論 文

Technical Papers

半教師あり学習のための生成・識別ハイブリッド分類器の設計法

A Hybrid Generative/Discriminative Classifier Design for Semi-supervised Learing

藤野 昭典 Akinori Fujino
日本電信電話株式会社　NTT コミュニケーション科学基礎研究所
NTT Communication Science Labolatories, NTT Corporation
a.fujino@cslab.kecl.ntt.co.jp

上田 修功 Naonori Ueda
（同　上）
ueda@cslab.kecl.ntt.co.jp

斉藤 和巳 Kazumi Saito
（同　上）
saito@cslab.kecl.ntt.co.jp

keywords: generative and bias correction models, maximum entropy principle, EM algorithm, naive Bayes model, multiclass and single-labeled classification

Summary

Semi-supervised classifier design that simultaneously utilizes both a small number of labeled samples and a large number of unlabeled samples is a major research issue in machine learning. Existing semi-supervised learning methods for probabilistic classifiers belong to either generative or discriminative approaches. This paper focuses on a semi-supervised probabilistic classifier design for multiclass and single-labeled classification problems and first presents a hybrid approach to take advantage of the generative and discriminative approaches. Our formulation considers a generative model trained on labeled samples and a newly introduced bias correction model, whose belongs to the same model family as the generative model, but whose parameters are different from the generative model. A hybrid classifier is constructed by combining both the generative and bias correction models based on the maximum entropy principle, where the combination weights of these models are determined so that the class labels of labeled samples are as correctly predicted as possible. We apply the hybrid approach to text classification problems by employing naive Bayes as the generative and bias correction models. In our experimental results on three English and one Japanese text data sets, we confirmed that the hybrid classifier significantly outperformed conventional probabilistic generative and discriminative classifiers when the classification performance of the generative classifier was comparable to the discriminative classifier.

1. はじめに

統計的手法に基づく分類器は，一般的に，クラスラベル y が既知の特徴ベクトル $\boldsymbol{x}$ (ラベルありデータ) を用いて学習される．また，このとき，多数のラベルありデータを学習に用いるほど分類器の精度が向上することが知られている．しかし，ラベルありデータの作成は，分類対象に精通した専門家によるラベルの付与が必要であり，多くの場合に高コストである．一方，ラベルなしデータは，低コストで集めることができることが多い．例えば web ページの分類問題では，インターネットから容易にデータを収集することができる．それ故，少数のラベルありデータと多数のラベルなしデータを用いて分類器の精度を向上させる半教師あり学習法は機械学習の重要な研究課題の１つであり，従来よりさまざまな分類器が提案されてきた [Nigam 00, Grandvalet 05, Szummer 01, Amini 02, Blum 98, Zhu 03]．文献 [Seeger 01] に，半教師あり学習に関する様々な従来研究が概説されている．

確率モデルに基づく半教師あり学習法は，従来，生成アプローチ，あるいは識別アプローチに基づいて提案されてきた．生成アプローチでは，特徴ベクトル $\boldsymbol{x}$ とクラスラベル y の同時確率分布 $p(\boldsymbol{x}, y)$ [*1] をモデル化する．さらに，ベイズ則によりクラス事後確率 $P(y|\boldsymbol{x})$ を求め，その確率を最大にするクラスラベル y を選択することで分類を行う．生成アプローチでは，ラベルなしデータを，クラスラベルに関する不完全データとして扱うことにより，分類器の学習に用いる [Nigam 00].

*1 本論文では，確率と確率密度関数をそれぞれ $P(y)$ と $p(\boldsymbol{x})$ のように大文字・小文字で区別して表記する．また，確率変数に対する確率と確率密度関数の分布をともに確率分布と呼ぶ．

識別アプローチでは，特徴ベクトルのクラス事後確率 $P(y|\boldsymbol{x})$ を直接モデル化する．ただし，識別アプローチでは $p(\boldsymbol{x})$ をモデル化しないため，ラベルなしデータを学習に利用する際には新たな仮定が必要となる．例えば，特徴ベクトル間の距離に基づいた仮定 [Szummer 01]，クラス事後確率のエントロピーに基づく仮定 [Grandvalet 05] などが用いられている (詳細は 2･2 節参照)．

一般的に，識別アプローチに基づく分類器 (識別分類器) は生成アプローチに基づく分類器 (生成分類器) と比較して高い分類性能をもつことが知られているが，ラベルありデータが少数であるとき，生成分類器が識別分類器よりしばしば高い分類性能をもつことが報告されている [Ng 02]．そこで本研究では，生成・識別アプローチの双方の利点を活かすためのハイブリッド法を検討する．

教師あり学習の枠組では，生成・識別アプローチのハイブリッドが試みられている [Tong 00, Raina 04]．文献 [Tong 00] では，生成モデルに基づくベイズ分類器で得られるクラス境界を識別基準で修正する手法を提案し，特徴ベクトルに欠損値があるデータに対する手法の有効性を報告している．しかし，クラスが欠損値となるラベルなしデータを分類器学習に用いる問題を扱っていない．文献 [Raina 04] では，特徴ベクトルの部分空間ごとに生成モデルを仮定し，これらの生成モデルをクラス事後確率を最大化するように重み付け結合して分類器を構築する手法を提案している．テキスト分類問題において，文書を本文とタイトルの 2 つの部分空間に分割して分類器を構築することで純粋な生成，識別分類器より高い精度が得られることを報告している．

本研究では，半教師あり学習の問題に対して，生成・識別アプローチのハイブリッドに基づく新しい分類器設計法を提案する．文献 [Raina 04] では，特徴ベクトルの部分空間の結合を行うのに対し，提案法では，ラベルあり・なしデータの結合を行うことを特徴とする．具体的には，まず，ラベルありデータにより生成モデルを学習する．ラベルありデータが少数であるとき，生成モデルから得られる分類器で得られるクラス境界は，真のクラス境界から大きくずれる (偏りが大きい)．そこで，この偏りを補正するための新たなモデル (偏り補正モデル) を導入する．そして，最大エントロピー原理 [Berger 96] に基づき，クラス事後確率を最大化するように生成モデルと偏り補正モデルを結合することにより，生成・識別ハイブリッド分類器を得る．基本的なアイデアは既に [Fujino 05a, 藤野 05b] で提起したが，本論文では EM アルゴリズム的なパラメータ学習法，および実験データを追加して有用性を確認する．

本論文では，まず 2 節で，従来の生成，識別アプローチに基づく半教師あり学習法について述べる．つぎに 3 節では，本研究で提案するハイブリッド法について述べる．さらに 4 節では，ハイブリッド法の生成モデルに多項分布に基づくナイーブベイズ (NB) モデルを用いてテキスト分類問題に適用したときの実験結果を示し，従来法との比較により提案法の有効性を考察する．

2. 従　来　法

本論文では，特徴ベクトル $\boldsymbol{x}$ のクラスラベル y を K 個の候補 $\{1,\ldots,k,\ldots,K\}$ の中から 1 つに決定する多クラス単一ラベル分類問題に焦点を当てる．半教師あり学習では，少数のラベルありデータ $D_l=\{\boldsymbol{x}_n,y_n\}_{n=1}^N$ と多数のラベルなしデータ $D_u=\{\boldsymbol{x}_m\}_{m=1}^M$ を用いて分類器を学習する．本節では，生成，識別のそれぞれのアプローチに基づく従来の半教師あり学習法について述べる．

2･1 生成アプローチ

生成アプローチでは，同時確率モデル (生成モデル) $p(\boldsymbol{x},y;\boldsymbol{\theta}_y)$ を仮定し，訓練データを用いてパラメータ $\Theta=\{\boldsymbol{\theta}_k\}_{k=1}^K$ を学習する．$\boldsymbol{x}$ のクラスラベル y は，ベイズ則によりクラス事後確率 (ベイズ事後確率) $P(k|\boldsymbol{x};\Theta)\propto p(\boldsymbol{x},k;\boldsymbol{\theta}_k)$ を求め，その確率を最大化する k を選択することで推定される．生成モデルは分類対象の特徴に合わせて，多項分布やガウス分布を用いて仮定できる．

生成モデルによる半教師あり学習では，混合モデル $p(\boldsymbol{x};\Theta)=\sum_{k=1}^K p(\boldsymbol{x},k;\boldsymbol{\theta}_k)$ を仮定して，ラベルなしデータをクラスラベルに関する不完全データとして扱う [Dempster 77]．訓練データ $D=\{D_l,D_u\}$ が与えられるとき，MAP 推定では，以下の目的関数を最大化する Θ をパラメータの推定値とする．

$$J_1(\Theta)=\sum_{n=1}^N\log p(\boldsymbol{x}_n,y_n;\boldsymbol{\theta}_{y_n})+\sum_{m=1}^M\log\sum_{k=1}^K p(\boldsymbol{x}_m,k;\boldsymbol{\theta}_k)+\log p(\Theta)\qquad(1)$$

上式の $p(\Theta)$ は Θ の事前確率分布である．$J_1(\Theta)$ によるパラメータ推定は EM アルゴリズムを用いて行うことができる．

一般に，Θ の推定は，$M>>N$ のとき，式 (1) の右辺の第二項に強く影響され，ラベルなしデータによる教師なし学習に近い推定となる．生成モデルと混合モデルの仮定が分類対象に対して適切でない場合，ラベルなしデータをパラメータ学習に用いることで分類精度がむしろ悪化する危険性がある．この問題に対処するために，式 (1) の右辺の第二項のパラメータ学習への寄与を下げる重みパラメータ λ を導入する方法 (EM-λ) が提案されている [Nigam 00]．この方法では，ラベルありデータの leave-one-out 交差検定法により，分類精度を最大にするパラメータ Θ を与える λ を探索する．探索された λ の下で，パラメータ Θ を推定する．

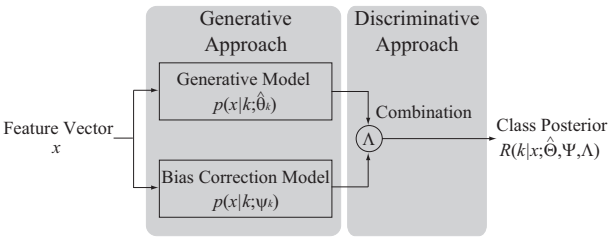

図 1 ハイブリッド法に基づく分類器の構成

2·2 識別アプローチ

識別アプローチでは，クラス事後確率 $P(y|\boldsymbol{x})$ を直接モデル化する．例えば，多項ロジスティック回帰 (MLR) モデル [Hastie 01] では，クラス事後確率は未知パラメータ $W=\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K\}$ を用いて以下のようにモデル化される．

$$P(y=k|\boldsymbol{x};W)=\frac{\exp(\boldsymbol{w}_k\cdot\boldsymbol{x})}{\sum_{k'=1}^{K}\exp(\boldsymbol{w}_{k'}\cdot\boldsymbol{x})} \qquad (2)$$

ここで，$\boldsymbol{w}_k\cdot\boldsymbol{x}$ は $\boldsymbol{w}_k$ と $\boldsymbol{x}$ の内積を表す．

識別アプローチでは，データの確率モデル $p(\boldsymbol{x})$ を仮定しないため，ラベルなしデータをパラメータ学習に用いるのに新たな仮定を必要とする．例えば，最小エントロピー正則項 (MER)[Grandvalet 05] に基づく方法では，クラスごとにデータがクラスターを形成しているという仮定に基づき，クラス事後確率のエントロピーを最小化することでラベルなしデータをよく分離するようにモデルを学習する．MER を用いた MLR の学習 (MLR/MER) では，以下の目的関数の最大化により W を求める．

$$\begin{aligned}J_2(W)=&\sum_{n=1}^{N}\log P(y_n|\boldsymbol{x}_n;W)\\&+\lambda\sum_{m=1}^{M}\sum_{k=1}^{K}P(k|\boldsymbol{x}_m;W)\log P(k|\boldsymbol{x}_m;W)\\&+\log p(W)\end{aligned} \qquad (3)$$

λ は重みパラメータであり，$p(W)$ は W の事前確率分布である．

3. ハイブリッド法

本研究では，生成アプローチと識別アプローチのハイブリッドに基づく半教師あり学習法 (ハイブリッド法) を提案する．ハイブリッド法では，生成モデルと新たに導入する偏り補正モデルを識別アプローチに基づいて結合することで分類器を構築する．以下にハイブリッド法の詳細を述べる．

3·1 分類器の構成

図 1 にハイブリッド法に基づく分類器の構成を示す．ハイブリッド法では，まず，クラスラベル y が k であるときの特徴ベクトル $\boldsymbol{x}$ の確率モデル (生成モデル)$p(\boldsymbol{x}|k;\boldsymbol{\theta}_k)$ を仮定し，教師あり学習と同様に，ラベルありデータを用いて学習する．ラベルありデータが少数であるとき，学習された生成モデル $p(\boldsymbol{x}|k;\hat{\boldsymbol{\theta}}_k)$ から得られるクラス境界は，真のクラス境界から大きくずれる (偏りが大きい)．そこで，生成モデルと同型の分布 (パラメータは異なる) で仮定する偏り補正モデル $p(\boldsymbol{x}|k;\boldsymbol{\psi}_k)$ を新たに導入する．これらを，訓練データのクラス事後確率を最大にするように，最大エントロピー (ME) 原理 [Berger 96] に基づいて結合することで，分類器の識別関数を与える．偏り補正モデルのパラメータ $\Psi=\{\boldsymbol{\psi}_k\}_{k=1}^{K}$ は，生成モデルの偏りを緩和するようにラベルなしデータを用いて学習する．以上の枠組により，ハイブリッド法では，生成過程に基づくデータの分布特性を積極的に利用する生成アプローチの長所と，識別率を最大化する識別アプローチの長所を兼ね備えた分類器を設計する．

3·2 生成モデルの学習

生成モデルのパラメータ $\Theta=\{\boldsymbol{\theta}_k\}_{k=1}^{K}$ は，ラベルありデータ D_l に対するパラメータの事後確率 $p(\Theta|D_l)$ を最大にするように推定する (MAP 推定)．ベイズ則 $p(\Theta|D_l)\propto p(D_l|\Theta)p(\Theta)$ により，以下の目的関数 $J(\Theta)$ を最大化するように Θ を計算して，推定値 $\hat{\Theta}=\{\hat{\boldsymbol{\theta}}_k\}_{k=1}^{K}$ を得る．

$$J(\Theta)=\sum_{n=1}^{N}\log p(\boldsymbol{x}_n|y_n;\boldsymbol{\theta}_{y_n})+\log p(\Theta) \qquad (4)$$

ここで，$p(\Theta)$ はパラメータ Θ の事前確率分布を表す．

3·3 ME 原理による識別関数の導出

ハイブリッド法では，分類器の識別関数を，生成モデルと偏り補正モデルの特性を反映する，エントロピー基準の下で最も一様なクラス事後確率分布で定義する．この分布は，生成モデルと偏り補正モデルに対する制約の下で ME 原理を満たすクラス事後確率分布 (ME モデル)$R(k|\boldsymbol{x})$ として与えることができる．

ME モデルに生成モデル $p(\boldsymbol{x}|k;\hat{\boldsymbol{\theta}}_k)$ の特性を反映させるため，訓練データの経験分布 $\tilde{p}(\boldsymbol{x},k)=\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n,k-y_n)/N$ による生成モデルの対数尤度の期待値と，$R(k|\boldsymbol{x})$ による生成モデルの対数尤度の期待値が等しい，という制約を与える．この制約は，$\boldsymbol{x}$ の経験分布 $\tilde{p}(\boldsymbol{x})=\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n)/N$ を用いて，以下の式で表される．

$$\begin{aligned}&\sum_{\boldsymbol{x},k}\tilde{p}(\boldsymbol{x},k)\log p(\boldsymbol{x}|k;\hat{\boldsymbol{\theta}}_k)\\&=\sum_{\boldsymbol{x},k}\tilde{p}(\boldsymbol{x})R(k|\boldsymbol{x})\log p(\boldsymbol{x}|k;\hat{\boldsymbol{\theta}}_k)\end{aligned} \qquad (5)$$

前述した偏り補正モデル $p(\boldsymbol{x}|k;\boldsymbol{\psi}_k)$ に対する制約は，式 (5) と同様の形で与えることができる．また，データが帰属するクラスの偏りを ME モデルに反映させるために，$R(k|\boldsymbol{x})$ によるクラス確率の推定値と，訓練データの経験分布によるクラス確率の推定値が等しい，という制約を

与える．この制約は以下の式で表すことができる．

$$\sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x},k) = \sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x}) R(k|\boldsymbol{x}), \forall k \qquad (6)$$

以上の制約の下で，クラス事後確率分布のエントロピー $H(R) = -\sum_{\boldsymbol{x},k} \tilde{p}(\boldsymbol{x}) R(k|\boldsymbol{x}) \log R(k|\boldsymbol{x})$ を最大化することにより，生成モデル，偏り補正モデルとクラスの偏りを反映した ME モデルのクラス事後確率分布:

$$R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma) = \frac{p(\boldsymbol{x}|k;\hat{\boldsymbol{\theta}}_k)^{\gamma_1} p(\boldsymbol{x}|k;\boldsymbol{\psi}_k)^{\gamma_2} e^{\mu_k}}{\sum_{k'=1}^{K} p(\boldsymbol{x}|k';\hat{\boldsymbol{\theta}}_{k'})^{\gamma_1} p(\boldsymbol{x}|k';\boldsymbol{\psi}_{k'})^{\gamma_2} e^{\mu_{k'}}} \qquad (7)$$

が導出できる．ここで，$\Gamma = \{\gamma_1, \gamma_2, \{\mu_k\}_{k=1}^K\}$ はラグランジュ乗数である．γ_1 と γ_2 は生成モデルと偏り補正モデルの結合の重みを，μ_k はクラス k の偏りを与えるパラメータである．仮に $\gamma_1 = 1, \gamma_2 = 0, \mu_k = P(k)$ とすれば，式 (7) は，$R(k|\boldsymbol{x};\hat{\Theta}) \propto p(\boldsymbol{x}|k;\hat{\boldsymbol{\theta}}_k) P(k)$ となる．すなわち，ME 原理により与えられるクラス事後確率分布は，特殊な場合に生成モデルのベイズ事後確率分布と一致する分布となっている．ハイブリッド法では，$R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ を分類器の識別関数とする．

ME 原理によれば，エントロピー $H(R)$ を最大化する R のパラメータ Γ は，R による訓練データの対数尤度 $L(\Gamma) = \sum_{\boldsymbol{x},k} \tilde{p}(\boldsymbol{x},k) \log R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ を最大化する Γ と一致する [Berger 96, Nigam 99]．このため，Ψ を与えると，式 (7) のクラス事後確率のパラメータ Γ は，ラベルありデータを用いた最尤推定により求められる．しかし，Γ の推定と $\hat{\Theta}$ の計算に同じラベルありデータを用いると，生成モデルの結合の重み γ_1 を過剰に大きく推定する危険性がある [Raina 04]．また，最尤推定は，一般的に，訓練データが少数のときに過学習を引き起こす．これら 2 つの問題を緩和するために，ラベルありデータの leave-one-out 交差検定法と，パラメータの事前確率分布を用いて Γ を推定する．具体的には，以下の式で表される目的関数 $F(\Gamma|\Psi)$ を最大化する Γ を計算する．

$$F(\Gamma|\Psi) = \sum_{n=1}^{N} \log R(y_n|\boldsymbol{x}_n;\hat{\Theta}^{(-n)},\Psi,\Gamma) + \log p(\Gamma) \qquad (8)$$

ここで，$\hat{\Theta}^{(-n)}$ はラベルありデータ $(\boldsymbol{x}_n, y_n)$ を除外して推定した生成モデルのパラメータ値を表し，$p(\Gamma)$ は Γ の事前確率分布を表す．ハイブリッド法ではガウス事前確率分布 [Chen 99] を用いて

$$p(\Gamma) \propto \prod_{j=1}^{2} \exp\left(-\frac{(\gamma_j - a_j)^2}{2\sigma_j^2}\right) \prod_{k=1}^{K} \exp\left(-\frac{\mu_k^2}{2\rho_k^2}\right) \qquad (9)$$

とする．

3·4 偏り補正モデルの学習

式 (7) の識別関数 $R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ は，特徴ベクトル $\boldsymbol{x}$ のクラスラベル y を，関数:

$$g_k(\boldsymbol{x};\boldsymbol{\psi}_k) = p(\boldsymbol{x}|k;\hat{\boldsymbol{\theta}}_k)^{\gamma_1} p(\boldsymbol{x}|k;\boldsymbol{\psi}_k)^{\gamma_2} e^{\mu_k} \qquad (10)$$

訓練データセット:
$D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$
1. 初期化: $\Psi^{(0)}$ の設定, $t \leftarrow 0$.
2. 式 (4) により $\hat{\Theta}$ と $\hat{\Theta}^{(-n)}$, $\forall n$ を計算
3. $\hat{\Theta}^{(-n)}$ と $\Psi^{(0)}$ の下で式 (8) により $\Gamma^{(0)}$ を計算
4. 反復学習 (収束条件 $\frac{|\Psi^{(t+1)} - \Psi^{(t)}|}{|\Psi^{(t)}|} < \epsilon$)
   - $\Gamma^{(t)}$ と $\hat{\Theta}$ の下で式 (12) により $\Psi^{(t+1)}$ を計算
   - $\Psi^{(t+1)}$ と $\hat{\Theta}^{(-n)}$ の下で式 (8) により $\Gamma^{(t+1)}$ を計算
   - $t \leftarrow t+1$.
5. $R(k|\boldsymbol{x},\hat{\Theta},\Psi^{(t)},\Gamma^{(t)})$ を出力

図 2 パラメータ学習アルゴリズム

を最大にする k として決定する．このとき，クラス間での $g_k(\boldsymbol{x};\boldsymbol{\psi}_k)$ の差が大きい $\boldsymbol{x}$ の分類は容易であるが，すべてのクラスで $g_k(\boldsymbol{x};\boldsymbol{\psi}_k)$ が小さな値をとる $\boldsymbol{x}$ では，クラス間の差が小さいため，$R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ による分類の信頼性は高くないと考えられる．そこで，クラスラベルが未知の特徴ベクトル $\boldsymbol{x}$ に対し，クラス間での $g_k(\boldsymbol{x};\boldsymbol{\psi}_k)$ の差が増大することを期待して，総和 $\sum_{k=1}^K g_k(\boldsymbol{x};\boldsymbol{\psi}_k)$ が最大になるように偏り補正モデルのパラメータ Ψ を学習する．具体的には，ラベルなしデータを用いた，以下の目的関数 $G(\Psi|\Gamma)$ の最大化により Ψ を推定する．

$$G(\Psi|\Gamma) = \sum_{m=1}^{M} \log \sum_{k=1}^{K} g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k) + \log p(\Psi) \qquad (11)$$

ここで，$p(\Psi)$ はパラメータ Ψ の事前確率分布を表す．

Γ の値が既知のとき，EM アルゴリズム [Dempster 77] のような反復計算を行うことで，$G(\Psi|\Gamma)$ を初期値近傍で最大化する Ψ を推定できる．この反復計算では，(t) ステップでの Ψ の推定値を $\Psi^{(t)}$ とするとき，目的関数:

$$Q(\Psi^{(t+1)},\Psi^{(t)}|\Gamma) = \gamma_2 \sum_{m=1}^{M} \sum_{k=1}^{K} R(k|\boldsymbol{x}_m;\hat{\Theta},\Psi^{(t)},\Gamma) \log p(\boldsymbol{x}_m|k;\boldsymbol{\psi}_k^{(t+1)}) + \log p(\Psi^{(t+1)}) \qquad (12)$$

の最大化により，$(t+1)$ ステップでの推定値 $\Psi^{(t+1)}$ を求める ($Q(\Psi^{(t+1)},\Psi^{(t)}|\Gamma)$ の導出は付録 A を参照)．$\Psi^{(t+1)}$ と $\Psi^{(t)}$ の差が十分小さくなるまで反復計算を行うことで Ψ の推定値が得られる．

しかし Γ は，前節で述べたように，Θ と Ψ が与えられたときに式 (8) を用いて学習されるパラメータである．一方，式 (12) による Ψ の学習のために，Γ の値を与える必要がある．すなわち，Γ と Ψ のパラメータ学習には相互に依存関係がある．そこで，Ψ と Γ を反復的に交互に学習する．具体的には，初期値 $\Psi^{(0)}$ を与えて，式 (8) により Γ を推定する．得られた Γ の推定値と $\Psi^{(0)}$ を用いて，式 (12) により $\Psi^{(1)}$ を推定する．そして，推定値 $\Psi^{(1)}$ を用いて Γ を再推定する．このような反復計算を繰返すことで，Γ と Ψ を推定する．以上のパラメータ学習のアルゴリズムを図 2 にまとめる．

4. 実　　験

4·1 データセット

ハイブリッド法の有効性を検証するため，テキスト分類実験のベンチマークテストに用いられる 3 つのテストコレクション (Reuters-21578, WebKB, 20 Newsgroups) と，Goo カテゴリに登録されている web ページから作成したデータセットを用いて評価実験を行った．

Reuters-21578 (Reuters) は，135 のトピックカテゴリからなる Reuters newswire の英文記事を集めたテストコレクションである [Yang 99]．文書分類のベンチマークテストでは，記事を多く含む 10 のトピックカテゴリを用いて分類実験が行われることが多い [Nigam 00]．そこで，この 10 カテゴリの中から単一のカテゴリに含まれる記事のみを抽出して評価実験用データセットを作成した．抽出の結果，2 カテゴリにはほとんど記事が含まれなかったため，データセットから除外した．本実験に用いたカテゴリは **acq, crude, earn, grain, interest, money-fx, ship, trade** の 8 つである．Reuters によるベンチマークテストでは，投稿時期の早い記事と遅い記事をそれぞれ訓練データ，テストデータとして用いることが多い．本実験でも同様に，投稿時期の遅い 2430 記事をテストデータとして用い，投稿時期の早い 5732 記事の中から学習に用いるラベルあり・なしデータをランダムに選択した．各記事の特徴ベクトルは単語の出現頻度から求め，その際に，停止語リスト [Salton 83] に含まれる単語と，1 つの記事のみに出現する単語を除外した．データセットの語彙の総数は 11822 であった．

WebKB は，アメリカの大学の web ページから作成されたテストコレクションであり，7 カテゴリから構成される．これらのカテゴリのうち，4 カテゴリ **course, faculty, project, student** がベンチマークテストに用いられることが多い [Nigam 99]．そこで，この 4 カテゴリに含まれる 4199 ページを用いて評価実験を行った．Reuters と同様に単語を除外した結果，データセットの語彙の総数は 18525 であった．

20 Newsgroups (20news) は，20 の議論グループに投稿された英文の記事から作成されたテストコレクションである．これらの議論グループのうち，従来研究 [Nigam 99] によるベンチマークテストと同様に，**comp** を親ディレクトリにもつ 5 グループを評価実験に用いた．記事の総数は 4881 であった．Reuters と同様に単語を除外した結果，データセットの語彙の総数は 19357 であった．

また，Goo カテゴリ[*2]に登録されている日本語の web ページを用いてデータセット (Goo) を作成した．「ビジネスと経済＞企業＞」配下から多くの web ページを含む 5 つのサブカテゴリを選択し，その中の 1 つのサブカテゴリのみに属する web ページを採集してデータセットを作成した．各ページに含まれる単語の抽出には茶筌[*3]を用いた．データセットに含まれる web ページの総数は 7791，語彙の総数は 36810 であった．

表 1 データサンプル数

	$\|D_l\|$	$\|D_u\|$	$\|D_t\|$
Reuters	16 - 1024 の 7 種類	4500	2430
WebKB	8 - 512 の 7 種類	2500	1000
20news	10 - 1280 の 8 種類	2500	1000
Goo	10 - 1280 の 8 種類	4000	2000

4·2 実 験 方 法

ハイブリッド法をテキスト分類に適用するため，生成モデルとしてナイーブベイズ (NB) モデル [上田 04] を用いた．NB モデルでは，各文書の特徴量として文書中に出現した単語の頻度を用いる．語彙の総数を V とするとき，文書の特徴ベクトルは文書中に含まれる単語の頻度ベクトル $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_V)$ で表現される．ここで，x_i は単語 i が文書中で出現した回数を表し，V は語彙の総数を表す．

NB モデルでは，文書中の各単語は独立に生成されると見なし，クラス k に属す文書の単語頻度ベクトル $\boldsymbol{x}$ の確率分布が，

$$P(\boldsymbol{x}|k;\boldsymbol{\theta}_k) \propto \prod_{i=1}^{V}(\theta_{ki})^{x_i}, \theta_{ki} > 0, \sum_{i=1}^{V}\theta_{ki} = 1 \quad (13)$$

で表される多項分布に従うと仮定する．ここで，$\theta_{ki} > 0$ はクラス k に属す文書中で単語 i が生起する確率を表す未知パラメータである．NB モデルでは，式 (4) の事前確率分布 $p(\Theta)$ として，一般的にディリクレ分布 $p(\Theta) \propto \prod_{k=1}^{K}\prod_{i=1}^{V}(\theta_{ki})^{\xi_k - 1}$ が用いられる [上田 04]．ξ_k はハイパーパラメータであり，定数として与える．

ハイブリッド法では，各文書で要素の和 $\sum_{i=1}^{V} x_i$ が一定となるように特徴ベクトル $\boldsymbol{x}$ を正規化した上で，生成モデル $p(\boldsymbol{x}|k;\boldsymbol{\theta}_k)$ と偏り補正モデル $p(\boldsymbol{x}|k;\boldsymbol{\psi}_k)$ に NB モデルを適用した．$\boldsymbol{\theta}_k$ を推定するためのハイパーパラメータ ξ_k は，ラベルありデータの leave-one-out 交差検定法により推定される生成モデルの対数尤度 $\log p(\boldsymbol{x}_n|y_n;\boldsymbol{\theta}_{y_n}^{(-n)})$ が最大になるように調節した．この調節は，EM アルゴリズム [Dempster 77] を援用して効率的に行った．

評価実験は，各データセットから表 1 に示す個数のデータサンプルを無作為に抽出して行った．$|D_l|$ 個のラベルありデータと $|D_u|$ 個のラベルなしデータを用いて分類器を学習し，$|D_t|$ 個のテストデータを用いて分類精度を算出した．分類精度は，正しくクラスが予測されたテストデータの数の総数に対する割合で与えられる．分類器の性能の評価は，データサンプルの無作為抽出から分類精度の算出までの実験を 10 回繰り返して得られる分類精度の平均値を用いて行った．

∗2 http://dir.goo.ne.jp/

∗3 http://chasen.naist.jp/hiki/ChaSen/

ハイブリッド法の性能は，EM-λ と MLR/MER の 2 つの半教師あり学習法の性能と比較することによって評価した．EM-λ の生成モデルには NB モデルを用いた．また，各データセットにおける教師あり学習での生成，識別分類器の性能を確かめるため，ラベルなしデータを用いずに学習を行った NB モデルと MLR モデルの分類精度も合わせて調べた．

EM-λ と MLR/MER では，ラベルなしデータのモデルパラメータ学習への寄与度を表す重みパラメータ λ を設定する必要がある．本実験では，EM-λ の重みパラメータの候補として，$\{0.01, 0.05, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8.0.9, 1.0\}$ の 14 種類の値を用いた．パラメータ値は，文献 [Nigam 00] で提案された方法に従って，ラベルありデータの leave-one-out 交差検定法において分類精度が最も高くなるパラメータ値を選択した．また，MLR/MER の重みパラメータ λ は $\{\{0.1\times10^{-n}, 0.2\times10^{-n}, 0.5\times10^{-n}\}_{n=0}^{4}, 1\}$ の 16 候補から選択した．パラメータ値は，EM-λ と同様に交差検定法を用いて訓練データのみから決定 [Grandvalet 05] すべきであるが，パラメータ学習の計算コストが高いため，テストデータの予測精度を最も高くするパラメータ値を選択して性能比較を行った．

4·3 結果と考察

各データセットに対する分類精度の結果を表 2(a)-(d) に示す．表中の数値は 10 回の実験による分類精度の平均値を示し，括弧内の数値は標準偏差を示す．$|D_l|/|D_u|$ は，分類器の学習に用いたラベルありデータのラベルなしデータに対する個数比を示す．

まず，教師あり学習による生成分類器 NB と識別分類器 MLR の性能を比較する．表 2 より，文献 [Ng 02] で報告されているように，ラベルありデータが少数の場合は，生成分類器の性能が識別分類器よりも高く，逆に多数の場合は識別分類器の性能の方が高い傾向が本実験でもみられた．20news 以外の 3 つのデータセットでは，Reuters の $|D_l| = 32$, WebKB の $|D_l| = 8$, Goo の $|D_l| = 10$ で NB の分類精度が MLR より若干悪いものの，全体的にラベルありデータが少数の場合は NB の分類精度が MLR よりも高い傾向があった．また，ラベルありデータが増えるに従って MLR の分類精度が NB を逆転する傾向がみられた．

しかし，20news では，ラベルありデータ数によらず，MLR の分類精度は NB よりも高かった．この結果を分析するため，以下の式で定義される，語彙数で正規化した NB モデルのテストパープレキシティ $\mathcal{P}$ をテストコレクションごとに調べた．

$$\mathcal{P} = \frac{1}{V}\exp\left(-\frac{\sum_{k=1}^{K}\sum_{s=1}^{S} z_{sk}\sum_{i=1}^{V} x_{si}\log\hat{\theta}_{ki}}{\sum_{s=1}^{S}\sum_{i=1}^{V} x_{si}}\right) \tag{14}$$

$\mathcal{P}$ は，学習に用いていないテストデータ $(\boldsymbol{x}_s, y_s)$ に対して分類器がどれだけ適合しているかを表す指標である．上式の $\hat{\theta}_{ki}$ はラベルありデータによって学習されたモデルパラメータである．z_{sk} は $y_s = k$ のときに 1，$y_s \neq k$ のときに 0 をとる変数である．$\mathcal{P}$ が小さいほど，モデルはテストデータに適合していることを示す．図 3 より，すべてのデータセットでラベルありデータが少数のときに $\mathcal{P}$ は大きく，ラベルありデータを増やすに従って $\mathcal{P}$ が小さくなった．20news では他の 3 つのデータセットと比べて，ラベルありデータが少数のときにとくに大きかった．また，100 過ぎまでラベルありデータを増やしても $\mathcal{P}$ が急激に小さくならず，他のデータセットとの差が広がる傾向があった．これは，20news ではラベルありデータが少数のときは学習された NB モデルはテストデータにあまり適合しないことを示している．逆に，他の 3 つのデータセットのように，ラベルありデータが少数のときに $\mathcal{P}$ が小さければ，NB は MLR の性能を上回る傾向があった．この結果，教師あり学習において，訓練データが少数で生成モデルのテストパープレキシティが十分小さいときに，生成分類器の性能は識別分類器を上回るといえる．

つぎに，半教師あり学習による生成，識別分類器とハイブリッド法による分類器の性能を比較する．まず，EM-λ は，すべてのデータセットにおいて，ラベルありデータが少数の場合は MLR/MER とほぼ同等，またはより高い分類精度を示し，逆に多数の場合は MLR/MER より低い分類精度を示した．このように，生成，識別分類器の比較では，教師あり学習での報告 [Ng 02] と同様の特徴が半教師あり学習の場合にもみられた．

EM-λ とハイブリッド法の比較では，教師あり学習において MLR の分類精度が NB を上回るときに，ハイブリッド法の分類精度が EM-λ を上回る傾向があった．この結果は，提案法では生成・識別アプローチのハイブリッドにより，識別アプローチの利点が活かされていることを示唆している．

MLR/MER とハイブリッド法の比較では，ラベルありデータが多数の場合を除いて，すべてのデータセットでハイブリッド法の分類精度が MLR/MER を上回った．これは，MLR/MER は少数のラベルありデータに対して過学習する傾向があるのに対し，ハイブリッド法では，分類対象のデータ特性に応じて生成モデルを仮定することで過学習が抑制されるからと考えられる．過学習を引き起こさないように十分なラベルありデータが与えられるときに，識別分類器の性能はハイブリッド法を上回ると考えられる．

最後に，EM-λ と MLR/MER, ハイブリッド法のパラメータ学習の計算コストについて述べる．図 4 は，WebKB による実験でパラメータ学習に要した計算時間の平均値を示したものである．ただし，MLR/MER については，実験では行っていないが，公平な比較のため，λ

表 2 各データセットでの分類精度 (%)

(a) Reuters

		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Hybrid	EM-λ	MLR/MER	NB	MLR
16	.0036	82.7 (5.4)	**86.9** (2.6)	73.9 (6.2)	71.1 (4.4)	67.8 (7.5)
32	.0071	**89.3** (1.7)	89.0 (2.9)	82.5 (4.1)	80.4 (1.8)	80.9 (3.5)
64	.014	**91.5** (0.9)	90.6 (3.1)	83.7 (5.0)	85.1 (2.2)	83.1 (4.6)
128	.028	**92.9** (0.8)	92.0 (0.8)	88.7 (1.5)	88.7 (0.6)	87.8 (1.3)
256	.057	**93.2** (0.9)	93.0 (1.0)	91.2 (1.1)	90.3 (1.0)	90.6 (0.8)
512	.11	**94.1** (0.4)	93.4 (0.5)	93.4 (0.5)	92.9 (0.6)	93.2 (0.7)
1024	.23	**94.7** (0.2)	94.3 (0.3)	**94.7** (0.2)	94.1 (0.4)	94.5 (0.2)

(b) WebKB

		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Hybrid	EM-λ	MLR/MER	NB	MLR
8	.0032	61.4 (5.5)	**63.4** (8.5)	54.0 (4.2)	53.0 (7.9)	53.6 (4.3)
16	.0064	66.7 (4.7)	**69.4** (2.6)	54.1 (5.9)	60.2 (3.7)	54.0 (5.2)
32	.013	72.5 (3.4)	**72.7** (2.5)	63.6 (4.6)	69.2 (2.8)	63.5 (4.3)
64	.026	**76.4** (2.2)	75.4 (2.4)	72.1 (2.7)	73.3 (0.9)	71.9 (2.3)
128	.051	**79.3** (1.5)	76.9 (2.2)	78.1 (2.4)	76.9 (2.1)	78.0 (2.2)
256	.10	81.2 (1.5)	79.7 (1.9)	**83.8** (1.8)	79.9 (1.4)	83.4 (1.7)
512	.20	83.1 (1.7)	82.2 (1.4)	**88.0** (1.3)	82.5 (1.3)	87.5 (1.1)

(c) 20news

		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Hybrid	EM-λ	MLR/MER	NB	MLR
10	.0040	**52.5** (13.0)	39.5 (9.0)	41.9 (8.4)	32.4 (6.1)	37.6 (5.5)
20	.0080	**64.2** (6.3)	50.3 (7.8)	45.2 (5.1)	41.3 (4.6)	44.6 (4.2)
40	.016	**70.0** (2.1)	56.2 (6.4)	52.5 (5.3)	47.4 (3.9)	50.9 (3.6)
80	.032	**73.2** (2.3)	61.3 (5.3)	59.7 (2.8)	53.4 (3.9)	59.0 (2.3)
160	.064	**75.7** (1.6)	67.2 (4.2)	68.1 (2.9)	61.1 (2.6)	66.6 (2.0)
320	.13	**78.2** (0.9)	70.8 (2.7)	74.4 (1.5)	68.8 (1.4)	72.7 (1.2)
640	.26	**81.2** (1.2)	75.8 (1.8)	79.2 (1.3)	74.9 (2.1)	77.7 (1.2)
1280	.51	**83.8** (1.1)	78.9 (2.1)	82.4 (1.3)	77.8 (1.6)	81.1 (1.3)

(d) Goo

		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Hybrid	EM-λ	MLR/MER	NB	MLR
10	.0025	**49.0** (7.4)	47.7 (13.4)	40.7 (5.0)	37.5 (5.0)	40.2 (4.9)
20	.0050	**66.7** (4.9)	62.5 (7.3)	46.3 (4.2)	47.7 (3.3)	46.3 (3.8)
40	.010	**74.4** (2.6)	72.0 (3.6)	57.3 (4.4)	59.2 (4.5)	56.9 (3.8)
80	.020	**78.1** (1.8)	77.5 (1.9)	64.8 (2.0)	67.6 (2.9)	64.7 (2.0)
160	.040	**80.3** (1.5)	79.4 (1.3)	72.0 (2.0)	74.8 (1.3)	71.9 (1.8)
320	.080	**82.5** (1.0)	81.9 (1.4)	77.7 (1.5)	79.7 (0.9)	77.5 (1.3)
640	.16	**84.6** (0.9)	83.4 (0.8)	81.9 (1.0)	83.1 (0.8)	81.4 (1.0)
1280	.32	**86.0** (0.7)	85.0 (0.9)	85.3 (0.7)	85.0 (0.8)	84.8 (0.5)

の最適な候補値を leave-one-out 交差検定法で推定するときに要する計算時間を見積ってプロットしている．

ハイブリッド法のパラメータは Θ, Ψ, Γ の 3 種類であり，Ψ, Γ については反復的に学習する．1 回の反復学習に要する計算コストは，Ψ と Γ の計算コストの和になる．ただし，Ψ の次元はクラス数 K と特徴ベクトルの次元数 V の積になり，Γ の次元 $K+2$ よりも遥かに大きい．本実験では，数十から数百回の反復計算を要した．

EM-λ では，EM アルゴリズムにより Θ を反復計算する．Θ の次元は Ψ と同数であり，Γ の学習は必要ない．したがって，1 回の反復で要する計算コストはハイブリッド法よりやや小さい．実験では，パラメータ学習に数回から数十回の反復計算を要した．このため，Θ の計算コストは，ハイブリッド法の Ψ と比べて 1/10 程度であった．ただし，EM-λ では，λ もまた推定すべきパラメータであり，交差検定法で最適な λ の値を探索する．したがって，leave-one-out 交差検定法を用いる場合，EM-λ の計算コストは Θ の計算コストにラベルありデータ数と λ の候補数を乗じたものになる．図 4 より，実験では，EM-λ の計算時間は，ラベルありデータ数 $|D_l|$ が 2^3 と少数の場合でもハイブリッド法より長く，ラベルありデータ数に比例する傾向があった．EM-λ では，交差検定法で λ を推定することで，ハイブリッド法より計算時間を要したといえる．

MLR/MER では，W の次元は EM-λ の Θ と等しい．また，EM-λ と同様に最適な W を得るために λ の推定を要する．ただし W の学習では，非線型最適化問題を解く必要があるため，Θ の学習と比べて非常に高い計算コストを要する．それ故，図 4 のように，MLR/MER では EM-λ と比較して多大な計算時間を要した．

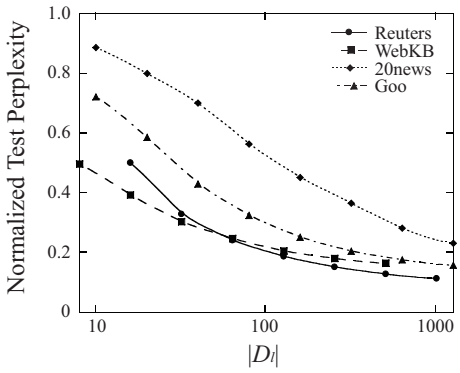

図 3 各テストコレクションでのテストパープレキシティ

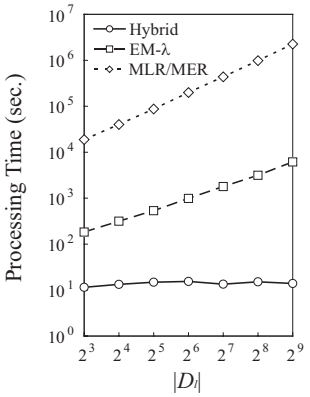

図 4 WebKB での計算時間

5. ま と め

生成・識別アプローチのハイブリッドに基づく，半教師あり学習のための分類器設計法を提案した．ハイブリッド法は，ラベルありデータにより学習する生成モデルと同型の分布でパラメータ値が異なる偏り補正モデルを新たに導入し，これらのモデルを最大エントロピー原理に基づいて結合することで分類器を構築することを特徴とする．実データによるテキスト分類実験を行った結果，教師あり学習による識別分類器が生成分類器と同等またはやや高い分類性能であったときに，ハイブリッド法が半

教師あり学習による生成分類器，識別分類器よりも高い性能をもつことを確認した．したがって，生成，識別の両アプローチがほぼ同等の性能であるときに，ハイブリッド法は有効であると考えられる．今後の課題は，本実験で用いた多項分布以外で生成モデルが与えられるような分類対象や，複数の生成モデルによりモデル化されるような分類対象に対して，ハイブリッド法の有効性を検証することである．

◇ 参 考 文 献 ◇

[Amini 02] Amini, M. R. and Gallinari, P.: Semi-supervised logistic regression, in *Proceedings of the 15th European Conference on Artificial Intelligence*, pp. 390–394 (2002)

[Berger 96] Berger, A., Della Pietra, S., and Della Pietra, V.: A maximum entropy approach to natural language processing, *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71 (1996)

[Blum 98] Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with Co-Training, in *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98)*, pp. 92–100 (1998)

[Chen 99] Chen, S. F. and Rosenfeld, R.: A Gaussian prior for smoothing maximum entropy models, Technical report, Carnegie Mellon University (1999)

[Dempster 77] Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1–38 (1977)

[Fujino 05a] Fujino, A., Ueda, N., and Saito, K.: A hybrid generative/discriminative approach to semi-supervised classifier design, in *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, pp. 764–769 (2005)

[藤野 05b] 藤野 昭典, 上田 修功, 斉藤 和巳：生成・識別ハイブリッドモデルに基づく半教師あり学習, 情報科学技術レターズ, Vol. 4, pp. 161–164 (2005)

[Grandvalet 05] Grandvalet, Y. and Bengio, Y.: Semi-supervised learning by entropy minimization, in *Advances in Neural Information Processing Systems 17*, pp. 529–536, MIT Press, Cambridge, MA (2005)

[Hastie 01] Hastie, T., Tibshirani, R., and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York Berlin Heidelberg (2001)

[Ng 02] Ng, A. Y. and Jordan, M. I.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, in *Advances in Neural Information Processing Systems 14*, pp. 841–848, MIT Press, Cambridge, MA (2002)

[Nigam 99] Nigam, K., Lafferty, J., and McCallum, A.: Using maximum entropy for text classification, in *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61–67 (1999)

[Nigam 00] Nigam, K., McCallum, A., Thrun, S., and Mitchell, T.: Text classification from labeled and unlabeled documents using EM, *Machine Learning*, Vol. 39, pp. 103–134 (2000)

[Raina 04] Raina, R., Shen, Y., Ng, A. Y., and McCallum, A.: Classification with hybrid generative/discriminative models, in *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA (2004)

[Salton 83] Salton, G. and McGill, M. J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, New York (1983)

[Seeger 01] Seeger, M.: Learning with labeled and unlabeled data, Technical report, University of Edinburgh (2001)

[Szummer 01] Szummer, M. and Jaakkola, T.: Kernel expansions with unlabeled examples, in *Advances in Neural Information Processing Systems 13*, pp. 626–632, MIT Press, Cambridge, MA (2001)

[Tong 00] Tong, S. and koller, D.: Restricted Bayes optimal classifiers, in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*, pp. 658–664 (2000)

[上田 04] 上田 修功, 斉藤 和巳：多重トピックテキストの確率モデル-テキストモデル研究の最前線- (1), (2), 情報処理, Vol. 45, pp. 184–190, 282–289 (2004)

[Yang 99] Yang, Y. and Liu, X.: A re-examination of text categorization methods, in *Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR-99)*, pp. 42–49 (1999)

[Zhu 03] Zhu, X., Ghahramani, Z., and Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions, in *Proceedings of the 20th International Conference on Machine learning (ICML-2003)*, pp. 912–919 (2003)

〔担当委員：津本 周作〕

2005 年 12 月 14 日　受理

◇ 付　　録 ◇

A. 偏り補正モデルのパラメータ学習のための Q 関数の導出

式 (11) の目的関数 $G(\Psi|\Gamma)$ から，式 (12) で示した Q 関数 $Q(\Psi^{(t+1)},\Psi^{(t)}|\Gamma)$ を導出する過程を述べる．略記のため，式 (10) で示した $g_k(\boldsymbol{x};\boldsymbol{\psi}_k)$ を用いて，

$$z_{mk}(\Psi) \equiv \frac{g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k)}{\sum_{k'=1}^{K} g_{k'}(\boldsymbol{x}_m;\boldsymbol{\psi}_{k'})} = R(k|\boldsymbol{x}_m;\hat{\Theta},\Psi,\Gamma) \quad \text{(A.1)}$$

と定義すると，$(t+1)$ ステップと (t) ステップでの目的関数 $G(\Psi|\Gamma)$ の第一項 $G'(\Psi|\Gamma)=\sum_{m=1}^{M}\log\sum_{k=1}^{K} g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k)$ の差は，

$$\begin{aligned}
&G'(\Psi^{(t+1)}|\Gamma) - G'(\Psi^{(t)}|\Gamma)\\
&= \sum_{m=1}^{M}\log\frac{\sum_{k=1}^{K} g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k^{(t+1)})}{\sum_{k=1}^{K} g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k^{(t)})}\\
&= \sum_{m=1}^{M}\sum_{k'=1}^{K}\left[z_{mk'}(\Psi^{(t)})\log\left\{ \frac{\sum_{k=1}^{K} g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k^{(t+1)})}{\sum_{k=1}^{K} g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k^{(t)})} \right.\right.\\
&\quad \left.\left. \times\frac{g_{k'}(\boldsymbol{x}_m;\boldsymbol{\psi}_{k'}^{(t+1)})}{g_{k'}(\boldsymbol{x}_m;\boldsymbol{\psi}_{k'}^{(t+1)})}\frac{g_{k'}(\boldsymbol{x}_m;\boldsymbol{\psi}_{k'}^{(t)})}{g_{k'}(\boldsymbol{x}_m;\boldsymbol{\psi}_{k'}^{(t)})} \right\}\right]\\
&= \sum_{m=1}^{M}\sum_{k'=1}^{K}\left[z_{mk'}(\Psi^{(t)})\right.\\
&\quad \left.\times\log\left\{ \frac{g_{k'}(\boldsymbol{x}_m;\boldsymbol{\psi}_{k'}^{(t+1)})}{g_{k'}(\boldsymbol{x}_m;\boldsymbol{\psi}_{k'}^{(t)})}\frac{z_{mk'}(\Psi^{(t)})}{z_{mk'}(\Psi^{(t+1)})} \right\}\right]\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} z_{mk}(\Psi^{(t)})\log\frac{p(\boldsymbol{x}_m|k;\boldsymbol{\psi}_k^{(t+1)})^{\gamma_2}}{p(\boldsymbol{x}_m|k;\boldsymbol{\psi}_k^{(t)})^{\gamma_2}}\\
&\quad + \sum_{m=1}^{M}\sum_{k=1}^{K} z_{mk}(\Psi^{(t)})\log\frac{z_{mk}(\Psi^{(t)})}{z_{mk}(\Psi^{(t+1)})} \qquad \text{(A.2)}
\end{aligned}$$

となる．ここで，式 (A.2) の第二項は，各 m での $z_{mk}(\Psi^{(t)})$ と $z_{mk}(\Psi^{(t+1)})$ のカルバック・ライブラー (KL) 距離の和を表し，

$$\sum_{k=1}^{K} z_{mk}(\Psi^{(t)})\log\frac{z_{mk}(\Psi^{(t)})}{z_{mk}(\Psi^{(t+1)})} \geq 0, \forall m \qquad \text{(A.3)}$$

である．このため，式 (12) のように $Q(\Psi^{(t+1)}, \Psi^{(t)}|\Gamma)$ を定義すると，$(t+1)$ ステップと (t) ステップでの目的関数 $G(\Psi|\Gamma)$ の差は，式 (A.2) で得た $G(\Psi|\Gamma)$ の第一項の差に第二項の差 $p(\Psi^{(t+1)}) - p(\Psi^{(t)})$ を加味して

$$\begin{aligned} & G(\Psi^{(t+1)}|\Gamma) - G(\Psi^{(t)}|\Gamma) \\ \geq\ & Q(\Psi^{(t+1)}, \Psi^{(t)}|\Gamma) - Q(\Psi^{(t)}, \Psi^{(t)}|\Gamma) \end{aligned} \tag{A.4}$$

と書ける．したがって，$(t+1)$ ステップで目的関数 $G(\Psi|\Gamma)$ の値を最も増加させるためには，$Q(\Psi^{(t+1)}, \Psi^{(t)}|\Gamma)$ を最大化するような $\Psi^{(t+1)}$ を推定すればよい．この推定を反復的に行うことで，初期値近傍で目的関数 $G(\Psi|\Gamma)$ を最大化する Ψ を求めることができる．

著 者 紹 介

藤野 昭典

1972 年生．1995 年京都大学工学部精密工学科卒業．1997 年同大学大学院工学研究科精密工学専攻修士課程修了．同年 NTT 入社．機械学習等の研究に従事．現在，NTT コミュニケーション科学基礎研究所に所属．電子情報通信学会 PRMU 研究奨励賞 (2004 年度)，FIT 論文賞 (2005 年) 各受賞．電子情報通信学会会員．

上田 修功

1958 年生．1982 年大阪大学工学部通信工学科卒業．1984 年同大学大学院修士課程修了．工学博士．同年 NTT 入社．1993 年より 1 年間 Purdue 大学客員研究員．画像処理，パターン認識・学習，ニューラルネットワーク，統計的学習，Web データマイニング等の研究に従事．現在，NTT コミュニケーション科学基礎研究所 知能情報研究部長 奈良先端大客員教授．電気通信普及財団賞 (1997, 2006 年)，電子情報通信学会論文賞 (2002, 2004 年) 等受賞，電子情報通信学会，情報処理学会，日本神経回路学会，IEEE 各会員．

斉藤 和巳(正会員)

1963 年生．1985 年慶応義塾大学理工学部数理科学科卒業．工学博士． 同年 NTT 入社．1991 年より 1 年間オタワ大学客員研究員．神経回路網，機械学習，複雑ネットワーク等の研究に従事．現在，NTT コミュニケーション科学基礎研究所 主任研究員 (特別研究員) 奈良先端大客員助教授．情報処理学会論文賞 (1997 年)，人工知能学会論文賞 (1999 年) 等受賞．電子情報通信学会，情報処理学会，日本神経回路学会，IEEE 各会員．