とする PLSI ベースのフィッシャー・カーネルなどの関数よりも, HP-TOP カーネルは計算量の点で優れている.

PLSI ベースのフィッシャー・カーネルでは, 各単語は隠れ変数上の確率分布を持つ. このような意味において, PLSI ベースのフィッシャー・カーネルは, より細かく類似性を測ろうとするが, 同時に上で示したように計算量にいては不利である.

与えられたカテゴリを隠れ変数として用いることにより, PLSI ベースのフィッシャー・カーネルを TOP カーネルに近付けることも可能であるが, 計算量の問題は解決されない.

6 実験

文書分類の実験を通して, 提案した HP-TOP カーネルを線形カーネルや PLSI ベースのフィッシャー・カーネルと比較する. 用いたデータセットは, ModApte-split (Dumais, 1998) を施した Reuters-21578 である. さらに, 本文が実質的に空であるような文書を削除した結果, 8815 訓練文書と, 3023 テスト文書を得た. 訓練文書全体で頻度 5 以上であるような単語のみを素性として用いた.

訓練データのサイズは, 1000 から 8000 まで 1000 ずつ変化させた. 各サイズについて, 10 個の異なるセットで実験を行った.

結果は, 頻度の高い 10 カテゴリ (表 1) の F 値の平均で評価した. 全カテゴリ数は 116 である. HP-TOP カーネルの構築には, まず元の素性空間で各カテゴリに対して分類平面を求める必要がある. しかし, 小さなカテゴリに関しては, 信頼できるモデルが得られるとは考えにくい. よって, 我々は高頻度の 10 カテゴリ以外のカテゴリを一つのカテゴリと見なした. つまり, 負例集合は, 10 個のモデルから成ることなる (頻度の高い 9 個のカテゴリと, 残りの多くのカテゴリから作った 1 個の新しいカテゴリである).

各カテゴリの確率分布を推定には, そのカテゴリに属する訓練文書のみを使って単純な最尤推定を行った. より詳細には, 各訓練文書について, 元の素性空間での分離平面の垂線方向成分を抽出して ($\mathbf{w}_{c'}\cdot\mathbf{d}-b_{c'}$ なる変換を行った), これら訓練文書を一次元データに変換した.

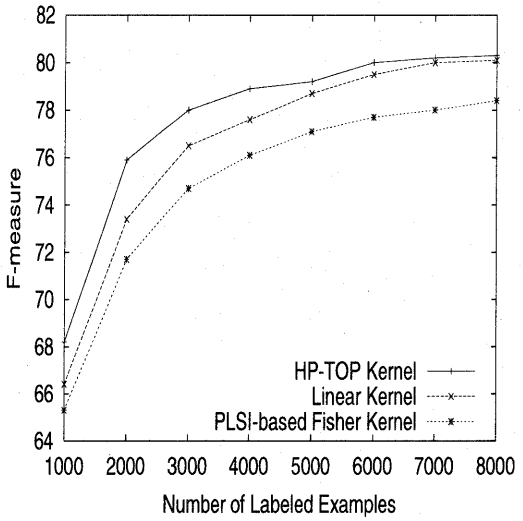

図 1: Macro-average of F-measure

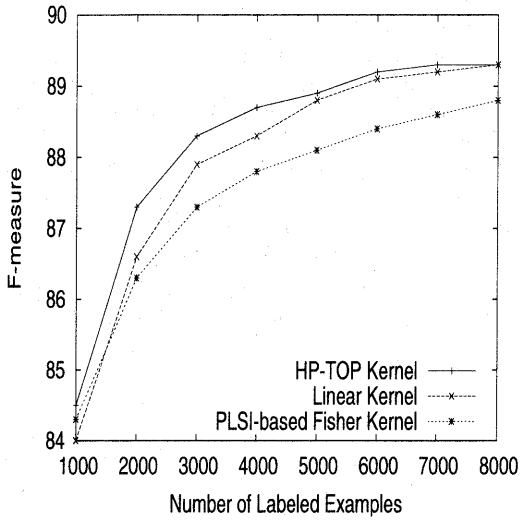

図 2: Micro-average of F-measure

カテゴリの事前分布は一様分布とした (Tsuda, 2002). PLSI ベースのフィッシャー・カーネルについては, 隠れ変数の数を 10, 20, 30 と変えて確率モデルを推定し, それらを足し合わせることでロバストなカーネル関数を作った (Hofmann, 2000). SVM のソフトマージン・パラメータ C は, 1.0 に固定した（異なる C についても簡単に実験を行ったが結果に顕著な違いは見られなかった）.

結果は図 1（マクロ平均）と図 2（マイクロ平均）に示す. HP-TOP カーネルは, 線形カーネルとフィッシャー・カーネルを全ての訓練データサイズにおいて上回っていることがわかる.

各訓練データサイズで, HP-TOP カーネルと線形カーネルの差について有意水準 5%でウィルコクソン検定を行った. その結果, 訓練データサイズ 1000 から 5000 については, 差が有意であった.

今回の実験設定では, PLSI ベースのフィッシャー・カーネルはあまりよい結果を出さなかった. しかし, ホフマン (Hofmann, 2000) が報告しているように, より小さな訓練データについてはフィッシャー・カーネルが比較的うまく働くものと思われる.

7 結論

分離平面を基にした TOP カーネルを提案した. 提案したカーネルは分離平面の垂線方向に想定した一次元ガウス型関数によって作られる. 文書分類において, 我々のカーネルは高い精度を示した.

線形カーネルなどに対する優位性はある程度示されたものの, さらなる研究が必要である. まず, 大きな訓練データサイズ（7000 と 8000）については, F 値の差が統計的に有意とはいえなかった. この点についてさらなる実験的あるいは理論的解析が必要であろう. また, カテゴリ構造のモデル化は他の確率モデルを用いても可能である. いくつかの確率モデルについて実験的に試してみる必要があるだろう. 特に, ガウス型の関数は我々の分布に対する直感に反して対称的であり, その使用については議論の余地が残る.

このモデルは EM アルゴリズムを用いるなどしてラベル無しデータを取り入れられるように拡張が可能である. また, 負例のカテゴリが与えられていない場合は提案手法は使えない. そのような場合のために, カテゴリ構造の推定に教師無しのクラスタリングなどが使用できると, 我々の手法はより応用可能性が広がる.

References

S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. *ACM-CIKM'98*, pp. 148–155, 1998.

T. Hofmann. Probabilistic Latent Semantic Indexing. *SIGIR'99*, pp. 50–57, Berkeley, California, August 1999.

T. Hofmann. Learning the similarity of documents: An information geometric approach to document retrieval and categorization. *NIPS 12*, pp. 914–920, 2000.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *NIPS 11*, pp. 487–493, 1998.

R. E. Kass and P. W. Vos. *Geometrical foundations of asymptotic inference.* New York : Wiley, 1997.

K. Tsuda and M. Kawanabe. The leave-one-out kernel. *ICANN* pp. 727–732, 2002.

K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.-R. Müller. A new discriminative kernel from probabilistic models. *Neural Computation*, 14(10), pp. 2397–2414, 2002.

V. Vapnik. *Statistical Learning Theory.* John Wiley, New York, 1998.