複数の構成要素データを扱う多クラス分類器の半教師あり学習法

藤野昭典[†] 上田修功[†] 斉藤和巳[†,☆]

テキストとリンク等を含むWebページのように，複数の構成要素からなるデータの多クラス単一ラベル分類問題に対して，生成モデル・識別の両アプローチのハイブリッドに基づく分類器の半教師あり学習法を提案する．提案法では，それぞれの構成要素のために，個別に生成モデルを設計するとともに，ラベルありデータが少数であることに起因する生成モデルの学習の偏りの影響を緩和するためのモデルを導入する．さらに，最大エントロピー原理に基づいてこれらのモデルを統合することにより分類器を構築する．リンク等の付加情報を含むテキストデータの分類実験により，生成モデルであるナイーブベイズモデルと識別モデルである多項ロジスティック回帰モデルの半教師あり学習で類似の分類性能が得られる場合，提案法では両モデルより高い分類性能を得られることを確認した．また，提案法に基づき複数の構成要素を用いて分類器を設計することと，提案法で設計した分類器の学習に多数のラベルなしデータを用いることが高い分類性能を得るのに効果があることを確認した．

Semi-supervised Learning of Multi-class Classifiers for Multi-component Data

Akinori Fujino,[†] Naonori Ueda[†] and Kazumi Saito[†,☆]

We present a method for semi-supervised learning of multi-class and single-label classifiers for multi-component data such as web pages consisting of text and links, based on a hybrid generative/discriminative approach. In our formulation, for each component, we design an individual generative model and introduce a model to reduce the effect of the bias associated with the generative model trained on few labeled samples. Then, we construct our classifier by combining these models based on the maximum entropy principle. In our experimental results for text classification using additional information such as links, we confirmed that our classifier outperformed generative and discriminative classifiers based on naive Bayes and multinomial logistic regression models, especially when the performance of the generative and discriminative classifiers was comparable. We also confirmed our formulation for dealing with multiple components was effective to obtain semi-supervised classifiers with good generalization ability. Moreover, we confirmed that using a large number of unlabeled data for training our classifier improved its classification performance.

1. はじめに

テキスト，ハイパーリンク，画像等の要素から構成されるWebページのように，複数の構成要素を含むデータ（Multi-Component Data，MCD）の分類問題では，主要な情報が高精度な分類器を設計するうえで最も重要な役割を果たすとともに，その他の付加情報についても分類精度の向上に寄与する可能性がある．このため，複数の構成要素を同時に扱う分類器を設計することは，分類精度を向上させるうえで重要であると考えられる．このような背景の下で，従来より，Webページのテキストとハイパーリンク[3),13),20)]，文書のテキストと引用[13)]，音楽情報とテキスト[2)]等，複数の構成要素を扱う分類器が提案されてきた．

教師あり学習の枠組みでは，確率モデルアプローチによる任意の複数の構成要素を扱う分類器（以下，MCD分類器と呼ぶ）として，生成モデル，識別の各アプローチと，両アプローチのハイブリッドに基づく分類器（生成分類器，識別分類器，ハイブリッド分類器）が提案されてきた．生成分類器は，データ $\boldsymbol{x}$ とクラスラベル y の同時確率密度 $p(\boldsymbol{x}, y)$ [☆☆]をモデル化

† 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 Laboratories, NTT Corporation
☆ 現在，静岡県立大学経営情報学部
Presently with School of Administration and Informatics, University of Shizuoka

☆☆ 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．

して構築される．同時確率モデルは，各クラスで構成要素の独立性を仮定し，構成要素ごとに生成モデルを設計して与えられる[2]．一方，識別分類器は，クラス事後確率 $P(y|\boldsymbol{x})$ を直接モデル化[10] して構築される．ハイブリッド分類器は，構成要素ごとに設計した生成モデルを識別学習で得られる重みを用いて統合することで構築される．ハイブリッド分類器の性能は，生成，識別分類器より平均的に高いことが実験的に確認されている[6),17)]．

一方，汎化性能の高い分類器を獲得するためには，膨大なラベルありデータを必要とする場合が多い．しかし，ラベルありデータの作成は，人手によるラベルの付与を必要とし，高コストである．それに対して，ラベルなしデータは比較的に容易に収集できる．このため，少数のラベルありデータに加えて多数のラベルなしデータを用いて分類器の汎化性能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，様々な手法が提案されてきた（文献 19) 参照)．

以上のように，MCD 分類器の設計法と半教師あり学習法はそれぞれの分野で別々に研究されてきたが，両技術を効果的に組み合わせることで分類器の高精度化を期待できる．MCD 分類器を半教師あり学習させる問題に対して，生成，識別分類器に，生成モデル，識別の各アプローチで提案されてきた半教師あり学習法[9),16)] をそれぞれ適用することが考えられる．また，高い分類精度が期待できるハイブリッド分類器では，半教師あり学習で獲得した構成要素の生成モデルを識別学習に基づいて統合することが考えられる．しかし，このハイブリッド分類器の半教師あり学習法は，生成モデルアプローチのみに基づいており，生成モデル・識別両アプローチの利点を活かしきっていない．

そこで，本論文では，確率モデルアプローチに焦点を当て，MCD 分類器の半教師あり学習の問題に対して，生成モデル・識別両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．提案法では，ラベルありデータで学習される構成要素の各々の生成モデルと，ラベルなしデータを分類器の学習に効果的に用いるために筆者らが最近考案した偏り補正モデル[7)] を構成要素ごとに導入する．そして，最大エントロピー（Maximum Entropy，ME）原理[1)] に基づいてこれらのモデルを統合して分類器を構築する．3 つのテストコレクションを用いた MCD の分類実験により，生成モデル，識別の各アプローチに基づく半教師あり学習法の応用と比べて，提案法が汎化性能の高い分類器を獲得するのに有用であることを確認する．

2. MCD 分類器への従来の半教師あり学習法の適用

2.1 MCD 分類器の半教師あり学習問題

本論文では，多クラス単一ラベル分類問題に対して，複数の構成要素からなるデータ（MCD）を扱うことができ，かつ半教師あり学習を行うことで，汎化性能の高い分類器を設計することを課題とする．

多クラス単一ラベル分類問題とは，各データに対して，K 個の候補の中から 1 つのクラスラベル $y \in \{1,\ldots,k,\ldots,K\}$ を選択する問題である．J 個の構成要素からなるデータは，構成要素 j の特徴ベクトル $\boldsymbol{x}^j$ を用いて，$\boldsymbol{x} = (\boldsymbol{x}^1,\ldots,\boldsymbol{x}^j,\ldots,\boldsymbol{x}^J)$ で表される．半教師あり学習では，ラベルありデータ集合 $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ とラベルなしデータ集合 $D_u = \{\boldsymbol{x}_m\}_{m=1}^M (M >> N)$ からなる訓練データ集合 $D = \{D_l, D_u\}$ を用いて分類器を学習する．

教師あり学習の枠組みでは，MCD の分類問題に対して，確率モデルアプローチに基づく生成，識別，ハイブリッドの各分類器が提案されている．本章では，これらの分類器に，生成モデル，識別の各アプローチに基づいて提案されてきた半教師あり学習法[9),16)] を適用する方法を述べる．

2.2 生成分類器の半教師あり学習

生成分類器は，データ $\boldsymbol{x}$ とクラスラベル y の同時確率密度 $p(\boldsymbol{x}, y)$ のモデルを用いて設計される．しかし，異なる種類のメディアから構成されるデータを直接モデル化することは容易ではない．そこで，それぞれのクラスで各構成要素は “独立” に生成されると仮定し，構成要素ごとに設計される生成モデル（以下，構成要素モデルと呼ぶ）$p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)$ を用いて同時確率密度を $p(\boldsymbol{x}, k; \Theta) = P(k)\prod_{j=1}^J p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)$ のようにモデル化する[2]．ここで，$\Theta = \{\boldsymbol{\theta}_k^j\}_{j,k}$ はクラス k における構成要素 j のモデルパラメータの集合を表す．訓練データを用いて Θ の推定値 $\hat{\Theta} = \{\hat{\boldsymbol{\theta}}_k^j\}_{j,k}$ を求めた後，データのクラス事後確率を，ベイズ則により，

$$P(y=k|\boldsymbol{x};\hat{\Theta}) = \frac{P(k)\prod_{j=1}^J p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j)}{\sum_{k'=1}^K P(k')\prod_{j=1}^J p(\boldsymbol{x}^j|k';\hat{\boldsymbol{\theta}}_{k'}^j)} \quad (1)$$

で与える．生成分類器は，データが属するクラスがクラス事後確率を最大にする k であると推定する．

生成分類器の半教師あり学習は，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見なし，混合モデルを仮定することで実現される[16)]．すなわち，ラベルなしデータの確率密度を，$p(\boldsymbol{x};\Theta) = \sum_{k=1}^K p(\boldsymbol{x}, k; \Theta)$

で表される分布でモデル化する．そして，訓練データ集合 D が与えられた下で，Θ の事後確率密度 $p(\Theta|D)$ を最大化させる値をパラメータ Θ の推定値として求める（MAP 推定，文献 21) 参照）．すなわち，ベイズ則により，$\log p(\Theta|D)$ に相当する目的関数：

$$\begin{aligned} F(\Theta) = &\sum_{n=1}^{N} \log P(y_n) \prod_{j=1}^{J} p(\boldsymbol{x}_n^j | y_n; \boldsymbol{\theta}_{y_n}^j) \\ &+ \sum_{m=1}^{M} \log \sum_{k=1}^{K} P(k) \prod_{j=1}^{J} p(\boldsymbol{x}_m^j | k; \boldsymbol{\theta}_k^j) \\ &+ \log p(\Theta) \end{aligned} \tag{2}$$

を最大化させる Θ を求める．ここで，$p(\Theta)$ は Θ の事前確率分布を表す．$F(\Theta)$ を最大化させる Θ は期待値最大化（EM）アルゴリズム[5] を用いて求められる．

Θ の推定は，N と M の大きさに影響される．$M >> N$ のとき，Θ の推定は，式 (2) の第 2 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．それゆえ，混合モデルの仮定が実データ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．この問題に対処すべく，ラベルなしデータのパラメータ学習への寄与を調整する重みパラメータ λ を導入する方法（EM-λ）が提案されている[16]．式 (2) の第 2 項に $\lambda \in [0, 1]$ を乗ずることで，EM-λ の目的関数が得られる．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，汎化性能の高いパラメータ Θ の推定値を与える λ を探索する．探索された λ を用いてパラメータ Θ を推定する．

2.3 識別分類器の半教師あり学習

識別分類器は，構成要素ごとにデータを分離せずに，データのクラス事後確率 $P(k|\boldsymbol{x})$ を直接モデル化して構築できる．たとえば，多項ロジスティック回帰（Multinomial Logistic Regression，MLR）モデル[10] を適用して，以下のように未知パラメータ集合 $W = \{\boldsymbol{w}_k\}_{k=1}^K$ を用いてクラス事後確率分布をモデル化できる．

$$P(y = k|\boldsymbol{x}; W) = \frac{\exp(\boldsymbol{w}_k \cdot \boldsymbol{x})}{\sum_{k'=1}^{K} \exp(\boldsymbol{w}_{k'} \cdot \boldsymbol{x})} \tag{3}$$

ここで，$\boldsymbol{w}_k \cdot \boldsymbol{x}$ は $\boldsymbol{w}_k$ と $\boldsymbol{x}$ の内積を表す．特徴ベクトル $\boldsymbol{x}$ に対して最大エントロピー（ME）原理を適用[15] しても同様に，式 (3) で表されるクラス事後確率モデルを得ることができる．

MLR の半教師あり学習を実現する一手法として，最小エントロピー正則化項（Mimimum Entropy Regularizer，MER)[9] が提案されている．この方法では，ラベルなしデータのクラス事後確率のエントロピーをクラス分離度の尺度として用い，こ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 MLR を学習する．MER を用いた MLR の学習では，以下の目的関数を最大化させる W をパラメータの推定値とする．

$$\begin{aligned} F(W) = &\sum_{n=1}^{N} \log P(y_n|\boldsymbol{x}_n; W) \\ &+ \lambda \sum_{m=1}^{M} \sum_{k=1}^{K} P(k|\boldsymbol{x}_m; W) \log P(k|\boldsymbol{x}_m; W) \\ &+ \log p(W) \end{aligned} \tag{4}$$

ここで，λ は重みパラメータであり，$p(W)$ は W の事前確率分布を表す．MLR の学習では，$p(W)$ としてガウス事前確率分布[4] $p(W) \propto \prod_{k=1}^{K} \prod_{l=1}^{L} \exp\left(-w_{kl}^2 / 2\sigma^2\right)$ を用いることができる．L は特徴ベクトル $\boldsymbol{x}$ とパラメータ $\boldsymbol{w}_k$ の次元数を表し，σ は学習の際に値を設定すべきハイパーパラメータである．本論文では，MER を用いた学習により得られる MLR を MLR/MER と呼ぶ．

2.4 ハイブリッド分類器への半教師あり学習法の単純な適用

ハイブリッド分類器では，クラスごとに構成要素モデル $p(\boldsymbol{x}^j|k; \boldsymbol{\theta}_k^j)$ を仮定し，訓練データにより学習された構成要素モデルを識別学習で得られる重みを用いて統合することでクラス事後確率分布をモデル化する[6),17]．すなわち，生成モデルアプローチによる構成要素のモデル化と，識別アプローチによるクラス事後確率のモデル化とのハイブリッドにより分類器を構築する．多クラス単一ラベル分類問題では，ME 原理から得られる以下の式を用いて，ハイブリッド分類器のクラス事後確率分布を与えることができる[6]．

$$\begin{aligned} &R(y = k|\boldsymbol{x}; \hat{\Theta}, \Gamma) \\ &= \frac{\exp\left\{\mu_k + \sum_{j=1}^{J} \gamma_j \log p(\boldsymbol{x}^j|k; \hat{\boldsymbol{\theta}}_k^j)\right\}}{\sum_{k'=1}^{K} \exp\left\{\mu_{k'} + \sum_{j=1}^{J} \gamma_j \log p(\boldsymbol{x}^j|k'; \hat{\boldsymbol{\theta}}_{k'}^j)\right\}} \\ &= \frac{e^{\mu_k} \prod_{j=1}^{J} p(\boldsymbol{x}^j|k; \hat{\boldsymbol{\theta}}_k^j)^{\gamma_j}}{\sum_{k'=1}^{K} e^{\mu_{k'}} \prod_{j=1}^{J} p(\boldsymbol{x}^j|k'; \hat{\boldsymbol{\theta}}_{k'}^j)^{\gamma_j}} \end{aligned} \tag{5}$$

ここで，$\hat{\Theta} = \{\hat{\boldsymbol{\theta}}_{j,k}\}_{j,k}$ は構成要素モデルのパラメータ推定値の集合を表す．$\Gamma = (\{\gamma_j\}_{j=1}^J, \{\mu_k\}_{k=1}^K)$ は，構成要素の統合の重みとクラスの出現の偏りを与えるパラメータである．式 (5) の 2 行目で示されるように，ハイブリッド分類器のクラス事後確率分布は，式 (3) の右辺で $\boldsymbol{x}$ に代えて構成要素モデルの対数尤度 $\log p(\boldsymbol{x}^j|k : \boldsymbol{\theta}_k^j)$ と 1 からなるベクト

ルを適用することで得られる分布 $P_k = \exp\{w_{k0} + \sum_{j=1}^{J} w_{kj} \log p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)\}/Z$（$Z$ は $\sum_{k=1}^{K} P_k = 1$ とするための正規化項）と類似の分布型を持つ．ただし，式 (3) から誘導される分布ではクラスごとに各構成要素の重み w_{kj} を与えるのに対し，式 (5) では構成要素ごとに 1 つの重み γ_j を与えている点が異なる．

Γ は，分類器の予測精度が最大になることを期待して，ラベルありデータの leave-one-out 交差検定法で得られるクラス事後確率の対数尤度と Γ の対数事前確率分布の和の最大化により推定する[6),17)]．この推定方法により，各構成要素にクラスラベルの予測能力に応じた統合の重みを与えることが期待できる．汎化性能が低い構成要素に小さな重みを与えることで，過学習の抑制を期待できる．

ハイブリッド分類器にラベルなしデータの情報を反映させる単純な手法として，2.2 節で述べた EM-λ を用いて構成要素モデル $p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)$ のパラメータ $\boldsymbol{\theta}_k^j$ を学習させることが考えられる．この方法では，ラベルあり・なしデータを用いて学習した構成要素モデルを，ラベルありデータのみで識別学習した重みで統合することで分類器を得る．すなわち，分類器の半教師あり学習を，生成モデルアプローチのみに基づいて行う．本論文では，この方法をカスケードハイブリッド法（CH）と呼ぶ．

3. 提 案 法

本論文では，MCD 分類器の半教師あり学習の問題に対して，生成モデル・識別の両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．提案法では，2.4 節で述べたハイブリッド分類器において，生成モデルと識別学習の双方の利点を活かした半教師あり学習を実現するために，最近，筆者らが考案した偏り補正モデル（Bias Correction Model，BCM）[7)] を導入して分類器を構築する．本論文では，便宜上，提案法を H-BCM（Hybrid with Bias Correction Models）と呼ぶ．以下に，H-BCM の定式化とパラメータ学習法について述べる．

3.1 構成要素モデルと偏り補正

H-BCM では，まず，クラスごとに構成要素モデル $p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)$ を設計し，ラベルありデータ集合 D_l を用いて学習する．ここで，$\Theta = \{\boldsymbol{\theta}_k^j\}_{j,k}$ をクラス k における構成要素 j のモデルパラメータ $\boldsymbol{\theta}_k^j$ の集合とする．構成要素モデルのパラメータは，MAP 推定に基づいて構成要素ごとに個々に学習する．具体的には，以下の目的関数 $F_1^j(\{\boldsymbol{\theta}_k^j\}_k)$ を最大化する $\{\boldsymbol{\theta}_k^j\}_k$ をモデルパラメータの推定値とする．

$$F_1^j(\{\boldsymbol{\theta}_k^j\}_k) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n^j|y_n;\boldsymbol{\theta}_{y_n}^j) + \sum_{k=1}^{K} \log p(\boldsymbol{\theta}_k^j),\ \forall j \quad (6)$$

ここで，$p(\boldsymbol{\theta}_k^j)$ は $\boldsymbol{\theta}_k^j$ の事前確率分布を表す．

パラメータ学習に十分な数のラベルありデータが与えられていない場合，パラメータの推定値 $\hat{\Theta} = \{\hat{\boldsymbol{\theta}}_k^j\}_{j,k}$ は，真のデータ分布の最適な近似を与えるパラメータ値から大きく外れる危険性がある．すなわち，少数のラベルありデータで学習される構成要素モデルは，しばしば高い偏りを持つ．この偏りが分類性能に与える悪影響を緩和するため，各クラスの構成要素ごとに偏り補正モデル $p(\boldsymbol{x}^j|k;\boldsymbol{\psi}_k^j)$ を導入する．偏り補正モデルには，構成要素モデルと同型の分布（パラメータ値は異なる）を与える．$\Psi = \{\boldsymbol{\psi}_k^j\}_{j,k}$ を偏り補正モデルのパラメータ集合とする．H-BCM では，構成要素モデルと偏り補正モデルを識別アプローチに基づいて統合することで分類器を構築する．

3.2 クラス事後確率分布

H-BCM では，構成要素モデルと偏り補正モデルの識別アプローチに基づく統合により得られる分布でクラス事後確率を定義する．この分布は，最大エントロピー（ME）原理[1)] に基づき，構成要素モデルと偏り補正モデルに関する制約を満たす，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としてモデル化（以下，ME モデルと呼ぶ）される[7)]．

ME モデル $R(k|\boldsymbol{x})$ 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，ラベルありデータの経験分布 $\hat{p}(\boldsymbol{x},k) = \sum_{n=1}^{N} I_{\boldsymbol{x}_n}(\boldsymbol{x}) I_{y_n}(k)/N$ 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 $\hat{p}(\boldsymbol{x}) = \sum_{n=1}^{N} I_{\boldsymbol{x}_n}(\boldsymbol{x})/N$ を用いて，以下の式で表される．

$$\sum_{\boldsymbol{x},k} \hat{p}(\boldsymbol{x},k) \log p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j) = \sum_{\boldsymbol{x},k} \hat{p}(\boldsymbol{x}) R(k|\boldsymbol{x}) \log p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j), \forall j \quad (7)$$

ここで，$I_{\boldsymbol{x}_n}(\boldsymbol{x})$ は，$\boldsymbol{x} = \boldsymbol{x}_n$ のときに 1，それ以外のときに 0 である指示関数を表す．偏り補正モデル $p(\boldsymbol{x}^j|k;\boldsymbol{\psi}_k^j)$ に関する制約もまた，式 (7) と同様の形で与える．さらに，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，ラベルあり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．

$$\sum_{\boldsymbol{x}} \hat{p}(\boldsymbol{x},k) = \sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x}) R(k|\boldsymbol{x}),\ \forall k \tag{8}$$

以上の制約の下で，クラス事後確率分布のエントロピー $H(R) = -\sum_{\boldsymbol{x},k} \hat{p}(\boldsymbol{x}) R(k|\boldsymbol{x}) \log R(k|\boldsymbol{x})$ を最大化させることにより，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：

$$R(y=k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma) = \frac{1}{Z} e^{\mu_k} \prod_{j=1}^{J} p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j)^{\gamma_{1j}} p(\boldsymbol{x}^j|k;\boldsymbol{\psi}_k^j)^{\gamma_{2j}} \tag{9}$$

を導出できる．ここで，Z は $\sum_{k=1}^{K} R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma) = 1$ とするための正規化項であり，$Z = \sum_{k'=1}^{K} \{e^{\mu_{k'}} \times \prod_{j=1}^{J} p(\boldsymbol{x}^j|k';\hat{\boldsymbol{\theta}}_{k'}^j)^{\gamma_{1j}} p(\boldsymbol{x}^j|k';\boldsymbol{\psi}_{k'}^j)^{\gamma_{2j}}\}$ である．また，$\Gamma = (\{\gamma_{1j},\gamma_{2j}\}_{j=1}^{J}, \{\mu_k\}_{k=1}^{K})$ はラグランジュ乗数であり，γ_{1j} と γ_{2j} は構成要素モデルと偏り補正モデルの統合の重みを，μ_k はクラス k の出現の偏りを与える．H-BCM では，構成要素モデルと偏り補正モデルの ME 原理に基づく統合により得られる $R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ を，分類器を表すクラス事後確率分布とする．

ME 原理では，制約を満たすエントロピー最大のクラス事後確率分布を与える ME モデル $R(k|\boldsymbol{x};\Gamma)$ のラグランジュ乗数 Γ の値は，$R(k|\boldsymbol{x};\Gamma)$ に関する経験分布 $\hat{p}(\boldsymbol{x},k)$ の対数尤度を最大化させる Γ の値と一致することが証明されている（文献 1)，文献 11) の 6.2 節参照）．このため，Ψ の値を与えると，ラベルありデータ集合 D_l を用いた最尤推定により式 (9) 内の Γ の推定値を求められる．しかし，3.1 節で述べたように，$\hat{\Theta}$ もまた D_l を用いて得られる推定値であり，Γ の推定に同じ D_l を用いることで，構成要素モデルの統合の重み $\{\gamma_1^j\}_j$ を過剰に大きく推定する危険性がある[17]．また，最尤推定には，一般に，ラベルありデータが少数のときに過学習を引き起こす危険性がある．これら 2 つの問題を緩和するために，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 $F_2(\Gamma|\Psi)$ を最大化させる Γ を計算する．

$$F_2(\Gamma|\Psi) = \sum_{n=1}^{N} \log R(y_n|\boldsymbol{x}_n;\hat{\Theta}^{(-n)},\Psi,\Gamma) + \log p(\Gamma) \tag{10}$$

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構成要素モデルのパラメータ値を表す．$p(\Gamma)$ は Γ の事前確率分布を表し，ガウス事前確率分布[4] を用いて

$$p(\Gamma) \propto \prod_{j=1}^{J}\prod_{l=1}^{2} \exp\left\{-\frac{(\gamma_{lj}-a_l)^2}{2\sigma_l^2}\right\} \times \prod_{k=1}^{K} \exp\left(-\frac{\mu_k^2}{2\rho^2}\right) \tag{11}$$

とする．$F_2(\Gamma|\Psi)$ を最大化させる Γ は，Ψ の値を与えると，準ニュートン法の一種である L-BFGS アルゴリズム[12] 等の勾配法を用いて探索できる．$F_2(\Gamma|\Psi)$ は Γ に関して上に凸の関数である（付録 A.1 節参照）ため，Γ の解の探索では大域的収束が保証される．$\gamma_{2j} < 0$ は偏り補正モデルの導入による分類性能の向上を期待できないことを意味するので，$\gamma_{2j} \geq 0,\ \forall j$ の領域で $F_2(\Gamma|\Psi)$ を最大化させる Γ を求める．

3.3 偏り補正モデルの学習

偏り補正モデルのパラメータ Ψ の学習は，ラベルありデータが十分に与えられないことに起因する分類器 $R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ の学習の偏りを緩和することを目的とする．H-BCM では，$R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ により与えられる分類器の識別関数：

$$g_k(\boldsymbol{x};\boldsymbol{\psi}_k) = e^{\mu_k} \prod_{j=1}^{j} p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j)^{\gamma_{1j}} p(\boldsymbol{x}^j|k;\boldsymbol{\psi}_k^j)^{\gamma_{2j}} \tag{12}$$

の値が最大である k をデータ $\boldsymbol{x}$ のクラスラベル y として決定する．このとき，すべてのクラスで $g_k(\boldsymbol{x};\boldsymbol{\psi}_k)$ の値が小さい $\boldsymbol{x}$ では，クラス間での $g_k(\boldsymbol{x};\boldsymbol{\psi}_k)$ の差が小さく，$R(k|\boldsymbol{x};\hat{\Theta},\Psi,\Gamma)$ 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データ $\boldsymbol{x}$ に対し，クラス間での $g_k(\boldsymbol{x};\psi_k)$ の差が増大することを期待して，識別関数の総和 $\sum_{k=1}^{K} g_k(\boldsymbol{x};\psi_k)$ 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 $F_3(\Psi|\Gamma)$ の最大化により Ψ を推定する．

$$F_3(\Psi|\Gamma) = \sum_{m=1}^{M} \log \sum_{k=1}^{K} g_k(\boldsymbol{x}_m;\boldsymbol{\psi}_k) + \log p(\Psi) \tag{13}$$

ここで，$p(\Psi)$ はパラメータ Ψ の事前確率分布を表す．$F_3(\Psi|\Gamma)$ の最大化による Ψ の推定は，2.2 節で述べた生成モデルアプローチにおける混合モデルの尤度最大化に基づくパラメータ推定と類似する．生成分類器の識別関数は $p(\boldsymbol{x}|k;\Theta)p(k)$ であり，識別関数の総和は混合モデルと一致する．H-BCM では，生成モデルアプローチと類似した手法で偏り補正モデルを学習することで，大量のラベルなしデータにより与えら

れる大域的なデータの分布特性を分類器の識別関数に反映させる．

Γ の値が既知のとき，EM アルゴリズム[5] のような反復計算を行うことで，$F_3(\Psi|\Gamma)$ を初期値近傍で最大化させる Ψ の局所解を推定できる[7]．反復計算では，(t) ステップでの Ψ の推定値を $\Psi^{(t)} = \{\boldsymbol{\psi}_k^{j,(t)}\}_{j,k}$ とするとき，目的関数：

$$Q(\Psi^{(t+1)}, \Psi^{(t)}|\Gamma) = \sum_{j=1}^{J} \gamma_{2j} \sum_{m=1}^{M} \sum_{k=1}^{K} \Big\{ R(k|\boldsymbol{x}_m; \hat{\Theta}, \Psi^{(t)}, \Gamma) \times \log p(\boldsymbol{x}_m^j|k; \boldsymbol{\psi}_k^{j,(t+1)}) \Big\} + \log p(\Psi^{(t+1)}) \tag{14}$$

の最大化により，$(t+1)$ ステップでの推定値 $\Psi^{(t+1)}$ を求められる．Γ の値が既知のとき，反復計算の各ステップでは，式 (14) の $R(k|\boldsymbol{x}_m; \hat{\Theta}, \Psi^{(t)}, \Gamma)$ と γ_{2j} の値は既知であるため，式 (6) による構成要素モデルのパラメータ推定と同様の計算方法で推定値 $\Psi^{(t+1)}$ を得られる．$\Psi^{(t+1)}$ 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 の推定値を得られる．

しかし，前節で述べたように，Γ は Θ と Ψ が与えられたときに式 (10) を用いて学習されるパラメータである．一方，式 (14) 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 (10) により Γ を推定する．得られた推定値 $\Gamma^{(0)}$ と $\Psi^{(0)}$ を用いて，式 (14) により Ψ 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り返して，$F_2(\Gamma|\Psi)$ と $F_3(\Psi|\Gamma)$ を同時に最大化させる Γ と Ψ の解を探索する．ただし，$F_3(\Psi|\Gamma)$ は Ψ に関して上に凸の関数であるとは限らないため，この反復計算ではパラメータ値が収束する保証はない．そこで，パラメータ値が収束しない場合，一定回数を超えた時点で反復計算を打ち切り，その時点で得られたパラメータ値を Γ と Ψ の推定値とする．以上のパラメータ学習のアルゴリズムを**図 1** にまとめる．

訓練データ集合：
$D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$
1. 初期化：$\boldsymbol{\psi}_k^{(0)}$, $\forall k$ の設定，$t \leftarrow 0$.
2. 式 (6) 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 (10) により $\Gamma^{(0)}$ を計算
4. 反復計算：$\sum_{j,k} ||\boldsymbol{\psi}_k^{j,(t+1)} - \boldsymbol{\psi}_k^{j,(t)}||/||\boldsymbol{\psi}_k^{j,(t)}|| + ||\Gamma^{(t+1)} - \Gamma^{(t)}||/||\Gamma^{(t)}|| < \epsilon$ または $t > t_{max}$ であれば反復計算を終了 ($||\cdot||$ は $\cdot$ の L2 ノルムを表す).
   - $\Gamma^{(t)}$ と $\hat{\Theta}$ の下で式 (14) により $\Psi^{(t+1)}$ を計算
   - $\Psi^{(t+1)}$ と $\hat{\Theta}^{(-n)}$ の下で式 (10) 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 1 パラメータ学習アルゴリズム
Fig. 1 Algorithm for parameter estimation.

3.4 H-BCM への NB モデルの適用

H-BCM をテキストとリンク情報からなるデータの分類に適用するため，構成要素 j を特徴ベクトル $\boldsymbol{x}^j = (x_1^j, \ldots, x_i^j, \ldots, x_{V_j}^j)$ で表し，ナイーブベイズ（NB）モデル[21] を用いて構成要素の生成確率をモデル化する．ここで，x_i^j は i 番目の特徴量を表し，V_j は構成要素 j に含まれる特徴量の数を表す．構成要素 j がテキスト（リンク情報）の場合，x_i^j を単語（URL）i の出現頻度，V_j を構成要素 j に含まれる単語（URL）の種類の総数として特徴ベクトルを与える．NB モデルでは，クラスが与えられた条件下では各々の単語（URL）が独立に生起すると仮定し，クラスラベルが k であるデータの構成要素 j の生成確率を

$$p(\boldsymbol{x}^j|k; \boldsymbol{\theta}_k^j) \propto \prod_{i=1}^{V_j} \left(\theta_{ki}^j\right)^{x_i^j} \tag{15}$$

で表される多項分布でモデル化する．式 (15) 中の θ_{ki}^j は構成要素 j のクラス k における単語（URL）i の生起確率を表す未知パラメータであり，$\sum_{i=1}^{V_j} \theta_{ki}^j = 1$, $\theta_{ki}^j > 0$, $\forall i$ の制約を持つ．H-BCM では，偏り補正モデル $p(\boldsymbol{x}^j|k; \boldsymbol{\psi}_k^j)$ にも同様に NB モデルを適用する．

H-BCM では，構成要素モデルと偏り補正モデルに NB モデルを適用するのに際して，構成要素の特徴ベクトル $\boldsymbol{x}^j$ を $|\boldsymbol{x}^j| = \sum_{i=1}^{V_j} x_i^j = 1$ となるように正規化する．式 (6) による構成要素モデルのパラメータ学習では，パラメータの事前確率分布 $p(\boldsymbol{\theta}_k^j)$ として，NB モデルでしばしば仮定されるディリクレ分布[21] $p(\boldsymbol{\theta}_k^j) \propto \prod_{i=1}^{V_j} (\theta_{ki}^j)^{\xi_k^j - 1}$ を用いる．ξ_k^j はハイパーパラメータである．式 (14) による偏り補正モデルのパラメータ学習においても同様に，パラメータ Ψ の事前確率分布としてディリクレ分布を用いる．H-BCM では，学習された構成要素モデルが学習に用いなかったデータに対して大きな尤度を与えることを期待して，ξ_k^j の値を，ラベルありデータ $(\boldsymbol{x}_n, y_n) \in D_l$ の leave-one-out 交差検定法により推定される構成要素モデルの対数尤度の和 $L = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n^j|y_n; \boldsymbol{\theta}_{y_n}^{j,(-n)})$ を最大にする値に調節する．この調節のように，訓練データを用いてハイパーパラメータ値を最適化することは，経験ベイズ法[8] で提案されている．H-BCM では，L を最大化させる ξ_k^j の値を，EM アルゴリズム[5]

を援用して効率的に探索する．

4. 評価実験

4.1 テストコレクション

評価実験には，テキスト分類問題のベンチマークテストにしばしば用いられる WebKB☆と Cora☆☆，20newsgroups (20news)☆☆☆の 3 つのテストコレクションを用いた．

WebKB は大学の Web ページを集めたものであり，7 つのカテゴリに分類されている．文献 15) の設定に従い，*student*，*faculty*，*course*，*project* の 4 つのカテゴリに含まれる 4,199 の Web ページを実験に用いた．Web ページの特徴ベクトルは，本文（MT），他のページへのリンク（OL），他のページからのリンク（IL），アンカーテキスト（AT）の 4 つの構成要素を抽出して作成した．IL はこのテストコレクションに含まれる Web ページのみから抽出し，AT として当該ページをリンクしている他のページが参照に用いているアンカーテキストを抽出した．MT と AT では単語の出現頻度，IL と OL では URL の出現頻度により構成要素の特徴ベクトルを作成した．特徴ベクトルの作成に際して，MT と AT では冠詞等の文書を特徴づける効果を持たない停止語（stop words）[18] と 1 つのページにしか出現しない低頻度語彙を除外し，IL と OL では 1 つのページにしか出現しない低頻度 URL を除外した．特徴ベクトルの次元は，それぞれ 18,525（MT），496（AT），4,131（OL），500（IL）であった．

Cora は 3 万以上の技術論文の概要と引用情報を集めたものであり，70 カテゴリのいずれかに分類されている．実験では，*/Artificial_Intelligence/Machine_Learning/** の 7 つのカテゴリに属する 4,240 論文を用いた．論文の特徴ベクトルは，本文（MT），著者名（AU），引用論文（CI）の 3 つの構成要素を抽出して作成した．MT では単語の出現頻度を，AU と CI ではそれぞれ著者名と引用論文の出現を表す特徴ベクトルを作成した．WebKB と同様に停止語と低頻度の特徴を除外した結果，特徴ベクトルの次元はそれぞれ 9,190（MT），1,495（AU），13,282（CI）であった．

20news は，UseNet の記事を集めたものであり，20 グループに分類されている．文献 15) の設定に従い，20 グループのうち，*comp.** の 5 つのグループに属する 4,881 記事を実験に用いた．構成要素として，各記事から “Subject:” のあとに続くタイトル（T）とそれ以外の本文（M）を抽出した．M と T の特徴ベクトルは単語の出現頻度により作成した．WebKB と同様に停止語と低頻度語彙を除外した結果，特徴ベクトルの次元は 19,273（M）と 1,775（T）であった．

分類実験では，H-BCM を適用するのに際して，各テストコレクションのすべての構成要素に対して，NB モデルを用いた．

4.2 従来の半教師あり学習法の応用との比較

4.2.1 実験方法

H-BCM の性能を評価するため，2 章で述べた従来の半教師あり学習法の応用に基づく 3 つの手法（EM-λ，MLR/MER，CH）との比較実験を行った．EM-λ と CH の構成要素モデルには NB モデルを適用した．また，NB モデルと MLR モデル，ハイブリッド法（HY）による分類器をラベルありデータのみで学習した場合の性能も調べた．比較手法の概要を**表 1** にまとめる．

性能比較には，各テストコレクションで，ラベルあり・なしデータとテストデータをランダムに選択して行った 10 回の実験の結果から算出される分類精度の平均値を用いた．分類精度は，テストデータの総数に対する，クラスラベルを正しく予測できたテストデータの個数の比率として求められる．実験に用いたテストデータ数は各コレクションで 1,000，訓練に用いたラベルなしデータ数 M は 2,500（WebKB），2,000（Cora），2,500（20news）である．実験では，ラベルありデータ数 N を変えて分類器を学習し，N ごとに性能比較を行った．

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値として，$\{0.01, 0.05, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 1.0\}$ の 14 種類の値を用いた．パラメータ値は，文献 16) で提案された方法に従って，ラベルありデータの leave-one-out 交差検定法による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 の値を $\{\{0.1 \times 10^{-n}, 0.2 \times 10^{-n}, 0.5 \times 10^{-n}\}_{n=0}^{4}, 1\}$ 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[9] 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，テストデータの分類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．

各分類器の学習に際して，ディリクレ事前確率分布と

☆ http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.gtar.gz

☆☆ http://www.cs.umass.edu/~mccallum/data/cora-classify.tar.gz

☆☆☆ http://people.csail.mit.edu/jrennie/20Newsgroups/20news-18828.tar.gz

表 1 比較手法の概要
Table 1 Outline of compared methods.

略記名	手法の名前	クラス事後確率のモデル化方法
H-BCM	提案法（偏り補正モデルを用いたハイブリッド）	構成要素・偏り補正モデル（NB モデル）の ME 原理に基づく統合によりモデル化（3 章参照）
CH	カスケードハイブリッド	EM-λ で学習した構成要素モデル（NB モデル）の ME 原理に基づく統合によりモデル化（2.4 節参照）
EM-λ	EM-λ	EM-λ で学習した構成要素モデル（NB モデル）を用いてベイズ則に基づきモデル化（2.2 節参照）
MLR/MER	最小エントロピー正則化項による多項ロジステック回帰	多項ロジステック回帰でモデル化し，最小エントロピー正則化項を用いて学習（2.3 節参照）
HY	ハイブリッド	ラベルありデータで学習した構成要素モデル（NB モデル）の ME 原理に基づく統合によりモデル化（2.4 節参照）
NB	ナイーブベイズ	ラベルありデータで学習した構成要素モデル（NB モデル）を用いてベイズ則に基づきモデル化（2.2 節参照）
MLR	多項ロジステック回帰	多項ロジステック回帰でモデル化し，ラベルありデータを用いて学習（2.3 節参照）

ガウス事前確率分布のハイパーパラメータを次のように設定した．CH と EM-λ，NB では，NB モデルパラメータ Θ のディリクレ事前確率分布 $p(\Theta) \propto \prod_{i,j,k} \left(\theta_{ki}^{j}\right)^{\eta}$ のハイパーパラメータ η の値を $\{\{0.1 \times 10^{-n}, 0.2 \times 10^{-n}, 0.5 \times 10^{-n}\}_{n=1}^{3}, 1\}$ の 10 候補から選択した．MLR/MER と MLR では，2.3 節で述べたガウス事前確率分布のハイパーパラメータ σ の値を $\{10^{n/2}\}_{n=1}^{7}$ の 7 候補から選択した．H-BCM では，Ψ のディリクレ事前確率分布のハイパーパラメータ η の値を $\{0.1 \times 10^{-n}, 0.2 \times 10^{-n}, 0.5 \times 10^{-n}\}_{n=2}^{4}$ の 9 候補から選択した．さらに，式 (11) で示した H-BCM のガウス事前確率分布のハイパーパラメータ ρ の値を $\{10^{n/2}\}_{n=-3}^{3}$ の 7 候補から選択し，他のハイパーパラメータ値を $\sigma_l = 1$，$a_l = 1$ に固定した．CH と HY においても，Γ の学習に際して，H-BCM と同様にガウス事前確率分布を用いてハイパーパラメータの値を設定した．各手法の比較は，コレクションごとに，各ハイパーパラメータを候補内の 1 つの値に固定したときに得られる分類精度で行った．比較では，各コレクションでテストデータの分類精度の平均値を最も高くするハイパーパラメータ値をそれぞれの手法で用いた．

4.2.2 実験結果と考察

表 2 に，各テストコレクションで，ラベルありデータ数 N を変えてそれぞれ 10 回の実験を行ったときの各手法の分類精度の平均値を示す．括弧内の数値は標準偏差を示す．表 2 より，WebKB の $N = 128$, 256, 512 と 20news の $N = 320$, 640, 1280 のときに，MLR/MER の分類精度が EM-λ よりも高かった．しかし，4.2.1 項で述べたように，MLR/MER の学習に際してテストデータの分類精度を最も高くする重みパラメータの候補値を選択したことが原因で，MLR/MER の分類精度の方が高くなった可能性がある．そこで，MLR/MER で重みパラメータの値を固定したときに得られる分類精度の平均値も参考のために調べた．重みパラメータの 16 候補値に対して，WebKB では 48.1–78.3（$N = 128$），50.4–84.0（$N = 256$），66.5–88.7（$N = 512$），20news では 64.5–72.2（$N = 320$），74.0–77.6（$N = 640$），78.3–80.7（$N = 1024$）の範囲で分類精度の平均値が変動した．

表 2 より，教師あり学習の場合では，文献 6)，17) で報告されているように，ハイブリッド分類器が MCD 分類の高精度化に有用であることが確認された．実験では，WebKB の $N = 32$，512 の場合を除いて，HY の分類精度が NB と MLR よりも高かった．

半教師あり学習の場合では，EM-λ，MLR/MER，CH，H-BCM の 4 つの手法の分類精度を調べた結果，WebKB で EM-λ と MLR/MER の分類精度の差が大きな場合を除いて，H-BCM の性能が他の手法を上回っていた．この結果より，MCD の分類問題において，ラベルなしデータを学習に用いて分類器の汎化性能を向上させるのに H-BCM が有用であるといえる．

H-BCM の性能は，MLR/MER と比較して，WebKB でラベルありデータ数が多数の場合を除いて高かった．この結果は，MLR/MER に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一方，H-BCM では，過学習を抑える生成モデルの特徴を有しているため，MLR/MER よりも高い分類精度が得られたと考えられる．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，MLR/MER は提案法よりも高い分類精度を与えると考えられる．

EM-λ と比較して，H-BCM は，WebKB でラベルありデータ数が少数の場合を除いて高い性能を示した．一般的に，大量のラベルありデータが与えられる場合，識別アプローチは生成モデルアプローチよりも高い分類精度を与えることが知られている[14]．この実験結

表 2 異なるラベルありデータ数 N に対する H-BCM と比較手法の分類精度 (%)
Table 2 Classification accuracies (%) with H-BCM and compared methods over various number of labeled data, N.

(a) WebKB（$M = 2500$, $K = 4$）

Training set		Semi-supervised				Supervised		
N	N/M	H-BCM	CH	EM-λ	MLR/MER	HY	NB	MLR
32	0.0128	69.9 (5.1)	**73.5** (4.1)	73.4 (2.8)	63.5 (5.0)	68.6 (3.6)	69.1 (2.9)	63.3 (4.5)
64	0.0256	76.2 (2.8)	**77.5** (2.3)	75.8 (1.7)	72.2 (2.7)	74.9 (1.1)	73.9 (1.1)	71.9 (2.2)
128	0.0512	**81.8** (1.7)	80.2 (2.0)	78.3 (2.4)	78.5 (2.2)	80.7 (1.9)	77.7 (1.8)	78.3 (2.0)
256	0.1024	**85.0** (1.3)	83.2 (1.5)	80.8 (1.6)	84.3 (1.7)	84.6 (1.3)	80.9 (1.2)	83.9 (1.7)
512	0.2048	87.2 (1.1)	84.8 (1.5)	82.4 (1.3)	**88.7** (1.3)	87.4 (1.1)	83.5 (1.3)	87.9 (1.3)

(b) Cora（$M = 2000$, $K = 7$）

Training set		Semi-supervised				Supervised		
N	N/M	H-BCM	CH	EM-λ	MLR/MER	HY	NB	MLR
56	0.028	**77.4** (3.5)	74.3 (4.9)	71.8 (4.1)	55.6 (4.4)	63.2 (3.5)	58.0 (2.5)	55.2 (3.8)
112	0.056	**83.4** (1.7)	80.7 (1.4)	77.9 (1.6)	63.9 (2.3)	72.6 (1.8)	66.7 (1.4)	63.5 (2.1)
224	0.112	**85.8** (0.9)	84.1 (0.7)	81.2 (1.1)	73.0 (1.0)	80.7 (0.8)	75.4 (1.4)	72.3 (1.1)
448	0.224	**87.3** (1.6)	86.1 (1.3)	83.3 (1.1)	78.8 (1.0)	84.7 (0.8)	80.1 (1.1)	77.3 (1.2)
896	0.448	**89.1** (0.6)	88.5 (1.2)	85.9 (1.0)	82.7 (1.0)	87.9 (1.1)	84.4 (0.8)	82.1 (1.2)

(c) 20news（$M = 2500$, $K = 5$）

Training set		Semi-supervised				Supervised		
N	N/M	H-BCM	CH	EM-λ	MLR/MER	HY	NB	MLR
40	0.016	**64.8** (3.0)	56.2 (4.6)	56.6 (6.9)	50.1 (5.1)	47.7 (3.7)	46.0 (3.9)	48.8 (3.9)
80	0.032	**71.6** (1.8)	63.7 (2.2)	60.8 (3.8)	57.3 (3.7)	56.3 (2.2)	51.8 (3.3)	56.4 (3.0)
160	0.064	**75.7** (1.5)	70.7 (2.7)	66.6 (4.6)	66.2 (3.6)	65.4 (1.4)	60.3 (2.7)	63.7 (2.3)
320	0.128	**78.5** (1.4)	74.9 (1.5)	71.0 (2.3)	72.6 (1.4)	72.8 (1.8)	68.0 (1.9)	71.0 (1.0)
640	0.256	**82.0** (1.2)	79.8 (1.3)	76.2 (1.8)	77.9 (0.9)	79.6 (1.2)	74.6 (1.6)	76.4 (0.9)
1280	0.512	**85.0** (1.0)	83.4 (1.5)	80.0 (2.0)	81.0 (1.2)	83.4 (0.9)	78.8 (1.7)	80.2 (0.9)

果は，λ の選択を識別的に行うことを除いて生成モデルアプローチの枠組で分類器を学習する EM-λ には，高精度な分類を達成できない本質的な限界があることを示唆している．

CH と比較して，H-BCM は，WebKB でラベルありデータ数が少数の場合を除いて高い性能を示した．H-BCM と CH では，ともに生成モデル，識別の両アプローチのハイブリッドに基づいて分類器を構築する．CH の H-BCM との相異点は，偏り補正モデルを用いずに半教師あり学習を行うことである．実験結果より，偏り補正モデルの導入は，ハイブリッド分類器の半教師あり学習に有効であり，汎化性能の向上に寄与することを確認した．

4.3 複数の構成要素を扱う効果

複数の構成要素を扱う効果を確認するため，文献 7) で提案した方法に，主要な構成要素 α のみを用いて分類器を設計した場合（便宜上，BCM-α と呼ぶ）と H-BCM の分類性能を比較した．BCM-α による分類器のクラス事後確率分布は，

$$R(y = k|\boldsymbol{x}^{\alpha}; \hat{\Theta}, \Psi, \Gamma) = \frac{1}{Z'} e^{\mu_k} p(\boldsymbol{x}^{\alpha}|k; \hat{\boldsymbol{\theta}}_k)^{\gamma_1} p(\boldsymbol{x}^{\alpha}|k; \boldsymbol{\psi}_k)^{\gamma_2} \quad (16)$$

で表される．ただし，$Z' = \sum_{k'=1}^{K} \{e^{\mu_{k'}} p(\boldsymbol{x}^{\alpha}|k'; \hat{\boldsymbol{\theta}}_{k'})^{\gamma_1} \times p(\boldsymbol{x}^{\alpha}|k'; \boldsymbol{\psi}_{k'})^{\gamma_2}\}$ である．さらに，CH，EM-λ，MLR/MER を主要な構成要素 α のみに適用した場合（便宜上，CH-α，EM-λ-α，MLR/MER-α と呼ぶ）と比較した．

各手法の性能比較には，4.2 節で述べた実験と同一の 10 種類の訓練・テストデータセットによる実験で得られた分類精度の平均値を用いた．また，4.2.1 項で述べた方法で，各手法のハイパーパラメータを設定した．BCM-α，CH-α，EM-λ-α，MLR/MER-α では，各テストコレクションで最も高い分類精度を与える構成要素を主要な構成要素 α として実験に用いた．主要な構成要素 α は，各テストコレクションでそれぞれ MT（WebKB），CI（Cora），M（20news）であった．

表 3 に，各テストコレクションで，ラベルありデータ数 N を変えて実験を行ったときの各手法の分類精度の平均値を示す．括弧内の数値は標準偏差を示す．

H-BCM の分類精度は，ラベルありデータ数が少数の場合を除いて BCM-α より高かった．また，ラベルありデータ数が大きいほど，H-BCM の分類精度は BCM-α を大きく上回る傾向があった．これは，ラベルありデータ数が少数の場合，分類器の汎化性能の向上に大きく寄与しない構成要素を加えることで，H-BCM は BCM-α より過学習を引き起こす危険性があることを示している．しかし，ラベルありデータ数の増加にともない，H-BCM は複数の構成要素を効果的に扱って高い分類精度を実現できることを確認した．

CH の分類精度は Cora の $N = 56$ の場合を除いて

表 3 複数の構成要素を扱う分類器と単一の構成要素を扱う分類器の異なるラベルありデータ数 N に対する分類精度 (%)

Table 3 Classification accuracies (%) with classifiers dealing with multi-component or single-component over various number of labeled data, N.

(a) WebKB（$M = 2500$, $K = 4$）

Training set		Multi-component				Single-component			
N	N/M	H-BCM	CH	EM-λ	MLR/MER	BCM-α	CH-α	EM-λ-α	MLR/MER-α
32	0.0128	69.9 (5.1)	**73.5** (4.1)	73.4 (2.8)	63.5 (5.0)	73.4 (3.0)	72.4 (4.0)	72.7 (2.8)	63.6 (4.7)
64	0.0256	76.2 (2.8)	**77.5** (2.3)	75.8 (1.7)	72.2 (2.7)	76.8 (2.1)	75.9 (2.1)	75.7 (2.4)	72.4 (2.7)
128	0.0512	**81.8** (1.7)	80.2 (2.0)	78.3 (2.4)	78.5 (2.2)	79.4 (1.7)	78.6 (2.1)	76.9 (2.4)	78.4 (2.3)
256	0.1024	**85.0** (1.3)	83.2 (1.5)	80.8 (1.6)	84.3 (1.7)	81.5 (1.7)	80.6 (1.1)	79.9 (1.6)	84.1 (1.7)
512	0.2048	87.2 (1.1)	84.8 (1.5)	82.4 (1.3)	**88.7** (1.3)	83.6 (1.6)	82.4 (1.6)	81.4 (1.4)	88.5 (1.3)

(b) Cora（$M = 2000$, $K = 7$）

Training set		Multi-component				Single-component			
N	N/M	H-BCM	CH	EM-λ	MLR/MER	BCM-α	CH-α	EM-λ-α	MLR/MER-α
56	0.028	77.4 (3.5)	74.3 (4.9)	71.8 (4.1)	55.6 (4.4)	**78.0** (3.2)	74.4 (5.2)	73.6 (4.2)	73.6 (4.0)
112	0.056	**83.4** (1.7)	80.7 (1.4)	77.9 (1.6)	63.9 (2.3)	81.8 (1.5)	79.7 (2.4)	79.5 (1.8)	78.6 (1.9)
224	0.112	**85.8** (0.9)	84.1 (0.7)	81.2 (1.1)	73.0 (1.0)	85.0 (0.8)	83.1 (0.8)	82.6 (1.4)	82.5 (1.1)
448	0.224	**87.3** (1.6)	86.1 (1.3)	83.3 (1.1)	78.8 (1.0)	86.0 (0.9)	85.5 (0.8)	84.2 (1.0)	84.2 (1.0)
896	0.448	**89.1** (0.6)	88.5 (1.2)	85.9 (1.0)	82.7 (1.0)	87.5 (1.0)	87.3 (0.8)	86.1 (0.9)	85.9 (1.2)

(c) 20news（$M = 2500$, $K = 5$）

Training set		Multi-component				Single-component			
N	N/M	H-BCM	CH	EM-λ	MLR/MER	BCM-α	CH-α	EM-λ-α	MLR/MER-α
40	0.016	64.8 (3.0)	56.2 (4.6)	56.6 (6.9)	50.1 (5.1)	**66.6** (3.3)	52.9 (7.9)	52.9 (7.5)	49.3 (4.9)
80	0.032	**71.6** (1.8)	63.7 (2.2)	60.8 (3.8)	57.3 (3.7)	69.7 (1.8)	60.4 (2.7)	56.4 (3.8)	56.1 (3.7)
160	0.064	**75.7** (1.5)	70.7 (2.7)	66.6 (4.6)	66.2 (3.6)	72.7 (1.4)	66.8 (3.1)	63.7 (4.7)	64.3 (3.1)
320	0.128	**78.5** (1.4)	74.9 (1.5)	71.0 (2.3)	72.6 (1.4)	74.9 (1.1)	71.0 (1.2)	67.3 (2.1)	70.9 (1.3)
640	0.256	**82.0** (1.2)	79.8 (1.3)	76.2 (1.8)	77.9 (0.9)	77.9 (1.0)	74.8 (1.2)	71.7 (1.6)	75.7 (1.1)
1280	0.512	**85.0** (1.0)	83.4 (1.5)	80.0 (2.0)	81.0 (1.2)	80.5 (0.9)	79.2 (1.6)	76.0 (2.4)	79.0 (1.2)

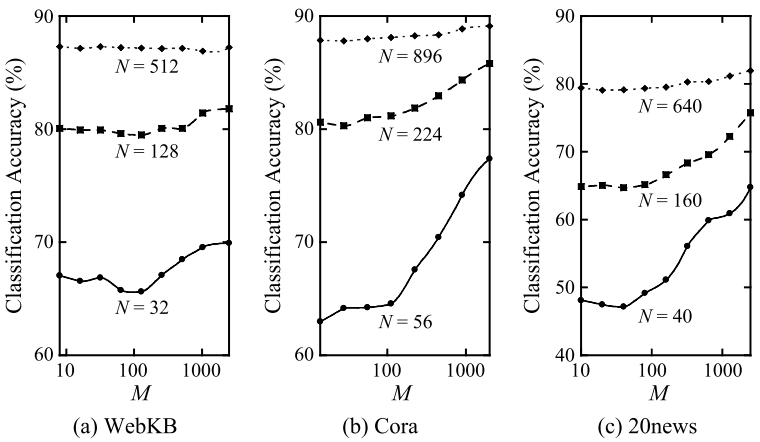

図 2 異なるラベルなしデータ数 M に対する H-BCM の分類精度 (%)

Fig. 2 Classification accuracies (%) with H-BCM over various number of unlabeled data, M.

CH-α より高く，複数の構成要素を扱うことで分類精度が向上する傾向があった．この傾向は，H-BCM と BCM-α の比較結果とほぼ同様であった．一方，EM-λ と MLR/MER の分類精度は，Cora では主要な構成要素 α のみを用いる EM-λ-α と MLR/MER-α よりも低く，WebKB ではほとんどの場合で分類精度が 1%弱高い程度であった．H-BCM と CH はともに複数の構成要素を識別学習で与えられる重みを用いて統合する．これらの実験結果は，H-BCM と CH で用いた構成要素の統合方法が，複数の構成要素を扱うのに有効であったことを示している．

4.4 ラベルなしデータの効果

分類器の汎化精度向上に対するラベルなしデータの効果を確認するため，ラベルなしデータ数を変えて学習したときの H-BCM の分類精度を調べた．**図 2** に，ラベルありデータ数を N に固定したときのラベルなしデータ数 M に対するテストデータの分類精度の平均値の変化を示す．実験では，テストコレクションごとに 1,000 個のテストデータを用いた．分類精度の平均値は，ラベルあり・なしデータとテストデータをラ

ンダムに選択して行った 10 回の実験の結果から算出した．H-BCM の学習には，4.2 節の実験で他手法との性能比較の際に用いたハイパーパラメータ値を使用した．

図 2 に示されるように，ラベルありデータ数を固定するとき，WebKB の $N = 512$ の場合を除いて，学習に用いるラベルなしデータ数を増やすことで分類精度が上昇する傾向が見られた．とくに，ラベルありデータが少数の場合，多数のラベルなしデータを学習に用いることで分類精度が大きく上昇した．H-BCM において，ラベルありデータの不足により生じる汎化性能の低さを改善するのに多数のラベルなしデータを学習に用いることが有効であることを確認した．

5. ま と め

本研究では，複数の構成要素を含むデータの多クラス単一ラベル分類問題に対して，生成モデル・識別の両アプローチのハイブリッドに基づく分類器の半教師あり学習法を提案した．提案法は，各構成要素に対して，個々に設計される生成モデルと，ラベルありデータが少数であることに起因する学習の偏りを緩和するために導入されるモデルを，最大エントロピー原理に基づいて統合することを特徴とする．3 つのテストコレクションを用いた分類実験で，生成モデルであるナイーブベイズモデルと識別モデルである多項ロジスティック回帰モデルの半教師あり学習で類似の分類精度が得られる場合にとくに，提案法では両モデルよりも高い分類精度を得られることを確認した．また，単一の構成要素のみを用いた分類器との性能比較により，提案法に基づいて複数の構成要素を扱うことの効果を確認した．さらに，提案法では，高い分類精度を得るのに十分な量のラベルありデータを与えていない場合に，多数のラベルなしデータを学習に用いることで分類性能が大幅に向上することを確認した．今後の課題は，テキストと画像といった異なる分布型でモデル化される異種情報データに対して提案法の有用性を確認することである．

参 考 文 献

1) Berger, A.L., Della Pietra, S.A. and Della Pietra, V.J.: A maximum entropy approach to natural language processing, *Computational Linguistics*, Vol.22, No.1, pp.39–71 (1996).
2) Brochu, E. and Freitas, N.: “Name that song!”: A probabilistic approach to querying on music and text, *Advances in Neural Information Processing Systems 15*, pp.1505–1512, MIT Press, Cambridge, MA (2003).
3) Chakrabarti, S., Dom, B. and Indyk, P.: Enhanced hypertext categorization using hyperlinks, *Proc. ACM International Conference on Management of Data* (*SIGMOD-98*), pp.307–318 (1998).
4) Chen, S.F. and Rosenfeld, R.: A Gaussian prior for smoothing maximum entropy models, Technical report, Carnegie Mellon University (1999).
5) 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).
6) 藤野昭典，上田修功，斉藤和巳：最大エントロピー原理に基づく付加情報の効果的な利用によるテキスト分類，情報処理学会論文誌，Vol.47, No.10, pp.2929–2937 (2006).
7) 藤野昭典，上田修功，斉藤和巳：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，人工知能学会論文誌，Vol.21, No.3, pp.301–309 (2006).
8) Good, I.J.: *The estimation of probabilities: An essay on modern Bayesian methods*, MIT Press, Cambridge, MA (1965).
9) Grandvalet, Y. and Bengio, Y.: Semi-supervised learning by entropy minimization, *Advances in Neural Information Processing Systems 17*, pp.529–536, MIT Press, Cambridge, MA (2005).
10) Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York, Berlin, Heidelberg (2001).
11) 北 研二：確率的言語モデル，東京大学出版会 (1999).
12) Liu, D.C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Math. Programming, Ser. B*, Vol.45, No.3, pp.503–528 (1989).
13) Lu, Q. and Getoor, L.: Link-based text classification, *IJCAI Workshop on Text-Mining & Link-Analysis* (*TextLink 2003*) (2003).
14) Ng, A.Y. and Jordan, M.I.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, *Advances in Neural Information Processing Systems 14*, pp.841–848, MIT Press, Cambridge, MA (2002).
15) Nigam, K., Lafferty, J. and McCallum, A.: Using maximum entropy for text classification, *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp.61–67 (1999).
16) 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).
17) Raina, R., Shen, Y., Ng, A.Y. and McCallum, A.: Classification with hybrid generative/discriminative models, *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA (2004).
18) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, New York (1983).
19) Seeger, M.: Learning with labeled and unlabeled data, Technical report, University of Edinburgh (2001).
20) Sun, A., Lim, E.P. and Ng, W.K.: Web classification using support vector machine, *Proc. 4th Int. Workshop on Web Information and Data Management* (*WIDM 2002*) *held in conj. with CIKM 2002*, pp.96–99 (2002).
21) 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).

付　　録

A.1 F_2 が上に凸の関数であることの証明

式 (10) で示した目的関数 $F_2(\Gamma|\Psi)$ が Γ に関して上に凸の関数であることを証明するには，$F_2(\Gamma|\Psi)$ の Γ に関するヘッセ行列が負定値であることを示せばよい．略記のため，式 (9)–(11) を

$$\begin{aligned}\boldsymbol{f}_{nk}^t &= (\{f_{nks}\}_{s=1}^S)\\ &= (\{\log p(\boldsymbol{x}_n|k;\hat{\boldsymbol{\theta}}_k^{j,(-n)}),\\ &\quad \log p(\boldsymbol{x}_n|k;\boldsymbol{\psi}_k^j)\}_{j=1}^J,\\ &\quad \{I_k(k')\}_{k'=1}^K) \qquad (17)\\ \boldsymbol{\lambda}^t &= (\{\lambda_s\}_{s=1}^S)\\ &= (\{\gamma_{1j},\gamma_{2j}\}_{j=1}^J,\{\mu_k\}_{k=1}^K) \qquad (18)\\ \boldsymbol{\Phi} &= diag(\{\phi_s\}_{s=1}^S)\\ &= diag(\{\sigma_1^{-2},\sigma_2^{-2}\}^J,\{\rho^{-2}\}^K) \qquad (19)\\ \boldsymbol{b}^t &= (\{b_s\}_{s=1}^S)\\ &= (\{a_1,a_2\}^J,\{0\}^K) \qquad (20)\end{aligned}$$

で示す行列表記を用いて書き換える．ただし，$S=2J+K$ であり，$\boldsymbol{f}_{nk}^t$ は $\boldsymbol{f}_{nk}$ の転置行列を，$diag(\{\phi_s\}_{s=1}^S)$ は s 番目の対角成分が ϕ_s である対角行列を表す．$I_k(k')$ は $k'=k$ のときに 1，それ以外に 0 である指示関数を表す．さらに，$\boldsymbol{\lambda}^t=\Gamma$ であるので，以下の説明では，F_2 が Γ に関して上に凸の関数であることを示すのに，Γ の代わりに $\boldsymbol{\lambda}$ を用いる．また，

$$R_{nk}(\boldsymbol{\lambda}) = R(y=k|\boldsymbol{x}_n;\hat{\Theta}^{(-n)},\Psi,\Gamma) \qquad (21)$$

を用いて略記する．

式 (17)–(21) を用いると，式 (10) の目的関数は

$$F_2(\boldsymbol{\lambda}) = \sum_{n=1}^N \log R_{ny_n}(\boldsymbol{\lambda}) - \log p(\boldsymbol{\lambda}) \qquad (22)$$

のように書き換えられる．ただし，

$$R_{ny_n}(\boldsymbol{\lambda}) = \frac{\exp\left(\boldsymbol{\lambda}^t\boldsymbol{f}_{ny_n}\right)}{\sum_{k'=1}^K \exp\left(\boldsymbol{\lambda}^t\boldsymbol{f}_{nk'}\right)} \qquad (23)$$

$$p(\boldsymbol{\lambda}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{b})^t\boldsymbol{\Phi}(\boldsymbol{\lambda}-\boldsymbol{b})\right\} \qquad (24)$$

である．

F_2 の $\boldsymbol{\lambda}$ に関するヘッセ行列 $\partial F_2/\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^t$ を得るために，F_2 を $\boldsymbol{\lambda}$ について偏微分すると，

$$\begin{aligned}\frac{\partial F_2}{\partial\boldsymbol{\lambda}} &= \sum_{n=1}^N\left(\boldsymbol{f}_{ny_n} - \sum_{k=1}^K R_{nk}(\boldsymbol{\lambda})\boldsymbol{f}_{nk}\right)\\ &\quad -\boldsymbol{\Phi}(\boldsymbol{\lambda}-\boldsymbol{b}) \qquad (25)\end{aligned}$$

となる．さらに，F_2 の 2 階偏微分を求めると，

$$\begin{aligned}\frac{\partial^2 F_2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^t} &= -\sum_{n=1}^N\sum_{k=1}^K\left\{R_{nk}(\boldsymbol{\lambda})\boldsymbol{f}_{nk}\right.\\ &\quad \left.\times\left(\boldsymbol{f}_{nk}-\sum_{k'=1}^K R_{nk'}(\boldsymbol{\lambda})\boldsymbol{f}_{nk'}\right)^t\right\}\\ &\quad -\boldsymbol{\Phi} \qquad (26)\end{aligned}$$

となる．式 (21) より，$\sum_{k=1}^K R_{nk}(\boldsymbol{\lambda})=1$，$R_{nk}(\boldsymbol{\lambda})\geq 0$ であるので，$\boldsymbol{h}_n(\boldsymbol{\lambda})=\sum_{k=1}^K R_{nk}(\boldsymbol{\lambda})\boldsymbol{f}_{nk}$ を用いて，式 (26) を

$$\begin{aligned}\frac{\partial^2 F_2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^t} &= -\sum_{n=1}^N\sum_{k=1}^K\Big[R_{nk}(\boldsymbol{\lambda})\left\{\boldsymbol{f}_{nk}-\boldsymbol{h}_n(\boldsymbol{\lambda})\right\}\\ &\quad \times\left\{\boldsymbol{f}_{nk}-\boldsymbol{h}_n(\boldsymbol{\lambda})\right\}^t\Big]-\boldsymbol{\Phi} \qquad (27)\end{aligned}$$

のように変換できる．このため，任意の S 次元ベクトル $\boldsymbol{u}=(u_1,\ldots,u_s,\ldots,u_S)^t$ に対して，

$$\begin{aligned}&\boldsymbol{u}^t\frac{\partial^2 F_2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^t}\boldsymbol{u}\\ &= -\sum_{n=1}^N\sum_{k=1}^K R_{nk}(\boldsymbol{\lambda})\left[\boldsymbol{u}^t\left\{\boldsymbol{f}_{nk}-\boldsymbol{h}_n(\boldsymbol{\lambda})\right\}\right]^2\\ &\quad -\boldsymbol{u}^t\boldsymbol{\Phi}\boldsymbol{u} \qquad (28)\end{aligned}$$

となる．$\boldsymbol{u}\neq\boldsymbol{0}$ のとき，式 (19) の定義より $\boldsymbol{u}^t\boldsymbol{\Phi}\boldsymbol{u}>0$ であるので，$\boldsymbol{u}^t\left(\partial^2 F_2/\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^t\right)\boldsymbol{u}<0$ がすべての $\boldsymbol{\lambda}$ に対して成り立つ．したがって，$\boldsymbol{\lambda}$ に関するヘッセ行列は負定値であり，目的関数 $F_2(\boldsymbol{\lambda})$ は $\boldsymbol{\lambda}$ に関して上に凸の関数であることが示された．

(平成 19 年 2 月 2 日受付)
(平成 19 年 3 月 24 日再受付)
(平成 19 年 4 月 27 日採録)

藤野 昭典（正会員）

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研究主任．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，IEEE 各会員．

上田 修功（正会員）

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．同年日本電信電話公社（現 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協創情報研究部長，奈良先端科学技術大学院大学客員教授．電気通信普及財団賞（1997 年，2006 年），電子情報通信学会論文賞（2002 年，2004 年）等受賞，電子情報通信学会，日本神経回路学会，IEEE 各会員．

斉藤 和巳（正会員）

1963 年生．1985 年慶應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．2007 年より静岡県立大学経営情報学部教授．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，人工知能学会，日本神経回路学会，IEEE 各会員．