情報処理学会論文誌　数理モデル化と応用　Vol. 2　No. 2　132–144 (Mar. 2009)

# 生成・識別モデルの統合に基づく半教師あり学習法とその多重分類への応用

藤野昭典[†1]　上田修功[†1]　磯崎秀樹[†1]

各データが複数のカテゴリに属する多重分類問題に対して，ラベルありデータとラベルなしデータを用いた半教師あり学習により分類器を設計する手法を提案する．提案法では，ラベルありデータで学習させた識別モデルとラベルなしデータで学習させた生成モデルを統合することによって分類器を構築する．多重テキスト分類のために，識別モデルと生成モデルにそれぞれ対数線形モデルとナイーブベイズモデルを用いて分類器を設計する．実テキストデータからなる 3 つのテストコレクションを用いた実験で，従来の対数線形モデルとナイーブベイズモデルの半教師あり学習法と比較して，提案法ではより高い汎化能力を持つ多重分類器を得られることを確認した．

# A Semi-supervised Learning Method Based on Generative/Discriminative Model Combination and Its Application to Multi-label Classification

Akinori Fujino,[†1] Naonori Ueda[†1]
and Hideki Isozaki[†1]

We propose a method for designing semi-supervised multi-label classifiers, which select one or more category labels for each data sample and are trained by using labeled and unlabeled samples. The proposed method provides a classifier based on a combination of discriminative and generative models trained on labeled and unlabeled samples, respectively. We design a multi-label text classifier by utilizing log-linear and naive Bayes models as the discriminative and generative models, respectively. Using three test collections consisting of real text data, we confirmed experimentally that the proposed method provided better multi-label classifiers with high generalization ability than conventional semi-supervised learning methods of log-linear and naive Bayes models.

## 1. はじめに

大量のデータをインターネット等で容易に取得できる現在，効率的にデータを管理するためにデータの内容を表すカテゴリラベルを付与する技術のニーズが高まっている．各データに 1 つ以上のカテゴリラベルを付与する問題は多重分類問題と呼ばれ，テキストデータ[15]や遺伝子データ[6]，画像データ[1] 等を対象として機械学習に基づく多重分類法の研究が行われてきた．

機械学習に基づく手法では，一般的に，属するクラス[⋆1]が既知のデータ（ラベルありデータ）を用いて分類器を学習させる．このとき，多数のラベルありデータを用いるほど汎化性能が高い分類器を得やすいことが知られている．しかし，ラベルありデータの作成には対象データに精通した専門家によるラベル付けが必要であり，大量に作成するには高いコストを要する．一方，属するクラスが未知のデータ（ラベルなしデータ）を集めるのは比較的容易である．このため，ラベルなしデータをラベルありデータと同時に学習に用いて分類器の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 つとなっている．

半教師あり学習に基づく分類器は，生成モデルと識別モデル，両モデルのハイブリッドの各アプローチに基づいて提案されてきた．生成モデルアプローチでは，データ $\boldsymbol{x}$ とクラス $\boldsymbol{y}$ の同時確率分布 $p(\boldsymbol{x}, \boldsymbol{y})$ のモデル（生成モデル）を学習させ，ベイズ則により得られる条件付き確率 $P(\boldsymbol{y}|\boldsymbol{x})$ を用いてデータを分類する．生成モデルはデータの種類に応じて設計される．また，ラベルなしデータをクラスに関する不完全データと見なし，生成モデルの混合によりデータの確率密度 $p(\boldsymbol{x})$ をモデル化する[17] ことで分類器の学習に用いる．

一方，識別モデルアプローチでは，データのクラス境界を直接的に学習する．たとえば，ロジスティック（LR）回帰モデルや最大エントロピーモデル[16] 等の対数線形モデルでは，条件付き確率分布 $P(\boldsymbol{y}|\boldsymbol{x})$ をモデル化し，ラベルありデータを高精度に識別できるように条件付き確率モデルを学習する．しかし，識別モデルアプローチではデータの確率密度 $p(\boldsymbol{x})$ をモデル化しない．それゆえ，半教師あり学習ではラベルなしデータを学習に用いるための仮定が必要になる．最小エントロピー正則化[9] を用いる LR モデルや transductive SVM (TSVM)[3),11)]，グラフに基づく方法[23] 等では，データの分布密度が低い領域にクラス境界が存在すると仮定し，ラベルなしデータをよく分離させるようにモデルを学習させる．

†1 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

⋆1 本論文では，カテゴリラベルの付与の有無で分けられたデータのサブ集合をクラスと呼ぶ．

これらの手法に対して，我々は，最近，生成モデルと識別学習のハイブリッドに基づく分類器設計法を提案した[8]．このハイブリッド法では，まず，クラスごとにラベルありデータで学習させる生成モデルを設計する．ラベルありデータが少数の場合，学習させた生成モデルがラベルありデータに過適合することによって新規のデータを高精度に分類できない危険性がある．そこで，過適合の影響を減少させるために，新たな生成モデルを導入し，これらの生成モデルの重み付き統合によって分類器を構築する．統合の重みの学習には識別学習を用い，新たに導入した生成モデルの学習にはラベルなしデータを用いる．実データを用いた実験により，生成モデルと識別モデルの両アプローチで同程度の分類精度が得られる場合に，ハイブリッド法で得られる分類精度が両アプローチよりも高いことを確認した．

このハイブリッド法は，データに 1 つのカテゴリラベルが付与される単一ラベル分類問題を対象とした手法であるため，多重分類問題に適用するには，カテゴリラベルごとに付与の可否を判定する 2 値分類に用いる必要がある．この適用方法では，ラベルありデータで学習させる生成モデルを各カテゴリラベルで個別に設計して用いる．しかし，このモデル設計では，カテゴリラベルの組合せ（多重ラベル）を考慮しないため，属するカテゴリが 1 つに限定されない多重ラベルのデータによく適合する生成モデルが与えられるとは限らない．一般的に，ラベルありデータでモデルを学習させる場合，データにあまり適合しない生成モデルでは，識別モデルよりも高い分類精度を与えられないことが知られている[13]．それゆえ，この適用方法では，生成モデルの適合性の問題が識別モデルよりも高精度な多重分類を実現できない要因になる．また，2 値分類による手法では，カテゴリラベルごとに独立に分類器を学習させるため，カテゴリラベルの組合せに応じてデータを分類するのに適したクラス境界が学習されるとは限らない．

そこで本研究では，識別モデルと生成モデルの統合に基づく半教師あり学習により多重分類器を設計する手法を提案する．提案法では，生成モデルを学習させるのにラベルなしデータを用いる．ラベルなしデータのクラスは未知であるため，生成モデルの学習にはラベルありデータで学習させた識別モデルを利用する．そして，学習させた生成モデルと識別モデルを統合して多重ラベルの条件付き確率モデルを与える．

本論文では，まず，2 章で生成モデルと識別モデルの概要を簡潔に述べ，3 章で生成モデルと識別モデルの統合に基づく半教師あり学習法の基本的な枠組みを述べる．そして，4 章で本手法に基づいて多重テキスト分類器を設計する方法を詳述する．さらに，実テキストデータからなる 3 つのテストコレクションを用いた評価実験の結果を 5 章で示し，従来法との比較により提案法の有効性を考察する．

## 2. 生成モデルと識別モデル

本研究では，各データに付与すべきカテゴリラベルを $K$ 個の候補の中から選択する問題に対し，データの特徴ベクトル $\boldsymbol{x}$ から，クラスベクトル $\boldsymbol{y} = (y_1, \ldots, y_k, \ldots, y_K)^T$，$y_k \in \{1, 0\}$ を推定する分類器を設計する．ここで，$y_{nk} = 1$（$y_{nk} = 0$）は $n$ 番目のデータ $\boldsymbol{x}_n$ に $k$ 番目のカテゴリラベルが付与される（付与されない）ことを表す．また，$\boldsymbol{a}^T$ は $\boldsymbol{a}$ の転置を表す．多重分類問題では各データの $\boldsymbol{y}$ は $K > 1$ かつ $\sum_{k=1}^{K} y_k \geq 1$ を満たし，単一ラベル分類問題では $\sum_{k=1}^{K} y_k = 1$ を満たす．$K = 1$ のときは 2 値分類問題となる．

半教師あり学習では，ラベルありデータ $D_l = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$ と多数のラベルなしデータ $D_u = \{\boldsymbol{x}_m\}_{m=1}^{M}$（$M >> N$）からなる訓練データ集合 $D = \{D_l, D_u\}$ を学習に用いることで，ラベルありデータのみを学習に用いる場合よりも汎化性能の高い分類器を得ることを目的とする．本論文では，半教師あり学習に基づいて多重分類器を設計することを課題とする．

本章では，分類器設計に用いる生成モデルと識別モデルの概要を簡潔に述べる．

### 2.1 生成モデル

生成モデルは，データの特徴ベクトル $\boldsymbol{x}$ とクラスベクトル $\boldsymbol{y}$ の同時確率分布 $p(\boldsymbol{x}, \boldsymbol{y})$ をモデル化したものであり，データの種類に応じて設計される．たとえば，多項分布モデルやガウス分布モデル，隠れマルコフモデルがそれぞれテキストデータや画像データ，系列データの生成モデルを設計するのによく利用される．生成モデルに基づく分類器は，生成モデルにベイズ則を適用して得られるクラス事後確率 $p(\boldsymbol{y}|\boldsymbol{x})$ を用いてデータの属するクラスを推定する．

ラベルありデータ集合 $D_l$ を用いた生成モデル $p(\boldsymbol{x}, \boldsymbol{y}; \Theta)$ の学習では，一般的に，パラメータ $\Theta$ の事後確率密度 $p(\Theta|D_l)$ を最大化させる値を $\Theta$ の推定値として求める（MAP 推定，文献 22) 参照）．すなわち，ベイズ則により，$\log p(\Theta|D_l)$ に相当する目的関数

$$J_g(\Theta) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n, \boldsymbol{y}_n; \Theta) + \log p(\Theta) \tag{1}$$

を最大化させる $\Theta$ を求める．第 1 項は生成モデルの対数尤度であり，第 2 項の $p(\Theta)$ は $\Theta$ の事前確率分布を表す．

### 2.2 識別モデル

識別モデルでは，ラベルありデータのクラスを高い精度で推定できるように識別関数を直接的に学習させる．識別関数 $f(\boldsymbol{x}, \boldsymbol{y}; W)$ は，データ $\boldsymbol{x}$ が属するクラスの推定値を

$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}; W)$ のように与える関数である．

対数線形（log-linear）モデルでは，パラメータ行列 $W$ と線形の識別関数 $f(\boldsymbol{x}, \boldsymbol{y}; W) = \boldsymbol{y}^T W \boldsymbol{x}$ を用いて，データ $\boldsymbol{x}$ に対するクラス $\boldsymbol{y}$ の条件付き確率分布を

$$P(\boldsymbol{y}|\boldsymbol{x}; W) = \frac{\exp\{f(\boldsymbol{x}, \boldsymbol{y}; W)\}}{\sum_{\boldsymbol{y}'} \exp\{f(\boldsymbol{x}, \boldsymbol{y}'; W)\}} \tag{2}$$

で与え，$P(\boldsymbol{y}|\boldsymbol{x}; W)$ を最大化させる $\boldsymbol{y}$ をデータ $\boldsymbol{x}$ のクラスの推定値とする．ラベルありデータ集合 $D_l$ を用いると，条件付き確率分布の対数尤度に基づく目的関数

$$J_d(W) = \sum_{n=1}^{N} \log P(\boldsymbol{y}_n|\boldsymbol{x}_n; W) + \log p(W) \tag{3}$$

の最大化により $W$ の推定値 $\hat{W}$ を求めることができる．$W$ の事前確率分布 $p(W)$ にはガウス事前確率分布[2] がよく用いられる．

## 3. 提案法の基本的な枠組み

本論文では，多重分類のために，生成モデルと識別モデルの統合に基づく半教師あり学習により分類器を設計する手法を提案する．提案法では，ラベルなしデータで学習させた生成モデルとラベルありデータで学習させた識別モデルを統合して分類器の条件付き確率モデルを与える．以下に，提案法の基本的な枠組みを述べる．

### 3.1 生成モデルの学習法と条件付き確率モデル

提案法では，ラベルなしデータで学習させた生成モデル $p(\boldsymbol{x}, \boldsymbol{y}; \Theta)$ を用いて分類器を構築する．仮に，ラベルなしデータ $\boldsymbol{x}_m$ の属するクラス $\boldsymbol{y}_m$ が既知であれば，MAP 推定により，$J_g(\Theta) = \sum_{m=1}^{M} \sum_{\boldsymbol{y}} I_{\boldsymbol{y}_m}(\boldsymbol{y}) \log p(\boldsymbol{x}_m, \boldsymbol{y}; \Theta) + \log p(\Theta)$ を最大化させる $\Theta$ を生成モデルのパラメータ推定値として求めることができる．ただし，$I_{\boldsymbol{y}_m}(\boldsymbol{y})$ は，$\boldsymbol{y} = \boldsymbol{y}_m$ の場合に 1，$\boldsymbol{y} \neq \boldsymbol{y}_m$ の場合に 0 の値をとる指示関数である．しかし，$\boldsymbol{y}_m$ は未知であるため，$I_{\boldsymbol{y}_m}(\boldsymbol{y})$ の値は定まらない．そこで，$I_{\boldsymbol{y}_m}(\boldsymbol{y})$ の期待値を与える条件付き確率分布 $R(\boldsymbol{y}|\boldsymbol{x})$ を導入し，$\Theta$ を推定するための目的関数を

$$J(R, \Theta) = \sum_{m=1}^{M} \sum_{\boldsymbol{y}} R(\boldsymbol{y}|\boldsymbol{x}_m) \log p(\boldsymbol{x}_m, \boldsymbol{y}; \Theta) + \log p(\Theta) \tag{4}$$

で与える．ただし，$R(\boldsymbol{y}|\boldsymbol{x})$ もまた推定すべき未知の確率分布である．

一方，$R(\boldsymbol{y}|\boldsymbol{x})$ は，ラベルありデータに対して，高精度なクラス推定を与える条件付き確率分布である必要がある．提案法では，$R(\boldsymbol{y}|\boldsymbol{x})$ を与えるのに，ラベルありデータで学習させた識別関数 $f(\boldsymbol{x}, \boldsymbol{y}; \hat{W})$ に基づく条件付き確率分布

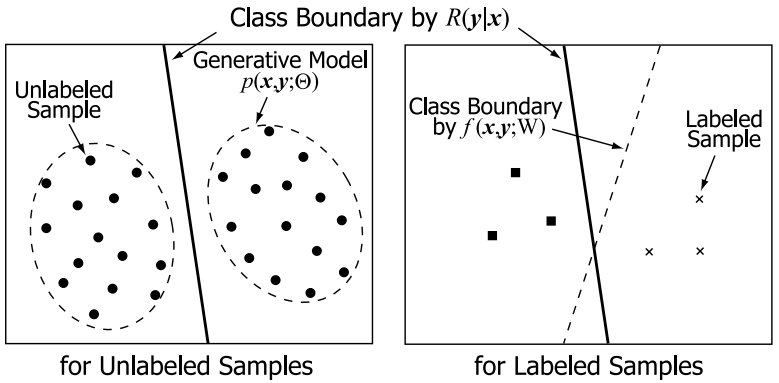


**図 1**　分類器設計の基本概念
Fig. 1　Basic idea of classifier design.

$$P_\gamma(\boldsymbol{y}|\boldsymbol{x}; \hat{W}) = \frac{\exp\{\gamma f(\boldsymbol{x}, \boldsymbol{y}; \hat{W})\}}{\sum_{\boldsymbol{y}'} \exp\{\gamma f(\boldsymbol{x}, \boldsymbol{y}'; \hat{W})\}} \tag{5}$$

を利用する．識別モデルでは，2.2 節で述べたように，ラベルありデータを高い精度で識別できるように識別関数を学習させる．このため，$P_\gamma(\boldsymbol{y}|\boldsymbol{x}; \hat{W})$ は，重みパラメータ $\gamma$ が正の実数値の場合，ラベルありデータを高い精度で識別する条件付き確率分布を与える．それゆえ，$P_\gamma(\boldsymbol{y}|\boldsymbol{x}; \hat{W})$ に近い確率分布で $R(\boldsymbol{y}|\boldsymbol{x})$ を与える．

提案法では，$R(\boldsymbol{y}|\boldsymbol{x})$ と $P_\gamma(\boldsymbol{y}|\boldsymbol{x}; \hat{W})$ の両確率分布の相違を表す尺度として，KL 距離（Kullback-Leibler divergence）

$$D(R(\boldsymbol{y}|\boldsymbol{x})||P_\gamma(\boldsymbol{y}|\boldsymbol{x}; \hat{W})) = \sum_{\boldsymbol{y}} R(\boldsymbol{y}|\boldsymbol{x}) \log \frac{R(\boldsymbol{y}|\boldsymbol{x})}{P_\gamma(\boldsymbol{y}|\boldsymbol{x}; \hat{W})} \tag{6}$$

を用いる．KL 距離の最小化に基づく制約の下で，式 (4) で与えた $J(R, \Theta)$ を最大化させる $R(\boldsymbol{y}|\boldsymbol{x})$ を推定することにより，**図 1** の概念図に示すような，ラベルなしデータに生成モデルを適合させ，かつラベルありデータに適したクラス境界を与える条件付き確率分布を求める．すなわち，$R(\boldsymbol{y}|\boldsymbol{x})$ と $\Theta$ を推定するための目的関数を

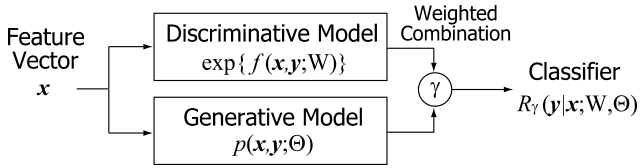


**図 2**　分類器の構成
Fig. 2　Outline of classifier formulation.

$$J_{\gamma,\beta}(R,\Theta)=\sum_{m=1}^{M}\sum_{\boldsymbol{y}}R(\boldsymbol{y}|\boldsymbol{x}_m)\log p(\boldsymbol{x}_m,\boldsymbol{y};\Theta)-\frac{1}{\beta}\sum_{m=1}^{M}D(R(\boldsymbol{y}|\boldsymbol{x}_m)||P_\gamma(\boldsymbol{y}|\boldsymbol{x}_m;\hat{W}))+\log p(\Theta) \tag{7}$$

で与える．$\beta$ は生成モデル学習の目的関数と KL 距離に基づく制約のバランスを定める正定数である．$\sum_{\boldsymbol{y}}R(\boldsymbol{y}|\boldsymbol{x})=1$ の制約の下でラグランジュ未定乗数法を適用すると，$\partial J_{\gamma,\beta}/\partial R=0$ を満たす条件付き確率分布

$$\begin{aligned}R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x};\Theta,\hat{W})&=\frac{P_\gamma(\boldsymbol{y}|\boldsymbol{x};\hat{W})p(\boldsymbol{x},\boldsymbol{y};\Theta)^\beta}{Z_{\boldsymbol{\gamma}}(\boldsymbol{x};\Theta,\hat{W})}\\&=\frac{\exp\{f(\boldsymbol{x},\boldsymbol{y};\hat{W})\}^\gamma p(\boldsymbol{x},\boldsymbol{y};\Theta)^\beta}{\sum_{\boldsymbol{y}'}\exp\{f(\boldsymbol{x},\boldsymbol{y}';\hat{W})\}^\gamma p(\boldsymbol{x},\boldsymbol{y}';\Theta)^\beta}\end{aligned} \tag{8}$$

が得られる．ただし，$Z_{\boldsymbol{\gamma}}(\boldsymbol{x};\Theta,\hat{W})=\sum_{\boldsymbol{y}'}P_\gamma(\boldsymbol{y}'|\boldsymbol{x};\hat{W})p(\boldsymbol{x},\boldsymbol{y}';\Theta)^\beta$ である．$\boldsymbol{\gamma}=(\gamma,\beta)^T$ は識別関数と生成モデルの統合の重みを与える．提案法では，上記の $R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x};\Theta,\hat{W})$ を分類器の条件付き確率モデルとして用いる．**図 2** に分類器の構成を図示する．

式 (8) で示した条件付き確率分布は，単一ラベル分類問題の場合，式中の $\exp\{f(\boldsymbol{x},\boldsymbol{y};\hat{W})\}$ をラベルありデータで学習させた生成モデルに置き換えることで我々が以前提案したハイブリッド法[8] の条件付き確率分布と一致する．生成モデルの代わりに識別関数を用いることは単純な改変であるが，1 章で述べた理由により，この改変は多重分類の精度向上に大きく寄与する可能性がある．事実，5 章の実験結果で示すように，識別モデルの識別関数を用いる提案法では，文献 8) の手法を応用した場合よりも汎化性能の高い多重分類器を得られた．

生成モデルのパラメータ $\Theta$ の推定値は，式 (7) に式 (8) を代入して得られる以下の目的関数の最大化に基づいて求められる．

$$J_{\boldsymbol{\gamma}}(\Theta)=\frac{1}{\beta}\sum_{m=1}^{M}\log Z_{\boldsymbol{\gamma}}(\boldsymbol{x}_m;\Theta,\hat{W})+\log p(\Theta) \tag{9}$$

$\boldsymbol{\gamma}$ の値を設定すると，EM アルゴリズム[4] のような繰返し計算を行うことで，$J_{\boldsymbol{\gamma}}(\Theta)$ を初期値近傍で最大化させる $\Theta$ を推定できる．繰返し計算の第 $(t)$ ステップでの $\Theta$ の推定値を $\Theta^{(t)}$ とすると，$J_{\boldsymbol{\gamma}}(\Theta^{(t+1)})-J_{\boldsymbol{\gamma}}(\Theta^{(t)})\geq Q_{\boldsymbol{\gamma}}(\Theta^{(t+1)},\Theta^{(t)})-Q_{\boldsymbol{\gamma}}(\Theta^{(t)},\Theta^{(t)})$ を満たす $Q$ 関数を

$$\begin{aligned}&Q_{\boldsymbol{\gamma}}(\Theta^{(t+1)},\Theta^{(t)})\\&\quad=\sum_{m=1}^{M}\sum_{\boldsymbol{y}}R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x}_m;\Theta^{(t)},\hat{W})\log p(\boldsymbol{x}_m,\boldsymbol{y};\Theta^{(t+1)})+\log p(\Theta^{(t+1)})\end{aligned} \tag{10}$$

で表せる（導出方法は文献 8) の付録を参照）．$Q_{\boldsymbol{\gamma}}(\Theta^{(t)},\Theta^{(t)})$ は $\Theta^{(t+1)}$ と独立な関数であるため，$Q_{\boldsymbol{\gamma}}(\Theta^{(t+1)},\Theta^{(t)})$ を最大化させる $\Theta^{(t+1)}$ は $\Theta^{(t)}$ よりも大きな $J_{\boldsymbol{\gamma}}(\Theta)$ の値を与える．それゆえ，$\Theta^{(t+1)}$ と $\Theta^{(t)}$ の差異が十分小さくなるまで繰返し計算を行うことで $\Theta$ の値を推定できる．

以上に述べた生成モデルの学習法は，EM アルゴリズム[4] および確定的アニーリング EM アルゴリズム[20] を拡張した手法と見なすことができる．$\gamma=0$ かつ $\beta=1$ の場合，$Z_{\boldsymbol{\gamma}}(\boldsymbol{x}_m;\Theta,\hat{W})$ は，EM アルゴリズムで最大化させる生成モデルの混合[17] と定数 $C$ の積 $C\sum_{\boldsymbol{y}}p(\boldsymbol{x}_m,\boldsymbol{y};\Theta)$ になる．また，$\gamma=0$ の場合，$Z_{\boldsymbol{\gamma}}(\boldsymbol{x}_m;\Theta,\hat{W})$ は確定的アニーリング EM アルゴリズムで最大化させる関数と定数の積 $C\sum_{\boldsymbol{y}}p(\boldsymbol{x}_m,\boldsymbol{y};\Theta)^\beta$ となる．すなわち，提案法で用いる学習法は，これらの EM アルゴリズムを KL 距離で制約される条件付き確率分布を用いて改変した手法となっている．そこで，本学習アルゴリズムを制約条件付き確率 EM（restricted conditional probability EM，RCP-EM）アルゴリズムと名付ける．

### 3.2　統合の重みの推定

識別関数と生成モデルの統合の重み $\boldsymbol{\gamma}=(\gamma,\beta)^T$ は，分類器の条件付き確率分布 $R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x};\Theta,\hat{W})$ を確定させるために決定すべき未知パラメータである．提案法では，ラベルありデータ集合 $D_l$ を用いた条件付き確率モデル $R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x};\Theta,\hat{W})$ の尤度最大化学習に基づいて $\boldsymbol{\gamma}$ の値を与える．ただし，$R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x};\Theta,\hat{W})$ に含まれる $\hat{W}$ も $D_l$ を用いて得る推定値であるため，同じ $D_l$ を $\boldsymbol{\gamma}$ の推定に用いることで生成モデルの重み $\beta$ と比べて一方的に大きな識別関数の重み $\gamma$ を与えてしまう危険性がある．この偏った重みの学習を避けるために，削除補間法[10] を応用した以下の目的関数の最大化により $\boldsymbol{\gamma}$ の値を与える．

1. 訓練データ集合 $D_l = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$, $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$ を入力.
2. 初期化：$\Theta^{(0)}$, $t \leftarrow 0$.
3. 式 (3) を用いて $\hat{W}$ と $\hat{W}^{(-n)}$ を計算.
4. $\Theta^{(0)}$ と $\hat{W}^{(-n)}$ の下で式 (11) を用いて推定値 $\hat{\boldsymbol{\gamma}}$ を計算.
5. 以下の計算を収束するまで繰り返す.
   - $\hat{\boldsymbol{\gamma}}$ と $\hat{W}$ の下で式 (10) を用いて $\Theta^{(t+1)}$ を計算.
   - $\Theta^{(t+1)}$ と $\hat{W}^{(-n)}$ の下で式 (11) を用いて $\hat{\boldsymbol{\gamma}}$ を再計算.
   - $t \leftarrow t+1$.
6. $R_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{y}|\boldsymbol{x}; \Theta^{(t)}, \hat{W})$ を出力.

**図 3**　パラメータ学習アルゴリズムの概要
Fig. 3　Outline of parameter estimation algorithm.

$$J_\Theta(\boldsymbol{\gamma}) = \sum_{n=1}^{N} \log R_{\boldsymbol{\gamma}}(\boldsymbol{y}_n|\boldsymbol{x}_n; \Theta, \hat{W}^{(-n)}) + \log p(\boldsymbol{\gamma}) \tag{11}$$

$\hat{W}^{(-n)}$ は $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ を含まないラベルありデータのサブ集合から得られる $W$ の推定値を表す．$p(\boldsymbol{\gamma})$ は $\boldsymbol{\gamma}$ の事前確率分布である．提案法では，$\Theta$ の値を与えた下で，準ニュートン法の実現法の 1 つである L-BFGS アルゴリズム[14] を用いて $J_\Theta(\boldsymbol{\gamma})$ を最大化させる $\boldsymbol{\gamma}$ を求める．

しかし，前節で述べたように，$\Theta$ もまた $\boldsymbol{\gamma}$ の値を与えた下で，式 (10) で推定する未知パラメータである．すなわち，$\Theta$ と $\boldsymbol{\gamma}$ のパラメータ学習には相互に依存関係がある．このため，$J_{\boldsymbol{\gamma}}(\Theta)$ と $J_\Theta(\boldsymbol{\gamma})$ を同時に最大化させる $\Theta$ と $\boldsymbol{\gamma}$ を探索する必要がある．提案法では，式 (10) を最大化させる $\Theta^{(t+1)}$ と式 (11) を最大化させる $\boldsymbol{\gamma}$ を交互に繰り返し計算することでパラメータ学習を行う．**図 3** にパラメータ学習のアルゴリズムをまとめる．

## 4. 多重テキスト分類への適用法

3 章で述べた手法を実問題に適用するには，データの特徴に応じて設計した識別モデルと生成モデルを用いる必要がある．多重テキスト分類問題では，複数のカテゴリに属するテキストデータの同時確率分布 $p(\boldsymbol{x}, \boldsymbol{y})$ を直接モデル化した生成モデル[19),21)] を適用することが考えられる．この単純な適用法では，式 (10) による生成モデルのパラメータ学習のために，すべての可能なクラス（カテゴリの組合せ）の条件付き確率 $R(\boldsymbol{y}|\boldsymbol{x})$ をラベルなしデータごとに推定する必要がある．しかし，可能なクラスは $\boldsymbol{y} = \boldsymbol{0}$ を除く $2^K - 1$ 通り存在するため，$K$ が大きい場合に条件付き確率の計算量が膨大になる．そこで，本論文では，カテゴリラベルごとにテキストデータをモデル化したナイーブベイズ（NB）モデル[17)] を生成モデルとして用いることで計算量の問題を回避する．識別モデルには，文献 5)，16) 等でテキスト分類に用いられている対数線形モデルを利用する．以下に，NB モデルと対数線形モデルを適用して設計する多重テキスト分類器の詳細を述べる．

### 4.1 NB モデル

本適用法では，$k$ 番目のカテゴリラベルが付与される場合（$y_k = 1$）と付与されない場合（$y_k = 0$）に対して，以下のような生成モデルを設計する．

$$p(\boldsymbol{x}, y_k; \Theta_k) = p(\boldsymbol{x}|y_k; \Theta_k)\pi_k(y_k) \tag{12}$$

ここで，$\pi_k(y_k)$ は $y_k$ の確率を表し，$\sum_{y_k=0}^{1} \pi_k(y_k) = 1$ を満たす．$p(\boldsymbol{x}|y_k; \Theta_k)$ は $y_k$ の条件下での $\boldsymbol{x}$ の確率密度を表す．$p(\boldsymbol{x}|y_k; \Theta_k)$ のモデル化に，NB モデル[17)] を用いる．

NB モデルでは，文書に含まれる単語が独立に出現すると仮定し，$i$ 番目の単語の出現頻度 $x_i$ で与えられる特徴ベクトル $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_V)^T$ を用いて文書を表現する．ここで，$V$ は文書集合全体に含まれる単語の種類（語彙）の総数を表す．そして，$y_k$ での文書の確率密度を

$$p(\boldsymbol{x}|y_k; \Theta_k) \propto \prod_{i=1}^{V} \{\theta_{ki}(y_k)\}^{x_i} \tag{13}$$

で表される多項分布でモデル化する．ここで，$\theta_{ki}(y_k)$ は $y_k$ で $i$ 番目の単語が出現する確率を表し，$\theta_{ki}(y_k) > 0$，$\sum_{i=1}^{V} \theta_{ki}(y_k) = 1$ を満たす．$\Theta_k = [\theta_{ki}(y_k)]_{i,y_k}$ は，NB モデルのパラメータである．

### 4.2 対数線形モデルの学習

NB モデルと統合する識別モデルには，2.2 節で述べた対数線形モデルを利用する．対数線形モデルの識別関数は，パラメータ行列 $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, \ldots, \boldsymbol{w}_K]^T$ と $f_k(\boldsymbol{x}, y_k; \boldsymbol{w}_k) = y_k \boldsymbol{w}_k^T \boldsymbol{x}$ を用いて $f(\boldsymbol{x}, \boldsymbol{y}; W) = \sum_{k=1}^{K} f_k(\boldsymbol{x}, y_k; \boldsymbol{w}_k)$ で表せる．$\boldsymbol{y} \in \{1, 0\}^K$ のとき $\sum_{\boldsymbol{y}} \exp\left\{\sum_{k=1}^{K} f_k(\boldsymbol{x}, y_k; \boldsymbol{w}_k)\right\} = \prod_{k=1}^{K} \sum_{y_k=0}^{1} \exp\{f_k(\boldsymbol{x}, y_k; \boldsymbol{w}_k)\}$ であるため，式 (2) の $P(\boldsymbol{y}|\boldsymbol{x}; W)$ を以下のような $K$ 個のロジスティック回帰モデルの積で表せる．

$$P(\boldsymbol{y}|\boldsymbol{x}; W) = \prod_{k=1}^{K} \frac{\exp\{f_k(\boldsymbol{x}, y_k; \boldsymbol{w}_k)\}}{\sum_{y'_k=0}^{1} \exp\{f_k(\boldsymbol{x}, y'_k; \boldsymbol{w}_k)\}} \tag{14}$$

それゆえ，式 (3) を用いた $W$ の学習では，カテゴリラベルごとに個別に $\boldsymbol{w}_k$ の値を推定できる．

### 4.3 条件付き確率モデル

3.1 節で述べた手法に従い，NB モデル $p(\boldsymbol{x}|y_k;\Theta_k)$ と対数線形モデルの識別関数 $f_k(\boldsymbol{x},y_k;\hat{\boldsymbol{w}}_k)$ の統合に基づく多重分類器を以下の目的関数を最大化させる条件付き確率分布 $R(\boldsymbol{y}|\boldsymbol{x})$ でモデル化する．

$$J_{\boldsymbol{\gamma}}(R,\Theta)=\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\neq\boldsymbol{0}}R(\boldsymbol{y}|\boldsymbol{x}_m)\log p(\boldsymbol{x}_m|y_k;\Theta_k)\pi_k(y_k) -\frac{1}{\beta}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\neq\boldsymbol{0}}R(\boldsymbol{y}|\boldsymbol{x}_m)\log\frac{R(\boldsymbol{y}|\boldsymbol{x}_m)}{P_\gamma(\boldsymbol{y}|\boldsymbol{x}_m;\hat{W})}+\log p(\Theta) \tag{15}$$

ただし，$\Theta=\{\Theta_k\}_{k=1}^K$ である．多重分類は各データに 1 つ以上のカテゴリラベルを付与する問題であるため，本手法では $\boldsymbol{y}=\boldsymbol{0}$ を除く $\boldsymbol{y}$ に対して上記の目的関数と $R(\boldsymbol{y}|\boldsymbol{x})$ を与えることに注意．$\sum_{\boldsymbol{y}\neq\boldsymbol{0}}R(\boldsymbol{y}|\boldsymbol{x})=1$ の条件下でラグランジュ未定乗数法を用いると，多重分類器の条件付き確率モデル

$$R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x};\hat{W},\Theta,\boldsymbol{\mu})=\frac{\prod_{k=1}^{K}h_k(\boldsymbol{x},y_k;\hat{\boldsymbol{w}}_k,\Theta_k,\mu_k,\boldsymbol{\gamma})}{\sum_{\boldsymbol{y}'\neq\boldsymbol{0}}\prod_{k=1}^{K}h_k(\boldsymbol{x},y'_k;\hat{\boldsymbol{w}}_k,\Theta_k,\mu_k,\boldsymbol{\gamma})} \tag{16}$$

を得ることができる．ただし，

$$h_k(\boldsymbol{x},y_k;\hat{\boldsymbol{w}}_k,\Theta_k,\mu_k,\boldsymbol{\gamma})=\exp\{f_k(\boldsymbol{x},y_k;\hat{\boldsymbol{w}})\}^{\gamma}p(\boldsymbol{x}|y_k;\Theta_k)^{\beta}\exp(y_k)^{\mu_k} \tag{17}$$

かつ $\mu_k=\beta\log\pi_k(y_k=1)/\pi_k(y_k=0)$ である．$\boldsymbol{\gamma}=(\gamma,\beta)^T$ と $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_k,\ldots,\mu_K)^T$ の値には，3.2 節で述べた条件付き確率モデルの尤度最大化に基づく手法で得られる推定値を用いる．

$\Theta$ の推定には，式 (15) に式 (16) を代入して得られる目的関数の $Q$ 関数

$$Q_{\boldsymbol{\gamma}}(\Theta^{(t+1)},\Theta^{(t)})=\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\neq\boldsymbol{0}}R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x}_m;\hat{W},\Theta^{(t)},\boldsymbol{\mu})\log p(\boldsymbol{x}_m|y_k;\Theta_k^{(t+1)})+\log p(\Theta^{(t+1)}) \tag{18}$$

を用いる．この $Q$ 関数を最大化させる $\Theta^{(t+1)}$ を直接求めるには，ラベルなしデータ集合の $\boldsymbol{x}$ と $\boldsymbol{y}$ のすべての組合せに対して $R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x};\hat{W},\Theta^{(t)},\boldsymbol{\mu})$ を計算する必要がある．しかし，可能性のある $\boldsymbol{y}$ は $2^K-1$ 通り存在するため，$K$ が大きい問題では $\Theta^{(t+1)}$ を推定するのに要する計算量が膨大になる．そこで，計算量を削減するために，カテゴリラベルごとに独立な関数

$$s_k(\boldsymbol{x},y_k;\Theta_k)=\frac{h_k(\boldsymbol{x},y_k;\hat{\boldsymbol{w}}_k,\Theta_k,\mu_k,\boldsymbol{\gamma})}{\prod_{y'_k=0}^{1}h_k(\boldsymbol{x},y'_k;\hat{\boldsymbol{w}}_k,\Theta_k,\mu_k,\boldsymbol{\gamma})} \tag{19}$$

を利用する．$\sum_{y_k=0}^{1}s_k(\boldsymbol{x},y_k;\Theta_k)=1$ かつ NB モデルのパラメータ $\Theta_k$ がカテゴリラベルごとに独立であるため，$Q$ 関数を $s_k(\boldsymbol{x},y_k;\Theta_k)$ と $p(\Theta_k)$ を用いて以下のように書き換えることができる．

$$Q_{\boldsymbol{\gamma}}(\Theta^{(t+1)},\Theta^{(t)})=\sum_{k=1}^{K}\left\{\sum_{m=1}^{M}\sum_{y_k=0}^{1}q_k(\boldsymbol{x}_m,y_k;\Theta^{(t)})\log p(\boldsymbol{x}_m|y_k;\Theta_k^{(t+1)})+\log p(\Theta_k^{(t+1)})\right\} \tag{20}$$

ただし，

$$q_k(\boldsymbol{x}_m,y_k=1;\Theta^{(t)})=\frac{s_k(\boldsymbol{x}_m,y_k=1;\Theta_k^{(t)})}{1-\prod_{k'=1}^{K}s_{k'}(\boldsymbol{x}_m,y'_{k'}=0;\Theta_{k'}^{(t)})} \tag{21}$$

かつ $q_k(\boldsymbol{x}_m,y_k=0;\Theta^{(t)})=1-q_k(\boldsymbol{x}_m,y_k=1;\Theta^{(t)})$ である（導出方法については付録 A.1 を参照）．式 (21) から得られる $q_k(\boldsymbol{x}_m,y_k;\Theta^{(t)})$ を用いることで，$\Theta$ の推定に要する計算量を $K$ に比例する量に削減できる．

提案法では，NB モデルのパラメータ値を推定するのにディリクレ事前確率分布 $p(\Theta_k)\propto\prod_{y_k=0}^{1}\prod_{i=1}^{V}\{\theta_{ki}(y_k)\}^{\xi-1}$ を用いる．このとき，$Q$ 関数を最大化させる $\Theta_k^{(t+1)}=[\theta_{ki}^{(t+1)}(y_k)]_{i,y_k}$ を以下の式で表せる．

$$\theta_{ki}^{(t+1)}(y_k)=\frac{\sum_{m=1}^{M}q_k(\boldsymbol{x}_m,y_k;\Theta^{(t)})x_{mi}+\xi-1}{\sum_{i=1}^{V}\left\{\sum_{m=1}^{M}q_k(\boldsymbol{x}_m,y_k;\Theta^{(t)})x_{mi}+\xi-1\right\}} \tag{22}$$

$\xi(>1)$ はディリクレ事前確率分布のハイパーパラメータである．$\xi$ の値を事前に設定し，式 (22) を用いて $\theta_{ki}^{(t+1)}(y_k)$ を計算することで NB モデルのパラメータ値を推定する．

## 5. 実　　験

### 5.1 テストコレクション

提案法の分類性能を実データを用いた実験で評価した．実験には，多重テキスト分類のベンチマークテストによく用いられる 3 つのテストコレクション（Reuters-21578，RCV1-2，WIPO-alpha）を用いた．

Reuters-21578 [⋆1]データセット（Reuters）は，135 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り，記事を多く含む 10 トピックカテゴリがベンチマークテストによく用いられる[17]．そこで，この 10 トピックカテゴリに含まれる記事を用いて実験を行った．投稿時期の早い 6,490 記事の中から学習に用いるラベルありデータとラベルなしデータを選択し，投稿時期の遅い 2,545 記事をテストデータとして用いた．2 つ以上のカテゴリラベルが付与されている記事の比率（多重度）は 9%であった．記事の特徴ベクトルの作成には，単語の出現頻度を用いた．ただし，停止語（stop word）リスト[18] に含まれる単語と 1 つの記事のみに出現する低頻度語を除外した．特徴ベクトルの次元数（語彙の総数）は 14,586 であった．

RCV1-2 [⋆2]データセットは，103 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り，訓練データセットとテストデータセットから構成されている[12]．実験には，訓練データセット内の多くの記事が属する 20 トピックカテゴリを利用した．訓練データセットとテストデータセットからこれらのカテゴリに属する 23,149 記事と 20,000 記事を抽出して実験に用いた．抽出した記事の多重度は 76%であった．Reuters と同様に停止語と低頻度語を除外した結果，特徴ベクトルの次元数は 39,242 であった．

WIPO-alpha データセット（WIPO）は，国際特許分類（IPC）体系で分類された特許文書からなり，訓練データセットとテストデータセットから構成されている[7]．評価実験では，訓練データセット内の多くの文書が属する 20 クラスを利用した．IPC は，セクション，クラス，サブクラス，グループ，サブグループの 5 階層から構成され，クラスは 2 階層目に相当する．実験には，訓練データセットとテストデータセットから 20 クラスのいずれかに属する 29,626 文書と 20,206 文書を抽出して用いた．抽出した文書の多重度は 18%であった．文書の特徴ベクトルを作成するのに，title と abstract に含まれる単語の頻度を用いた．Reuters と同様に停止語と低頻度語を除外した結果，特徴ベクトルの次元数は 35,941 であった．

### 5.2 実験設定

4 章で示した RCP-EM アルゴリズムに基づく多重分類器（RCP）の性能を評価するため，実験では，5 つの半教師あり学習に基づく分類器（JLL/MCL，LL/MER，TSVM，NB/DC，NB/EM-λ）および教師あり学習に基づく分類器（LL）の性能と比較した．

RCP の学習すべきパラメータは $W$，$\Theta$，$\boldsymbol{\gamma}$，$\boldsymbol{\mu}$ である．$W$ の推定にはガウス事前確率分布 $p(W) \propto \prod_{k=1}^{K} \exp(-\boldsymbol{w}_k^T \boldsymbol{w}_k / 2\sigma^2)$ を用い，$\boldsymbol{\gamma}$ と $\boldsymbol{\mu}$ の推定にはガウス事前確率分布 $p(\boldsymbol{\gamma}, \boldsymbol{\mu}) \propto \exp\{-(\gamma-1)^2/2 - (\beta-1)^2/2\} \prod_{k=1}^{K} \exp(-\mu_k^2/2\rho^2)$ を用いた．また，式 (22) 中の $\xi$ に定数値を与えて $\Theta$ の学習を行った．本実験では，事前確率分布のハイパーパラメータ値を候補値 $\sigma \in \{10^{n/2}\}_{n=3}^{5}$，$\rho \in \{10^{n/2}\}_{n=-2}^{2}$，$\xi \in \{1 + 10^{-n}\}_{n=1}^{4}$ の中から選択した．

NB/DC は，文献 8) で提案された生成モデルと識別学習のハイブリッド法に基づく単一ラベル分類器であり，識別学習に基づく NB モデルの重み付き統合により構築される．本実験では，NB/DC をカテゴリラベルの付与を判定する 2 値分類に用いた．NB/DC の学習には，RCP の $\rho$ と $\xi$ に相当するハイパーパラメータ値を設定する必要がある．本実験では，これらのハイパーパラメータ値を RCP と同じ候補値の中から選択した．

NB/EM-λ は，NB モデルに基づく単一ラベル分類器であり，EM-λ アルゴリズム[17] を用いて NB モデルを学習させる．本実験では，NB/EM-λ をカテゴリラベルの付与を判定する 2 値分類に用いた．NB/EM-λ では，ラベルありデータに対する NB モデルの対数尤度とラベルなしデータに対する対数尤度の重み付き和を最大化させるようにモデルを学習させる．本実験では，ラベルなしデータの寄与度を与える重みパラメータの値を 13 候補 $\lambda \in \{0.01, 0.02, 0.05, \{n \times 0.1\}_{n=1}^{10}\}$ の中から選択して NB モデルを学習させた．また，NB モデルのパラメータの事前確率分布にはディリクレ分布を用い，ディリクレ分布のハイパーパラメータ値 $\xi = \eta + 1$ を 10 候補 $\eta \in \{1, \{1 \times 10^{-n}, 2 \times 10^{-n}, 5 \times 10^{-n}\}_{n=1}^{3}\}$ の中から選択した．

JLL/MCL は，文献 5) に従い，同時確率分布の対数線形モデル（joint log-linear model, JLL モデル）$p(\boldsymbol{x}, \boldsymbol{y})$ を用いて設計した分類器であり，multi-conditional learning（MCL）と呼ばれる手法で JLL モデルを学習させる．この手法では，同時確率分布の周辺化で得られる $p(\boldsymbol{x})$ と $p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{y})/p(\boldsymbol{x})$ で得られる条件付き確率分布を用いて半教師あり学習を行う．具体的には，ラベルありデータに対する $p(\boldsymbol{y}|\boldsymbol{x})$ の対数尤度とラベルなしデータを含む訓練データに対する $p(\boldsymbol{x})$ の対数尤度の重み付き和を最大化させるように JLL モデルを学習させる．本実験では，$p(\boldsymbol{y}|\boldsymbol{x})$ の重みパラメータの値を $\alpha = 1$ に設定し，$p(\boldsymbol{x})$ の重みパラメータの値を 10 候補 $\beta \in \{10^{-n}\}_{n=0}^{9}$ の中から選択して JLL モデルを学習させた．また，JLL モデルのパラメータ $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, \ldots, \boldsymbol{w}_K]^T$ の事前確率分布にはガウス分布 $p(W) \propto \prod_{k=1}^{K} \exp(-\boldsymbol{w}_k^T \boldsymbol{w}_k / 2\sigma^2)$ を用い，ハイパーパラメータ値を 9 候補 $\sigma \in \{10^{n/2}\}_{n=1}^{9}$ の中から選択した．

LL/MER は，式 (14) で表される条件付き確率分布の対数線形（LL）モデルに基づく分類

⋆1 http://www.daviddlewis.com/resources/testcollections/reuters21578/reuters21578.tar.gz
⋆2 http://www.daviddlewis.com/resources/testcollections/rcv1/

器であり，LL モデルを最小エントロピー正則化（minimum entropy regularization，MER）法で学習[9] させる．この手法では，ラベルありデータに対するモデルの対数尤度と，ラベルなしデータに対する条件付きエントロピーの負値の重み付き和を最大化させるようにモデルを学習させる．本実験では，ラベルなしデータの寄与度を与える重みパラメータの値を 16 候補 $\lambda \in \{1, \{1\times10^{-n}, 2\times10^{-n}, 5\times10^{-n}\}_{n=1}^{4}\}$ の中から選択して LL モデルを学習させた．また，JLL/MCL と同じガウス事前確率分布とハイパーパラメータの候補値を用いた．

LL は，式 (14) で表される条件付き確率分布の対数線形（LL）モデルに基づく分類器であり，LL モデルをラベルありデータのみで学習させる．パラメータ学習には，JLL/MCL，LL/MER と同じガウス事前確率分布とハイパーパラメータの候補値を用いた．

以上に述べた 6 分類器では，ラベルありデータによる 10 分割交差検定法を用いて，パラメータの事前確率分布のハイパーパラメータと重みパラメータの値を決定した．また，NB/EM-$\lambda$ を除く RCP，NB/DC，JLL/MCL，LL/MER，LL では，データの特徴ベクトルを $\sum_{i=1}^{V} x_i = 1$ になるように正規化して適用した．

さらに，本実験では，transductive SVM（TSVM）をカテゴリラベルの付与を判定する 2 値分類に用いた場合の分類性能も調べた．TSVM の実行プログラムには，concave-convex procedure（CCCP）を用いてパラメータ学習を行う Univer SVM [⋆1] を利用した[3]．文献 5) の設定に従い，ラベルなしデータ用の Ramp Loss 関数のパラメータ値を $s = -0.5$ に設定し，線形カーネルを用いた．実験では，データの特徴ベクトルを $\sum_{i=1}^{V} x_i = 10$ になるように正規化し，ラベルありデータに対するコストパラメータを 7 候補 $C \in \{10^n\}_{n=-3}^{3}$ の中から選択した．また，ラベルなしデータに対するコストパラメータを 5 候補 $C^* \in \{10^n\}_{n=-3}^{1}$ の中から選択した．公平な比較のためにはラベルありデータによる交差検定法を用いてコストパラメータ値を選択すべきであるが，TSVM の学習には多大な計算時間を要する．そこで，テストデータに対して最良の分類性能を与えるコストパラメータ値を選択した．

分類器の性能評価には，多重分類問題でよく用いられる分類精度（CA），マイクロ平均 $F$ 値（$F_\mu$），マクロ平均 $F$ 値（$F_M$）の 3 つの評価値とラベリング $F$ 値（$F_L$）を用いた．これらの評価値は以下の式で計算される．

$$CA = \frac{1}{N_T}\sum_{\tau=1}^{N_T} I\boldsymbol{y}_\tau(\hat{\boldsymbol{y}}_\tau) \tag{23}$$

$$F_\mu = \frac{2\sum_{\tau=1}^{N_T}\sum_{k=1}^{K} y_{\tau k}\hat{y}_{\tau k}}{\sum_{\tau=1}^{N_T}\sum_{k=1}^{K} y_{\tau k} + \sum_{\tau=1}^{N_T}\sum_{k=1}^{K} \hat{y}_{\tau k}} \tag{24}$$

$$F_M = \frac{1}{K}\sum_{k=1}^{K}\frac{2\sum_{\tau=1}^{N_T} y_{\tau k}\hat{y}_{\tau k}}{\sum_{\tau=1}^{N_T} y_{\tau k} + \sum_{\tau=1}^{N_T} \hat{y}_{\tau k}} \tag{25}$$

$$F_L = \frac{1}{N_T}\sum_{\tau=1}^{N_T}\frac{2\sum_{k=1}^{K} y_{\tau k}\hat{y}_{\tau k}}{\sum_{k=1}^{K} y_{\tau k} + \sum_{k=1}^{K} \hat{y}_{\tau k}} \tag{26}$$

式中の $N_T$ はテストデータの個数を表す．また，$\boldsymbol{y}_\tau = (y_{\tau 1}, \ldots, y_{\tau k}, \ldots, y_{\tau K})^T$ は正解のクラスベクトルを，$\hat{\boldsymbol{y}}_\tau = (\hat{y}_{\tau 1}, \ldots, \hat{y}_{\tau k}, \ldots, \hat{y}_{\tau K})^T$ は分類器で推定されるクラスベクトルを表す．$F_M$ 値がカテゴリに属するデータを検索する精度の高さを表すのに対し，$F_L$ 値はデータに付与すべきカテゴリラベルを検索する精度の高さを表す．

本論文では，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う実験を同一条件で 10 回繰り返し，その実験から得られる評価値の平均値を用いて分類器の性能を比較する．実験では，訓練データセットから選択した 5,000 個のラベルなしデータと $N$ 個のラベルありデータを分類器の学習に用い，テストデータセットから選択した $N_T$ 個のテストデータを分類器の性能評価に用いた．Reuters ではすべてのテストデータ（$N_T = 2545$）を分類器の性能評価に用い，RCV1-2 と WIPO では $N_T = 5000$ とした．

### 5.3　実験結果と考察

#### 5.3.1　分 類 性 能

**表 1**，**表 2**，**表 3** に，各条件で 10 回の実験を行って得られた分類器の評価値（$CA$ 値，$F_\mu$ 値，$F_M$ 値，$F_L$ 値）の平均値を示す．括弧内の数値は標準偏差を示す．

RCP の分類性能は，ほとんどの条件で JLL/MCL，LL/MER よりも高かった．JLL/MCL と LL/MER ではラベルあり・なしデータで学習させた対数線形モデルのみを用いて分類器を構築するのに対し，RCP ではラベルなしデータによる分類器の学習に生成モデルを利用する．すなわち，RCP では，データの種類に応じて設計される生成モデルの確率分布をラベルなしデータの分布の事前知識として利用する．RCP は，この事前知識をラベルなしデータによる学習に利用する点で有利であったと考えられる．

また，RCP と JLL/MCL，LL/MER では，ラベルありデータとラベルなしデータの学習への寄与度を与える重みの決定法に違いがある．RCP では，3.2 節で述べたように，ラベルありデータで学習させる識別モデルとラベルなしデータで学習させる生成モデルを勾

⋆1 http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html

**表 1**　Reuters データセットでの各分類器の評価値（$CA$ 値，$F_\mu$ 値，$F_M$ 値，$F_L$ 値，単位：%）
Table 1　$CA$-, $F_\mu$-, $F_M$-, and $F_L$-scores (%) of each classifier for Reuters data set.

| | $N$ | RCP | JLL/MCL | LL/MER | TSVM | NB/DC | NB/EM-$\lambda$ | LL |
|---|---|---|---|---|---|---|---|---|
| $CA$ | 80 | **83.4** (1.9) | 74.7 (2.6) | 72.4 (1.1) | 72.6 (3.0) | 74.8 (2.0) | 73.0 (3.3) | 72.1 (1.4) |
| | 320 | **88.2** (0.8) | 82.6 (0.6) | 82.8 (0.6) | 83.4 (0.9) | 80.5 (1.4) | 78.2 (1.4) | 82.5 (0.5) |
| | 1,280 | **90.7** (0.4) | 87.1 (0.7) | 87.7 (0.5) | 88.1 (0.3) | 83.8 (0.4) | 82.3 (0.6) | 87.4 (0.4) |
| $F_\mu$ | 80 | **87.9** (1.9) | 82.6 (2.4) | 82.1 (1.0) | 83.1 (2.0) | 83.5 (2.1) | 81.4 (2.7) | 82.2 (1.1) |
| | 320 | **92.1** (0.6) | 89.2 (0.5) | 89.4 (0.4) | 90.5 (0.6) | 88.2 (0.8) | 86.8 (0.9) | 89.3 (0.4) |
| | 1,280 | **94.1** (0.2) | 92.6 (0.3) | 93.1 (0.2) | 93.3 (0.2) | 90.8 (0.2) | 89.2 (0.4) | 92.7 (0.3) |
| $F_M$ | 80 | **74.7** (3.5) | 59.4 (3.8) | 62.4 (4.0) | 71.0 (3.2) | 67.3 (3.4) | 67.4 (3.2) | 62.6 (4.0) |
| | 320 | **83.1** (1.3) | 76.1 (1.9) | 75.4 (1.6) | 81.6 (2.0) | 76.4 (1.4) | 76.1 (1.8) | 75.6 (1.7) |
| | 1,280 | **87.5** (0.4) | 85.6 (0.8) | 85.3 (0.5) | 86.9 (0.5) | 80.6 (0.7) | 79.1 (0.8) | 85.1 (0.5) |
| $F_L$ | 80 | **89.8** (1.7) | 81.1 (2.5) | 76.5 (1.3) | 81.4 (2.3) | 84.1 (1.3) | 82.7 (1.6) | 76.6 (1.3) |
| | 320 | **93.3** (0.5) | 90.0 (1.0) | 87.2 (0.7) | 89.2 (0.9) | 89.2 (0.7) | 87.8 (0.6) | 87.2 (0.7) |
| | 1,280 | **95.1** (0.1) | 92.4 (0.2) | 92.2 (0.4) | 92.7 (0.4) | 90.6 (0.2) | 90.6 (0.3) | 91.9 (0.4) |

**表 2**　RCV1-2 データセットでの各分類器の評価値（$CA$ 値，$F_\mu$ 値，$F_M$ 値，$F_L$ 値，単位：%）
Table 2　$CA$-, $F_\mu$-, $F_M$-, and $F_L$-scores (%) of each classifier for RCV1-2 data set.

| | $N$ | RCP | JLL/MCL | LL/MER | TSVM | NB/DC | NB/EM-$\lambda$ | LL |
|---|---|---|---|---|---|---|---|---|
| $CA$ | 160 | 31.8 (1.8) | 30.9 (1.5) | 30.7 (1.7) | **33.2** (2.0) | 29.0 (2.0) | 25.3 (3.2) | 30.7 (1.7) |
| | 640 | 45.3 (0.9) | 45.5 (1.2) | 45.1 (0.8) | **46.3** (1.0) | 39.2 (0.7) | 35.4 (1.2) | 44.8 (1.0) |
| | 2,560 | **55.4** (0.7) | 54.6 (0.9) | 54.6 (0.8) | 54.6 (0.6) | 45.9 (0.6) | 40.6 (0.8) | 54.5 (0.9) |
| $F_\mu$ | 160 | 66.3 (1.4) | 64.6 (1.2) | 64.6 (1.3) | **69.5** (1.2) | 66.6 (1.3) | 63.6 (2.0) | 64.8 (1.3) |
| | 640 | **76.8** (0.6) | 76.1 (0.7) | 75.2 (0.7) | **76.8** (0.3) | 74.1 (0.7) | 71.5 (0.7) | 75.2 (0.7) |
| | 2,560 | **81.9** (0.5) | 80.6 (0.6) | 80.8 (0.5) | 81.1 (0.4) | 77.5 (0.6) | 75.1 (0.5) | 80.8 (0.5) |
| $F_M$ | 160 | 50.9 (2.6) | 43.7 (1.8) | 43.2 (2.3) | **58.1** (2.0) | 53.3 (1.7) | 47.9 (2.4) | 43.5 (2.4) |
| | 640 | 66.6 (1.7) | 64.9 (1.5) | 62.3 (1.7) | **67.2** (0.6) | 63.7 (1.1) | 59.7 (1.0) | 62.4 (1.7) |
| | 2,560 | **74.4** (0.9) | 73.2 (0.7) | 72.5 (0.9) | 71.7 (0.9) | 68.9 (0.7) | 64.4 (0.6) | 72.6 (0.8) |
| $F_L$ | 160 | 67.4 (1.3) | 63.8 (1.2) | 61.9 (1.5) | **68.5** (1.4) | 65.0 (1.3) | 62.4 (2.4) | 62.0 (1.6) |
| | 640 | **76.7** (0.5) | 74.4 (0.7) | 73.4 (0.8) | 76.1 (0.4) | 73.0 (0.7) | 71.6 (0.6) | 73.4 (0.8) |
| | 2,560 | **81.7** (0.4) | 79.7 (0.5) | 79.7 (0.6) | 80.6 (0.4) | 77.0 (0.4) | 75.4 (0.4) | 79.6 (0.5) |

配法で推定した重み $\boldsymbol{\gamma}$ で統合する．一方，JLL/MCL と LL/MER では，重みパラメータ $\beta$，$\lambda$ の値を候補値の中から選択して決定する必要がある．モデル統合の重みを勾配法で推定できる RCP の特徴が，高性能な分類器を得るのに有利であるといえる．

RCP の分類性能は，Reuters と WIPO では TSVM よりも高かったが，RCV1-2 では必ずしも高いとは限らなかった．この原因を考察するため，学習に用いたラベルなしデータが

**表 3**　WIPO データセットでの各分類器の評価値（$CA$ 値，$F_\mu$ 値，$F_M$ 値，$F_L$ 値，単位：%）
Table 3　$CA$-, $F_\mu$-, $F_M$-, and $F_L$-scores (%) of each classifier for WIPO data set.

| | $N$ | RCP | JLL/MCL | LL/MER | TSVM | NB/DC | NB/EM-$\lambda$ | LL |
|---|---|---|---|---|---|---|---|---|
| $CA$ | 160 | **37.9** (3.0) | 18.1 (3.8) | 11.9 (2.5) | 19.4 (1.9) | 16.5 (2.0) | 9.7 (5.7) | 12.1 (2.4) |
| | 640 | **46.8** (1.4) | 28.5 (1.9) | 24.8 (1.9) | 31.3 (1.6) | 27.2 (2.3) | 21.5 (2.6) | 24.8 (1.9) |
| | 2,560 | **53.1** (0.7) | 37.6 (1.3) | 36.9 (1.1) | 40.6 (1.3) | 34.1 (1.4) | 30.4 (1.7) | 36.9 (1.1) |
| $F_\mu$ | 160 | **49.2** (1.9) | 38.5 (5.0) | 28.9 (3.0) | 40.3 (2.1) | 42.0 (1.7) | 36.6 (2.8) | 29.1 (3.0) |
| | 640 | **57.1** (2.2) | 49.4 (1.9) | 45.0 (1.7) | 53.4 (1.1) | 53.0 (1.6) | 47.3 (2.1) | 45.1 (1.7) |
| | 2,560 | **63.7** (0.9) | 58.6 (1.0) | 57.7 (1.1) | 61.9 (0.8) | 60.1 (0.8) | 58.2 (0.8) | 57.7 (1.0) |
| $F_M$ | 160 | **36.4** (1.4) | 27.4 (1.2) | 18.0 (1.5) | 32.4 (1.7) | 28.8 (1.5) | 22.1 (1.4) | 18.3 (1.4) |
| | 640 | **48.0** (1.5) | 39.7 (1.2) | 34.4 (1.7) | 45.9 (1.3) | 41.9 (1.3) | 35.9 (1.3) | 34.6 (1.6) |
| | 2,560 | **56.3** (0.7) | 49.6 (0.7) | 49.2 (1.0) | 54.1 (1.0) | 52.1 (0.8) | 48.3 (0.7) | 49.3 (1.0) |
| $F_L$ | 160 | **47.1** (2.1) | 36.3 (7.1) | 17.0 (2.2) | 31.9 (2.4) | 35.4 (1.6) | 25.3 (2.8) | 17.2 (2.1) |
| | 640 | **55.6** (2.2) | 48.3 (2.1) | 32.4 (1.9) | 44.7 (1.5) | 45.7 (1.9) | 41.7 (1.1) | 32.5 (1.9) |
| | 2,560 | **62.8** (1.0) | 55.8 (0.9) | 48.7 (1.2) | 55.4 (1.1) | 52.8 (1.0) | 53.1 (0.8) | 48.8 (1.2) |

**表 4**　各テストコレクションのクラス内分散 $V_w$ とクラス間分散 $V_b$
Table 4　Within-class and between-class variances ($V_w$ and $V_b$) for each test collection.

| | $V_w$ | $V_b$ |
|---|---|---|
| Reuters | 0.049 | 0.0050 |
| RCV1-2 | 0.027 | 0.0030 |
| WIPO | 0.042 | 0.0029 |

属するカテゴリの正解を用いて，ラベルなしデータのクラス内分散 $V_w$ とクラス間分散 $V_b$ を以下の式を用いて調べた．

$$V_w = \frac{1}{M}\sum_{\boldsymbol{y}}\sum_{s=1}^{S_{\boldsymbol{y}}}(\boldsymbol{x}_{\boldsymbol{y}s}-\boldsymbol{m}_{\boldsymbol{y}})^T(\boldsymbol{x}_{\boldsymbol{y}s}-\boldsymbol{m}_{\boldsymbol{y}}) \tag{27}$$

$$V_b = \frac{1}{M}\sum_{\boldsymbol{y}} S_{\boldsymbol{y}}(\boldsymbol{m}_{\boldsymbol{y}}-\boldsymbol{m})^T(\boldsymbol{m}_{\boldsymbol{y}}-\boldsymbol{m}) \tag{28}$$

$S_{\boldsymbol{y}}$ はクラスが $\boldsymbol{y}$ であるデータ $\boldsymbol{x}_{\boldsymbol{y}s}$ の数を表し，$\boldsymbol{m}_{\boldsymbol{y}}$ は平均ベクトル $\boldsymbol{m}_{\boldsymbol{y}} = \sum_{s=1}^{S_{\boldsymbol{y}}}\boldsymbol{x}_{\boldsymbol{y}s}/S_{\boldsymbol{y}}$ を表す．$\boldsymbol{m}$ はすべてのラベルなしデータの平均ベクトルを表す．$\boldsymbol{x}_{\boldsymbol{y}s}$ を $\sum_{i=1}^{V} x_{\boldsymbol{y}si} = 1$ で正規化し，10 回の実験で用いたラベルなしデータセットのクラス内分散，クラス間分散の平均値を計算した結果を**表 4** に示す．表より，各手法で高い分類精度が得られた Reuters では RCV1-2 と WIPO よりもクラス間分散が大きかった．一方，ラベルありデータが少数

の場合に RCP よりも TSVM の分類精度が高かった RCV1-2 では，ラベルなしデータのクラス内分散が小さかった．TSVM は，属するクラスが未知のラベルなしデータを分布密度が低い領域で分けるようにクラス境界を学習する．したがって，TSVM では，ラベルなしデータが各クラスで密集しているほど，同一クラスのラベルなしデータを分離させるクラス境界を与えるリスクが小さいと考えられる．逆に，クラス内分散が大きい場合，同一クラスに属するラベルなしデータの間で分布が疎な領域にクラス境界を与えてしまうリスクが大きいと考えられる．それゆえ，クラス内分散が小さい RCV1-2 では，生成モデルを用いてラベルなしデータのクラス境界を間接的に学習する RCP よりも，ラベルなしデータを疎な領域で分離させる TSVM の方が，クラス境界を学習するのに有利な場合があった，と推測できる．

RCP は NB/DC よりも分類性能が高く，NB/DC は NB/EM-$\lambda$ よりも分類性能が高かった．NB/DC の $CA$ 値は，ラベルありデータが少数の場合を除いて，識別モデルに基づく JLL/MCL と LL/MER，TSVM よりも低い傾向があった．NB/DC は，RCP と同様に複数のモデルの統合により分類器を与えるが，識別モデルと生成モデルの統合に基づく RCP とは異なり，統合するモデルに生成モデルのみを用いる．実験結果は，生成モデルのみの統合に基づく NB/DC では，高性能な多重分類器を得るのに限界があることを示唆している．

### 5.3.2　複数ラベルデータの分類精度

**表 5** に，2 つ以上のカテゴリラベルが付与された（複数ラベルの）テストデータに対する各手法の分類精度 $CA$ を示す．括弧内の数値は，すべてのテストデータに対する分類精度を表す．表 5 の (a) と (b) では，5.2 節で述べた交差検定法で各手法のハイパーパラメータ値と重みパラメータ値を設定するのに用いたデータが異なる．表 5 (a) は，表 1 〜 3 と同様に，すべてのラベルありデータを用いた場合の結果を表すのに対し，表 5 (b) は複数ラベルのラベルありデータのみを用いた場合の結果を表す．ただし，TSVM では，5.2 節で述べた理由で，ラベルありデータではなくテストデータを用いてコストパラメータ値を設定した．

表 5 (a) より，すべてのラベルありデータを用いてハイパーパラメータを調節した場合，複数ラベルのテストデータに対する RCP の分類精度は，他手法よりも必ずしも高いとは限らなかった．しかし，表 5 (b) に示すように，複数ラベルのラベルありデータを用いてハイパーパラメータを調節した場合，Reuters と WIPO では，複数ラベルのテストデータに対する RCP の分類精度が JLL/MCL，LL/MER，TSVM と比較してほぼ同等か高い傾向があった．また，Reuters と WIPO では，複数ラベルのラベルありデータでハイパーパラメータを調節しても，RCP によってすべてのテストデータに対して高い分類精度が得られる傾向は変わらなかった．RCV1-2 では，複数ラベルのテストデータに対する分類精度とすべてのテストデータに対する分類精度がともに，$N = 160$，640 の場合に TSVM の方が RCP よりも高く，逆に $N = 2560$ では RCP の方が TSVM よりも高かった．

表 5 (b) では，Reuters の $N = 80$ の場合に，複数ラベルのテストデータに対する分類精度が RCP よりも TSVM で高かった．これは，Reuters では複数ラベルのデータの割合が少なく，交差検定法による調節により RCP のハイパーパラメータ値が少数のラベルありデータに過適合したことが原因であると考えられる．TSVM では，テストデータを用いてコストパラメータ値を設定しており，このような過適合の問題はない．TSVM と同様の方法で

**表 5**　複数ラベルのテストデータとすべてのテストデータ（括弧内）の分類精度（%）
Table 5　$CA$-scores (%) for test samples assigned to two more category labels and for all test samples (within parenthesis).

(a) すべてのラベルありデータでハイパーパラメータを調節した場合

| | $N$ | RCP | JLL/MCL | LL/MER | TSVM | NB/DC | NB/EM-$\lambda$ | LL |
|---|---|---|---|---|---|---|---|---|
| (i) | 80 | 26.9 (**83.4**) | 19.2 (74.7) | 22.8 (72.4) | **28.2** (72.6) | 26.5 (74.8) | 24.4 (73.0) | 22.8 (72.1) |
| | 320 | 43.8 (**88.2**) | 29.8 (82.6) | 33.4 (82.8) | **43.9** (83.4) | 37.5 (80.5) | 32.3 (78.2) | 33.9 (82.5) |
| | 1,280 | 54.2 (**90.7**) | 50.3 (87.1) | 49.9 (87.7) | **55.3** (88.1) | 47.8 (83.8) | 41.1 (82.3) | 50.5 (87.4) |
| (ii) | 160 | 26.0 (31.8) | 22.5 (30.9) | 21.9 (30.7) | **27.7** (**33.2**) | 23.1 (29.0) | 17.6 (25.3) | 22.2 (30.7) |
| | 640 | 38.3 (45.3) | 40.0 (45.5) | 38.5 (45.1) | **40.4** (**46.3**) | 32.1 (39.2) | 28.9 (35.4) | 38.6 (44.8) |
| | 2,560 | 49.4 (**55.4**) | 48.9 (54.6) | 48.9 (54.6) | **49.5** (54.6) | 38.8 (45.9) | 35.8 (40.6) | 49.1 (54.5) |
| (iii) | 160 | **21.3** (**37.9**) | 19.0 (18.1) | 13.8 (11.9) | 16.6 (19.4) | 16.5 (16.5) | 9.2 (9.7) | 13.9 (12.1) |
| | 640 | 26.1 (**46.8**) | 15.5 (28.5) | 20.4 (24.8) | 26.2 (31.3) | **27.0** (27.2) | 13.2 (21.5) | 20.1 (24.8) |
| | 2,560 | 26.4 (**53.1**) | 24.6 (37.6) | 24.5 (36.9) | 29.3 (40.6) | **32.4** (34.1) | 28.2 (30.4) | 24.6 (36.9) |

(i) Reuters, (ii) RCV1-2, (iii) WIPO

(b) 複数ラベルのラベルありデータでハイパーパラメータを調節した場合

| | $N$ | RCP | JLL/MCL | LL/MER | TSVM | NB/DC | NB/EM-$\lambda$ | LL |
|---|---|---|---|---|---|---|---|---|
| (i) | 80 | 30.0 (**81.4**) | 20.4 (64.2) | 21.4 (71.4) | **30.7** (71.6) | 27.3 (74.9) | 24.2 (70.0) | 21.4 (70.9) |
| | 320 | **47.6** (**87.2**) | 34.3 (81.0) | 33.0 (82.5) | 45.3 (83.2) | 39.2 (79.5) | 35.9 (76.4) | 34.0 (82.5) |
| | 1,280 | **58.7** (**90.0**) | 54.3 (87.2) | 51.1 (87.6) | 57.4 (87.6) | 48.5 (82.7) | 41.6 (79.6) | 51.3 (87.5) |
| (ii) | 160 | 27.2 (31.9) | 22.5 (30.9) | 22.1 (30.8) | **28.3** (**33.0**) | 24.2 (29.2) | 19.9 (24.8) | 22.5 (30.9) |
| | 640 | 40.7 (45.3) | 40.8 (45.9) | 38.7 (44.9) | **40.9** (**45.8**) | 32.8 (38.8) | 30.7 (35.4) | 38.9 (44.8) |
| | 2,560 | **50.3** (**55.4**) | 48.7 (53.7) | 48.7 (54.2) | 49.5 (54.6) | 39.1 (45.5) | 35.8 (40.5) | 49.0 (54.3) |
| (iii) | 160 | 23.6 (**33.2**) | 22.9 (13.2) | 13.1 (9.9) | 18.9 (12.3) | 21.0 (15.3) | **25.6** (10.0) | 13.2 (10.6) |
| | 640 | 28.9 (**42.6**) | 29.2 (17.0) | 19.7 (22.2) | 26.6 (28.1) | 26.8 (25.9) | **31.2** (10.0) | 19.5 (22.2) |
| | 2,560 | 31.4 (**48.8**) | 25.5 (36.5) | 24.6 (36.9) | 30.1 (39.9) | **33.0** (33.9) | 30.8 (28.5) | 24.7 (37.0) |

(i) Reuters, (ii) RCV1-2, (iii) WIPO

RCP のハイパーパラメータ値を設定したところ，RCP による複数ラベルのテストデータに対する分類精度は $N = 80$，320，1280 の場合にそれぞれ 33.9%，49.8%，60.2%，すべてのテストデータに対しては，81.8%，87.4%，89.8%となり，いずれも TSVM より高い結果が得られた．

複数ラベルのテストデータに対する NB/DC と NB/EM-$\lambda$ の分類精度は，表 5 (b) に示すように，Reuters と RCV1-2 では RCP よりも低かった．しかし，WIPO では，NB/DC と NB/EM-$\lambda$ の方が RCP よりも高い場合があった．ただし，WIPO では，すべてのテストデータに対する NB/DC と NB/EM-$\lambda$ の分類精度は RCP よりも大きく下回った．複数ラベルのテストデータに対して NB/EM-$\lambda$ が RCP よりも高い分類精度を与える $N = 160$，640 の場合では，NB/EM-$\lambda$ は 22.8%，25.8%のテストデータに複数のカテゴリラベルを付与したのに対し，RCP ではその比率は 17.2%，20.1%であった．また，$N = 2560$ の場合に，複数のカテゴリラベルが付与されたテストデータの比率は，NB/DC で 36.0%，RCP で 22.5%であった．したがって，NB/DC と NB/EM-$\lambda$ で複数ラベルのテストデータの分類精度が高かったは，多くのテストデータが複数のカテゴリラベルを付与される傾向があったためと考えられる．

以上より，複数ラベルのラベルありデータを用いてハイパーパラメータを調節した場合では，一部の例外を除いて，複数ラベルのテストデータに対する分類精度とすべてのテストデータに対する分類精度との間で，RCP と他手法の優劣に大きな傾向の違いがなかったといえる．表 5 (a) のような結果が得られたのは，ハイパーパラメータの調節にすべてのラベルありデータを用いることによって，RCP では付与すべきカテゴリラベルが 1 つのデータにより適したハイパーパラメータ値が選択されたためと考えられる．複数ラベルのデータが少ない Reuters と WIPO では特にその傾向が顕著にみられた．

#### 5.3.3 計算時間

**表 6** に，WIPO でラベルありデータ数 $N$ を変えて分類器 RCP，JLL/MCL，LL/MER，TSVM，NB/DC を学習したときに要した計算時間を示す．表中の値は，各条件で 10 回の実験を行って得られた計算時間の平均値を表す．本実験では，5.2 節に述べた方法で選択されたハイパーパラメータ値，JLL/MCL と LL/MER の重みパラメータ値，TSVM のコストパラメータ値を与えた状況下でパラメータ学習を行ったときの計算時間を計測した．計測には，Linux コンピュータ（Core 2 DUO 2.66 MHz and 8 GB RAM）を用いた．

**表 6**　WIPO データセットでパラメータ学習に要した平均計算時間（秒）
Table 6　Average processing time (sec.) of parameter estimation for WIPO data set.

| $N$ | RCP | JLL/MCL | LL/MER | TSVM | NB/DC |
|---|---|---|---|---|---|
| 160 | 58.6 | 21.2 | 3.7 | 6,617.1 | 10.0 |
| 640 | 105.2 | 22.0 | 4.6 | 6,033.8 | 11.1 |
| 2,560 | 198.0 | 16.7 | 7.9 | 2,951.3 | 12.0 |

表 6 より，RCP の学習に要する計算時間は，JLL/MCL と LL/MER よりも長かった．この計算時間の差は，ラベルありデータとラベルなしデータの学習への寄与度を与える重みの決定法の違いに主に起因する．JJ/MCL と LL/MER では重みパラメータ $\beta$，$\lambda$ の値を与えたうえでパラメータ学習を行うのに対し，RCP では重みパラメータ $\gamma$ と NB モデルパラメータの学習を交互に繰り返す．この繰返し学習のため，RCP では計算時間を要する．しかし，JLL/MCL と LL/MER の重みパラメータも訓練データから推定すべきパラメータであり，重みパラメータの候補値ごとにパラメータ学習を行ったうえで最適な値を選択する必要がある．5.2 節に示した重みの候補値すべてのために要した計算時間は，$N = 160$，640，2560 の場合でそれぞれ 312.2，426.8，478.4 秒（JLL/MCL），70.6，82.8，133.8 秒（LL/MER）であった．

RCP は NB/DC よりもパラメータ学習に計算時間を要した．RCP と NB/DC ではともに，重みパラメータと NB モデルパラメータの繰返し学習を必要とし，その学習法は類似する．NB/DC では，2 値分類器としてカテゴリごとに学習させるため，同時に学習させるパラメータ数が RCP よりも少ない．それゆえ，RCP よりも少ない繰返し数で重みパラメータと NB モデルパラメータの学習が終了することが多く，パラメータ学習に要する計算時間が短かった．

また，RCP の学習に要する計算時間は TSVM よりも短かかったことも確認した．本実験では，ラベルなしデータのコストパラメータ $C^*$ に大きな値を設定するほど TSVM の学習に時間を要する傾向があった．$N = 2560$ では，小さい $C^*$ の値が選択されることが多かったため，$N = 160$，640 よりも TSVM の計算時間の平均値が短い結果が得られた．

## 6. ま と め

本論文では，識別モデルと生成モデルの統合に基づく半教師あり学習により多重分類器を設計する手法を提案した．提案法は，ラベルありデータで学習させた識別モデルとラベルなしデータで学習させた生成モデルを統合して分類器を与えることを特徴とする．識別モデルと生成モデルにそれぞれ対数線形モデルとナイーブベイズ（NB）モデルを用いて多重テキスト分類器を設計した．実テキストデータからなる 3 つのテストコレクションを用いた実験

により，生成モデルと識別学習のハイブリッドに基づく単一ラベル分類器を各カテゴリの2値分類に利用する方法や，従来の対数線形モデルと NB モデルの半教師あり学習法と比較して，提案法ではより汎化性能の高い多重分類器を得られることを確認した．

今後の課題は，ドメイン適応問題等，異なる情報源からラベルありデータとラベルなしデータが与えられる場合の半教師あり学習法を検討することである．

## 参 考 文 献


1) Boutella, M.R., Luo, J., Shena, X. and Brown, C.M.: Learning multi-label scene classification, *Pattern Recognition*, Vol.37, No.9, pp.1757–1771 (2004).
2) Chen, S.F. and Rosenfeld, R.: A Gaussian prior for smoothing maximum entropy models, Technical report, Carnegie Mellon University (1999).
3) Collobert, R., Sinz, F., Weston, J. and Bottou, L.: Large scale transductive SVMs, *Journal of Machine Learning Research*, Vol.7, pp.1687–1712 (2006).
4) 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).
5) Druck, G., Pal, C., Zhu, X. and McCallum, A.: Semi-supervised classification with hybrid generative/discriminative methods, *Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'07*), pp.280–289 (2007).
6) Elisseeff, A. and Weston, J.: A kernel method for multi-labelled classification, *Advances in Neural Information Processing Systems 14*, pp.681–687, MIT Press, Cambridge, MA (2002).
7) Fall, C.J., Törcsvári, A., Benzineb, K. and Karetka, G.: Automated categorization in the international patent classification, *ACM SIGIR Forum*, Vol.37, No.1, pp.10–25 (2003).
8) Fujino, A., Ueda, N. and Saito, K.: Semi-supervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle, *IEEE Trans. Pattern Analysis and Machine Intelligence* (*TPAMI*), Vol.30, No.3, pp.424–437 (2008).
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) Jelinek, F. and Mercer, R.: Interpolated estimation of Markov source parameters from sparse data, *Pattern Recognition in Practice*, Gelsema, E.S. and Kanal, L.N. (Eds.), pp.381–402, Amsterdam, the Netherlands, North-Holland Publishing Company (1980).
11) Joachims, T.: Transductive inference for text classification using support vector machines, *Proc. 16th International Conference on Machine Learning* (*ICML 1999*), pp.200–209 (1999).
12) Lewis, D., Yang, Y., Rose, T. and Li, F.: RCV1: A New benchmark collection for text categorization research, *Journal of Machine Learning Research*, Vol.5, pp.361–397 (2004).
13) Liang, P. and Jordan, M.I.: An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators, *Proc. 25th International Conference on Machine learning* (*ICML 2008*), pp.584–591 (2008).
14) 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).
15) McCallum, A.K.: Multi-label text classification with a mixture model trained by EM, *AAAI'99 Workshop on Text Learning* (1999).
16) 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).
17) 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).
18) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, New York (1983).
19) Sato, I. and Nakagawa, H.: Knowledge discovery of multiple-topic document using parametric mixture model with Dirichlet prior, *Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'07*), pp.590–598 (2007).
20) Ueda, N. and Nakano, R.: Deterministic annealing EM algorithm, *Neural Networks*, Vol.11, pp.271–282 (1998).
21) Ueda, N. and Saito, K.: Parametric mixture models for multi-topic text, *Advances in Neural Information Processing Systems 15*, pp.737–744, MIT Press, Cambridge, MA (2003).
22) 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），情報処理，Vol.45, pp.184–190 (2004).
23) Zhu, X., Ghahramani, Z. and Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions, *Proc. 20th International Conference on Machine Learning* (*ICML 2003*), pp.912–919 (2003).

## 付　　録

### A.1　カテゴリラベルごとに独立な関数を用いた $Q$ 関数の導出方法

式 (18) から式 (20) を導出する方法を述べる．式 (19) で定義されたカテゴリラベルごとに独立な関数 $s_k(\boldsymbol{x}, y_k; \Theta_k)$ は，$\boldsymbol{y} \in \{1,0\}^K$ に対して $\sum_{\boldsymbol{y}} \prod_{k=1}^{K} s_k(\boldsymbol{x}, y_k; \Theta_k) = \prod_{k=1}^{K} \sum_{y_k=0}^{1} s_k(\boldsymbol{x}, y_k; \Theta_k) = 1$ を満たす．このため，$s_k(\boldsymbol{x}, y_k; \Theta_k)$ を用いて，以下のように $R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x}_m; \hat{W}, \Theta^{(t)}, \boldsymbol{\mu})$ を書き換えることができる．

$$
\begin{aligned}
R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x}_m; \hat{W}, \Theta^{(t)}, \boldsymbol{\mu}) &= \frac{\prod_{k=1}^{K} h_k(\boldsymbol{x}_m, y_k; \hat{\boldsymbol{w}}_k, \Theta_k^{(t)}, \mu_k, \boldsymbol{\gamma})}{\prod_{k=1}^{K} \sum_{{y''}_k=0}^{1} h_k(\boldsymbol{x}_m, {y''}_k; \hat{\boldsymbol{w}}_k, \Theta_k^{(t)}, \mu_k, \boldsymbol{\gamma})} \\
&\quad \times \frac{\prod_{k=1}^{K} \sum_{{y''}_k=0}^{1} h_k(\boldsymbol{x}_m, {y''}_k; \hat{\boldsymbol{w}}_k, \Theta_k^{(t)}, \mu_k, \boldsymbol{\gamma})}{\sum_{\boldsymbol{y}' \neq \boldsymbol{0}} \prod_{k=1}^{K} h_k(\boldsymbol{x}_m, {y'}_k; \hat{\boldsymbol{w}}_k, \Theta_k^{(t)}, \mu_k, \boldsymbol{\gamma})} \\
&= \frac{\prod_{k=1}^{K} s_k(\boldsymbol{x}_m, y_k; \Theta_k^{(t)})}{Z_m} \qquad (29)
\end{aligned}
$$

ただし，$Z_m = 1 - \prod_{k=1}^{K} s_k(\boldsymbol{x}_m, {y'}_k = 0; \Theta_k^{(t)})$ である．

次に，式 (21) で定義した $q_k(\boldsymbol{x}_m, y_k; \Theta^{(t)})$ と式 (29) を用いると，以下の等式を得ることができる．

$$
\begin{aligned}
&\sum_{\boldsymbol{y} \neq \boldsymbol{0}} R_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{x}_m; \hat{W}, \Theta^{(t)}, \boldsymbol{\mu}) \log p(\boldsymbol{x}_m|y_k; \Theta_k^{(t+1)}) \\
&= \sum_{\boldsymbol{y}} \frac{\prod_{k'=1}^{K} s_{k'}(\boldsymbol{x}_m, y_{k'}; \Theta_{k'}^{(t)})}{Z_m} \log p(\boldsymbol{x}_m|y_k; \Theta_k^{(t+1)}) \\
&\quad - \frac{\prod_{k'=1}^{K} s_{k'}(\boldsymbol{x}_m, y_{k'} = 0; \Theta_{k'}^{(t)})}{Z_m} \log p(\boldsymbol{x}_m|y_k = 0; \Theta_k^{(t+1)}) \\
&= \sum_{y_k=0}^{1} \frac{s_k(\boldsymbol{x}_m, y_k; \Theta_k^{(t)})}{Z_m} \log p(\boldsymbol{x}_m|y_k; \Theta_k^{(t+1)}) \\
&\quad - \frac{\prod_{k'=1}^{K} s_{k'}(\boldsymbol{x}_m, y_{k'} = 0; \Theta_{k'}^{(t)})}{Z_m} \log p(\boldsymbol{x}_m|y_k = 0; \Theta_k^{(t+1)}) \\
&= \sum_{y_k=0}^{1} q_k(\boldsymbol{x}_m, y_k; \Theta^{(t)}) \log p(\boldsymbol{x}_m|y_k; \Theta_k^{(t+1)}) \qquad (30)
\end{aligned}
$$

上式を式 (18) に代入することで式 (20) を導出できる．

(平成 20 年 8 月 21 日受付)
(平成 20 年 10 月 11 日再受付)
(平成 20 年 12 月 12 日採録)

**藤野　昭典**（正会員）

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．博士（情報学）．同年 NTT 入社．機械学習，テキスト処理等の研究に従事．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）等受賞．電子情報通信学会，IEEE 各会員．

**上田　修功**（正会員）

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主席研究員，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（1997，2006 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，日本神経回路学会，IEEE 各会員．

**磯崎　秀樹**（正会員）

1983 年東京大学工学部計数工学科卒業．1986 年同工学系大学院修士課程修了．博士（工学）．同年 NTT 入社．1990 年より 1 年間 Stanford 大学客員研究員．論理的推論，自然言語処理，質問応答の研究に従事．現在，NTT コミュニケーション科学基礎研究所主幹研究員，2003 年度情報処理学会論文賞，山下記念研究賞，言語処理学会第 12，13 回年次大会最優秀発表賞，Coling-ACL 2006 AFNLP Meritorious Asian Paper Award 等受賞．電子情報通信学会，人工知能学会，言語処理学会各会員．