2007－NL－181 (5)
2007／9／25

F値最大化学習に基づく文書の多重ラベリング

藤野 昭典 磯崎 秀樹 鈴木 潤
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

概要: テキスト分類問題は，一般的に各文書に複数のラベルを付与する問題 (多重ラベリング問題) として定義される。本稿では，機械学習アプローチに基づく多重ラベリング法に焦点を当て，モデル統合と F_1 値最大化に基づく分類器設計法を提案する．提案法では，まず，データへのラベル付与の可否を判定する2値分類のためのモデルをラベル毎に複数個設計する．次に，これらのモデルを訓練データに対して F_1 値を最大化させるように統合する．3つの実テキストデータセットを用いた実験により，提案法は，1つのデータに多くのラベルが付与され，多数のラベルの組合せが存在するデータセットに対して特に有用であることを確認した．

Multi-labeling of Documents based on F-score Maximization

Akinori FUJINO Hideki ISOZAKI Jun SUZUKI
NTT Communication Science Laboratories, NTT Corporation

Abstract: Text categorization is generally defined as a multi-labeling problem, where multiple category labels are assigned to each text document. We focus on machine learning approaches to multi-labeling problems and present a classifier design method based on model combination and F_1-score maximization. In our formulation, we first design multiple models for binary classification per category label, which determine whether a category label is assigned or not to each data sample. Then, we combine these models to maximize the F_1-score of a training dataset. Using three real text datasets, we confirmed experimentally that our proposed method was useful especially for the datasets where many category labels were assigned to each data sample and which consisted of many combinations of category labels.

1 はじめに

テキスト分類問題は，一般的に各文書に複数のラベルを付与する問題 (多重ラベリング問題) とみなせる．例えば，日本語特許文書は，技術分野を表す国際特許分類 (IPC) 記号や，Fタームと呼ばれる複数の観点から定義された分類記号を複数付与することで分類される [3]．それ故，多重ラベリング問題に対して高い汎化能力をもつ分類器を設計することは重要な課題の1つである．

機械学習アプローチにより多重ラベリング問題に対処する最も単純な方法は，2値分類器を用いることである．この方法では，ラベル間の独立性を仮定して，ラベル毎に文書へのラベル付与の可否を判定する2値分類器を設計する．2値分類器としてロジスティック回帰モデル (LRM) やサポートベクトルマシン (SVM) [5]，ナイーブベイズ [10] などが用いられる．

テキスト分類の精度の評価には，F_1 値がしばしば用いられる．近年，F_1 値を直接最大化するようにSVM [6] や LRM [4] を学習させる手法が提案され，その有効性が確かめられてきた．このため，F_1 値最大化学習により得られる分類器を用いることで，多重ラベリングの精度向上が期待できる．

一方，高い汎化性能を得る手法として，複数のモデルを統合する枠組 [14, 12] が提案されている．これらの枠組では，複数のモデルを個々に設計し，それらを重み付きで統合することにより分類器を構築する．各モデルを個々に用いる場合と比較して，モデル統合により高い汎化性能を持つ分類器を得られることが報告されている．

本稿では，多重ラベリング問題に対して，複数のモデルをF値最大化学習により統合して分類器を設計する手法を提案する．提案法では，まず，ラベル毎に2値分類のためのモデルを複数個設計する．

次に，多重ラベリングの精度の評価にしばしば用いられるマイクロ平均，マクロ平均 F_1 値を最大化するようにそれらのモデルを重み付きで統合する．モデル統合の重みの計算には，文献 [4] で提案された LRM の F_1 値最大化学習アルゴリズムを拡張した手法を用いる．3 つの実データを用いた実験により，提案法が従来の 2 値分類器に基づく方法より有用であることを確認する．

多重ラベリング問題に対処する別の方法として，文書と付与すべきラベルの組合せの関係を直接学習する手法が提案され，2 値分類器に基づく方法よりも高精度であることが報告されている [7]．そこで，本稿では，この手法と提案法を比較した結果も合わせて述べる．

2 LRM の F_1 値最大化学習

本節では，LRM の F_1 値最大化学習 [4] について述べる．LRM は特徴ベクトル $\boldsymbol{x}$ で表されるデータへのラベル付与の可否 $y \in \{1, 0\}$ を判定する 2 値分類器である．ここで，$y^{(n)} = 1$ ($y^{(n)} = 0$) は n 番目の特徴ベクトル $\boldsymbol{x}^{(n)}$ にラベルを付与する (付与しない) ことを意味する．

線形モデルに基づく 2 値分類器の識別関数は一般的に以下のように定義される．

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}_1^t \boldsymbol{x} + \theta_0 \tag{1}$$

ここで，$\boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_1^t)^t$ はモデルパラメータベクトルであり，$\boldsymbol{\theta}_1^t \boldsymbol{x}$ は $\boldsymbol{\theta}_1$ と $\boldsymbol{x}$ の内積を，t はベクトルの転置を表す．$f(\boldsymbol{x}^{(n)}; \boldsymbol{\theta}) \geq 0$ ($f(\boldsymbol{x}^{(n)}; \boldsymbol{\theta}) < 0$) のとき，2 値分類器は $\boldsymbol{x}^{(n)}$ へのラベル付与の可否を $\hat{y}^{(n)} = 1$ ($\hat{y}^{(n)} = 0$) であると判定する．

LRM では，一般的に式 (1) の識別関数をロジスティック関数：

$$g(z) = \frac{1}{1 + \exp(-z)} \tag{2}$$

に代入して得られる分布を，ラベル付与の確率を与えるクラス事後確率分布として $P(y = 1|\boldsymbol{x}; \boldsymbol{\theta}) = g(f(\boldsymbol{x}; \boldsymbol{\theta}))$ のように定義する．ここで，$P(y = 0|\boldsymbol{x}; \boldsymbol{\theta}) = 1 - P(y = 1|\boldsymbol{x}; \boldsymbol{\theta}) = g(-f(\boldsymbol{x}; \boldsymbol{\theta}))$ であり，$g(0) = 0.5$ である．それ故，$P(y = 1|\boldsymbol{x}^{(n)}; \boldsymbol{\theta}) \geq 0.5$ ($P(y = 1|\boldsymbol{x}^{(n)}; \boldsymbol{\theta}) < 0.5$) の場合，LRM は $y^{(n)} = 1$ ($y^{(n)} = 0$) であると判定する．モデルパラメータベクトル $\boldsymbol{\theta}$ の値は，訓練データ集合 $D = \{\boldsymbol{x}^{(m)}, y^{(m)}\}_{m=1}^M$ の $P(y|\boldsymbol{x}; \boldsymbol{\theta})$ に対する尤度と $\boldsymbol{\theta}$ の事前確率密度を表す以下の目的関数の最大化により推定できる．

$$J_R(\boldsymbol{\theta}) = \sum_{m=1}^{M} \log P(y^{(m)}|\boldsymbol{x}^{(m)}; \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{3}$$

本稿では，式 (3) を用いて学習される LRM を便宜上，LRM-L と呼ぶ．

これに対して，F_1 値最大化に基づく LRM の学習 [4] では，ロジスティック関数を用いて訓練データ集合 D の F_1 値の近似関数を与え，その近似関数の最大化により識別関数を学習する．

F_1 値は，適合率 $PR = C/A$ と再現率 $RE = C/B$ を用いて，$F_1 = 2(1/PR + 1/RE)^{-1}$ のように定義される評価値である．ここで，A は分類器がラベルを付与すべきと予測するデータ ($\hat{y}^{(n)} = 1$ であるデータ) の数を，B は真にラベルを付与すべきデータ ($y^{(n)} = 1$ であるデータ) の数を表す．C は $\hat{y}^{(n)} = y^{(n)} = 1$ であるデータの数を表す．これらは，$A = \sum_{m=1}^M \hat{y}^{(m)}$，$B = \sum_{m=1}^M y^{(m)}$，$C = \sum_{m=1}^M y^{(m)} \hat{y}^{(m)}$ により計算できる．

ここで，$\hat{y}^{(m)}$ を以下のように式 (1) と式 (2) で示した識別関数とロジスティック関数を用いて

$$\hat{y}^{(m)} \approx g(\gamma f(\boldsymbol{x}^{(m)}; \boldsymbol{\theta})),\ \gamma > 0 \tag{4}$$

のように近似すると，$\lim_{\gamma \to \infty} g(\gamma f(\boldsymbol{x}^{(m)}; \boldsymbol{\theta})) = \hat{y}^{(m)}$ より，γ が十分大きいとき，式 (4) で与えた近似関数は $\hat{y}^{(m)}$ と一致する．式 (4) の近似を用いることで，訓練データ集合 D に対する F_1 値の近似関数を以下のように与えることができる．

$$\tilde{F}_1(\boldsymbol{\theta}) = \frac{2 \sum_{m=1}^M g(\gamma f(\boldsymbol{x}; \boldsymbol{\theta})) y^{(m)}}{\sum_{m=1}^M y^{(m)} + \sum_{m=1}^M g(\gamma f(\boldsymbol{x}; \boldsymbol{\theta}))} \tag{5}$$

上式を目的関数として勾配法を適用することにより，初期値周辺で $\tilde{F}_1(\boldsymbol{\theta})$ を最大化させるパラメータ $\boldsymbol{\theta}$ の局所最適解を求めることができる．本稿では，この手法で学習される LRM を LRM-F と呼ぶ．

3 提案法

本稿では，多重ラベリング問題に対して，F_1 値最大化学習によるモデル統合に基づく分類器設計法を提案する．本節では，提案法におけるモデル統合の枠組みと LRM と SVM の 2 つのモデルを用いた応用法を示す．

3.1 多重ラベリングのためのモデル統合

多重ラベリング問題は，K 個の候補の中から各データに付与すべきラベルを複数個選択する問題で

あり，特徴ベクトル $\boldsymbol{x}$ に対するラベル付与ベクトル $\boldsymbol{y} = (y_1, \ldots, y_k, \ldots, y_K)^t$ を予測する問題として定義される．ここで，$y_k \in \{1, 0\}$ は，n 番目の特徴ベクトル $\boldsymbol{x}^{(n)}$ にラベルが付与される (付与されない) 場合に $y_k^{(n)} = 1$ ($y_k^{(n)} = 0$) となる変数である．

提案法では，まず，ラベル毎に 2 値分類のためのモデルを J 個 ($J \geq 1$) 設計し，個々にモデルを学習することで $J \times K$ 個の識別関数を得る．ここで，k 番目のラベルのために設計された j 番目のモデルの識別関数を $f_{jk}(\boldsymbol{x}; \boldsymbol{\theta}_{jk})$ で表し，$\boldsymbol{\theta}_{jk}$ をモデルパラメータベクトルとする．また，モデルパラメータ集合を $\Theta = \{\boldsymbol{\theta}_{jk}\}_{j,k}$ で表す．モデルパラメータベクトルを個々に学習して得られる推定値を $\hat{\Theta} = \{\hat{\boldsymbol{\theta}}_{jk}\}_{jk}$ で表す．

次に，提案法では，各モデルの識別関数を以下のように線形に統合することで，分類器の識別関数を新たに定義する．

$$f_k(\boldsymbol{x}; \hat{\Theta}, \boldsymbol{w}) = \sum_{j=1}^{J} w_j f_{jk}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{jk}) + w_0, \ \forall k \qquad (6)$$

ここで，$\boldsymbol{w} = (w_0, w_1, \ldots, w_j, \ldots, w_J)^t$ は重みパラメータであり，w_j ($j \geq 1$) は j 番目のモデルの統合の重みを，w_0 はラベル付与の判定基準を調節する閾値を与える．

$\boldsymbol{w}$ の値は，多重ラベリング問題の評価によく用いられるマイクロ平均 F_1 値を最大化させるように与えられる．訓練データ集合 $D = \{\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)}\}_{m=1}^{M}$ のマイクロ平均 F_1 値は以下の式で計算できる．

$$F_\mu = \frac{2\sum_{m=1}^{M}\sum_{k=1}^{K} y_k^{(m)} \hat{y}_k^{(m)}}{\sum_{m=1}^{M}\sum_{k=1}^{K} y_k^{(m)} + \sum_{m=1}^{M}\sum_{k=1}^{K} \hat{y}_k^{(m)}} \qquad (7)$$

ここで，式 (4) と同様に，$\hat{y}_k^{(m)}$ を

$$\hat{y}_k^{(m)} \approx g(\gamma f_k(\boldsymbol{x}^{(m)}; \hat{\Theta}, \boldsymbol{w})), \ \gamma > 0 \qquad (8)$$

のように近似して式 (7) に代入することで，マイクロ平均 F_1 値の近似関数 $\tilde{F}_\mu(\hat{\Theta}, \boldsymbol{w})$ を与える．$\tilde{F}_\mu(\hat{\Theta}, \boldsymbol{w})$ を最大化させる $\boldsymbol{w}$ の値を $\boldsymbol{w}$ の推定値とする．

しかし，訓練データ集合 D は，各モデルのパラメータ集合 Θ の学習にも用いられる．Θ と $\boldsymbol{w}$ の学習に同一の訓練データを用いると，$\boldsymbol{w}$ の過学習を引き起こす危険性がある．そこで，文献 [14] 等で適用されているように訓練データ集合の n 分割交差検定法を用いて $\boldsymbol{w}$ の推定を行う．$\hat{\Theta}^{(-m)}$ をデータ $\{\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)}\}$ を含まない $n-1$ 個の訓練データのサブ集合を用いて学習したパラメータ推定値として，$\boldsymbol{w}$ のための目的関数を

$$J_\mu(\boldsymbol{w}) = \frac{2\sum_{m,k} y_k^{(m)} g(\gamma f_k(\boldsymbol{x}; \hat{\Theta}^{(-m)}, \boldsymbol{w}))}{\sum_{m,k} y_k^{(m)} + \sum_{m,k} g(\gamma f_k(\boldsymbol{x}; \hat{\Theta}^{(-m)}, \boldsymbol{w}))} + r(\boldsymbol{w}) \qquad (9)$$

で与える．ここで，$r(\boldsymbol{w})$ は分類器の過学習を抑制するためのペナルティ項である．提案法では，以下のガウス分布を用いてペナルティ項を与える．

$$r(\boldsymbol{w}) = \prod_{j=0}^{J} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(w_j - \rho_j)^2}{2\sigma_j^2}\right\} \qquad (10)$$

ここで，σ_j と ρ_j は学習時に定数値を設定すべきハイパーパラメータである．準ニュートン法の一種である L-BFGS アルゴリズム [8] を適用することで，$\boldsymbol{w}$ の初期値周辺で $J_\mu(\boldsymbol{w})$ を最大化させる $\boldsymbol{w}$ の推定値を求める．本稿では，式 (9) による手法を MC-F_μ と呼ぶ．

3.2 マクロ平均 F_1 値，ラベリング F_1 値最大化による学習

多重ラベリング問題では，マイクロ平均 F_1 値の他にマクロ平均 F_1 値やラベリング F_1 値 [13, 7] も分類器の評価に用いられる．マクロ平均 F_1 値がラベル毎にデータ選択の正確性を示す F_1 値を平均したものであるのに対し，ラベリング F_1 値はデータ毎にラベル選択の正確性を示す F_1 値を平均したものである．訓練データ集合 D に対するこれらの F_1 値は以下の式で計算できる．

$$F_M = \frac{1}{K}\sum_{k=1}^{K} \frac{2\sum_{m=1}^{M} y_k^{(m)} \hat{y}_k^{(m)}}{\sum_{m=1}^{M} y_k^{(m)} + \sum_{m=1}^{M} \hat{y}_k^{(m)}} \qquad (11)$$

$$F_L = \frac{1}{M}\sum_{m=1}^{M} \frac{2\sum_{k=1}^{K} y_k^{(m)} \hat{y}_k^{(m)}}{\sum_{k=1}^{K} y_k^{(m)} + \sum_{k=1}^{K} \hat{y}_k^{(m)}} \qquad (12)$$

F_M と F_L の近似関数 $\tilde{F}_M(\hat{\Theta}, \boldsymbol{w})$，$\tilde{F}_L(\hat{\Theta}, \boldsymbol{w})$ を F_μ と同様に式 (8) を用いて与えることで，式 (9) と類似の形で表される目的関数を得ることができる．そこで，次節で述べる評価実験では，$\tilde{F}_M(\hat{\Theta}, \boldsymbol{w})$，$\tilde{F}_L(\hat{\Theta}, \boldsymbol{w})$ を最大化させるように $\boldsymbol{w}$ を学習して得られる分類器の性能についても確認する．本稿では，$\tilde{F}_M(\hat{\Theta}, \boldsymbol{w})$，$\tilde{F}_L(\hat{\Theta}, \boldsymbol{w})$ の最大化に基づきモデルを統合する手法をそれぞれ MC-F_M，MC-F_L と呼ぶ．

3.3 文書の多重ラベリング問題への応用

提案法を文書の多重ラベリング問題に適用するため，本稿では LRM と SVM を 2 値分類のモデルとし

て用いた．文書の特徴ベクトルは，単語の独立性を仮定して単語頻度ベクトル $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_V)^t$ で与えた．ここで，x_i は i 番目の単語の出現頻度を，V はテキストデータ集合に含まれる語彙の総数 (特徴ベクトルの次元) を表す．LRM，SVM 共に線形の識別関数を用いた．

LRM のパラメータベクトルの推定値は，式 (3) と同様の形で表される目的関数の最大化により与えた．但し，パラメータベクトルの事前確率分布にはガウス事前確率分布 [1] を用いた．SVM のパラメータベクトルの推定には SVM^{light}[1] を用いた．

4 実験

4.1 テストコレクション

提案法の評価実験は，Reuters-21578 (Reuters) と WIPO-alpha (WIPO)，日本語特許データ (JPAT) の 3 つのテストコレクションを用いて行った．Reuters と WIPO は英語文書からなるデータセットであり，多重ラベリング問題に対する分類器のベンチマークテストにしばしば用いられてきた．

Reuters は，Reuters newswire のニュース記事を集めたものであり，135 トピックカテゴリからなる．文献 [15] の設定に従い，7770 個と 3019 個の記事をそれぞれ訓練，テストデータとして抽出した．これらのデータはすべて 90 トピックカテゴリのいづれかに属していた．実験では，停止語リスト [11] に含まれる語彙と 1 つの記事のみに含まれる低頻度語彙を取り除いて，各記事の特徴ベクトルを作成した．特徴ベクトルの次元は 16365 であった．

WIPO は，国際特許分類 (IPC) 体系のラベルが付与された特許文書を集めたものである [2]．IPC は，技術分野を表すカテゴリを *Section*, *Class*, *Sub class*, *Main group* の 4 階層と各 Main group に付随する Sub group 分類木により体系化したものであり，各文書には属するカテゴリのラベルが付与されている．評価実験には，*Section D* に属する文書に付与すべき Main group のラベルを 160 個の候補から選択する問題を用いた．データセットの設定に従い，それぞれ 1352，358 文書を訓練，テストデータとして用いた．Reuters と同様の方法で語彙を除去して，各特許文書の特徴ベクトルを作成した．特徴ベクトルの次元は 45895 であった．

JPAT は，NTCIR[2] 特許検索タスクの分類サブタスク [3] で配布されたテストコレクションであり，

[1] http://svmlight.joachims.org/

[2] http://research.nii.ac.jp/ntcir/index-ja.html

表 1: データセットの統計情報：1 つのデータに付与されるラベル数の平均値 N_{av} と最大値 N_{max}，データセット中のラベル数 K，データ数 N_{ds}，ラベルの組合せの総数 N_{LC}

	Reuters	WIPO	JPAT
N_{av}	1.17	1.28	10.5
N_{max}	15	6	40
K	90	160	268
N_{ds}	10789	1710	2464
N_{LC}	468	378	2430
N_{ds}/N_{LC}	23.1	4.52	1.01

1993 年から 1999 年までの間に特許庁に提出された日本語特許文書を集めたものである．これらの文書は上位の「テーマ」と下位の「F ターム」の 2 階層からなる分類体系のラベルが付与されている．評価実験には，テーマ *5J104* に属する文書に付与すべき F タームを 268 個の候補から選択する問題を用いた．1993 年から 1997 年までに提出された 1920 文書を訓練データ，1998 年と 1999 年に提出された 544 文書をテストデータとして用いた．形態素解析器 MeCab[3] を用いて，各文書から名詞，動詞，形容詞を抽出して特徴ベクトルを作成した．1 つの文書のみに含まれる低頻度語彙を取り除いた結果，特徴ベクトルの次元は 21135 であった．

表 1 に，3 つのデータセットのラベル付与に関する統計情報をまとめた結果を示す．Reuters と WIPO では，1 データ当りに付与されるラベル数の平均値 N_{av} は 1 に近く，JPAT と比べて非常に小さかった．データに付与されるラベルの組合せの総数 N_{LC} は，Reuters，WIPO と比較して，JPAT で非常に多かった．これらの結果より，データ毎に異なる多くのラベルが付与されている JPAT は Reuters や WIPO よりも複雑なデータセットであるといえる．

4.2 実験方法

3 節で述べた MC-F_μ，MC-F_M，MC-F_L の提案法を評価するため，2 つの LRM に基づく手法 LRM-L，LRM-F，3 つの SVM に基づく手法 SVM，SVM-J，SVM-F との比較実験を行った．これらの手法でテストデータにラベル付けを行ったときのマイクロ平均，マクロ平均，ラベリング F_1 値を調べた．実験では，特徴ベクトル $\boldsymbol{x}$ を $\sum_{i=1}^{V} x_i = 1$ を満たすように正規化して各手法に適用した．

提案法と SVM，SVM-J の各手法で用いた SVM には線形カーネル関数を適用し，SVM^{light} でペナルティコスト C をデフォルト設定にして学習させた．

[3] http://mecab.sourceforge.net/

SVM-J では，文献 [9] の設定に従って，正例と負例の重みを調節する J パラメータの値を与えて SVM を学習させた．SVM-F では，SVM^{perf}[4]を用いて訓練データの F_1 値の最大化により SVM を学習させた．SVM-F のペナルティコスト C の値には，候補値 $\{10^n\}_{n=1}^{5}$ のうち，テストデータのマイクロ平均 F_1 値を最大化させる値を用いた．提案法のペナルティ項のハイパーパラメータについては，モデルの重み w_j $(j \neq 0)$ に関する σ_j と ρ_j を 1，閾値 w_0 に関する ρ_0 を 0 に固定し，σ_0 の値を候補値 $\{10^n\}_{n=-3}^{5}$ の中から交差検定法を用いて選択した．

また，特徴ベクトルからラベル付与ベクトルへの写像を直接学習する最大マージンラベリング法 (MML) [7] の性能も調べた．MML のハイパーパラメータ値には，候補値 $\{10^n\}_{n=-1}^{2}$ の中からテストデータのマイクロ平均 F_1 値を最大化させる値を用いた．

4.3 実験結果と考察

4.3.1 従来法との比較

表 2 に，提案法と 4.2 節で述べた比較手法を 3 つのデータセットに適用して得られたラベリングの精度を示す．実験では，各手法でテストデータにラベル付けしたときのマイクロ平均 F_1 値・適合率・再現率 (F_μ, P_μ, R_μ)，マクロ平均 F_1 値・適合率・再現率 (F_M, P_M, R_M)，ラベリング F_1 値・適合率・再現率 (F_L, P_L, R_L) の 9 つの評価値を調べた．F_M と P_M を算出する際，すべてのテストデータに付与されなかったラベルの F_1 値と適合率を 0 とした．同様に，F_L と P_L を算出する際，ラベルが全く付与されなかったデータの F_1 値と適合率を 0 とした．表 2 に SVM-F の実験結果の記載がないのは，SVM-F の計算コストが高く，JPAT では最適なペナルティコスト値を得られなかったためである．

表 2 より，すべてのデータセットで LRM-F の F_M 値は LRM-M より大きかった．SVM-F は Reuters と WIPO で SVM より大きい F_M 値を与えた．これらの結果は，文献 [6, 4] の実験結果と一致する．また，LRM-F (SVM-F) の F_L 値は LRM-L (SVM and SVM-J) より大きく，SVM-F の F_μ 値は SVM, SVM-J よりも大きかった．Reuters 以外のデータセットでは，LRM-F も LRM-L より大きい F_μ 値を与えた．以上の結果より，ラベル毎に F_1 値を最大化するように 2 値分類器 LRM，SVM を学習させることは，マクロ平均 F_1 値のみならず，マイクロ平均，ラベリング F_1 値を向上させる効果があったといえる．

[4] http://svmlight.joachims.org/svm_perf.html

表 2: 提案法と比較手法によるマイクロ平均，マクロ平均，ラベリング F_1 値 (%)

(a) Reuters

Method	F_μ (P_μ/R_μ)	F_M (P_M/R_M)	F_L (P_L/R_M)
MC-F_μ	87.1 (87.9/86.4)	47.5 (57.5/44.4)	89.7 (89.8/92.1)
MC-F_M	83.6 (77.6/90.7)	**51.5** (52.7/55.5)	88.8 (86.7/94.9)
MC-F_L	86.6 (85.4/87.8)	49.5 (57.5/47.7)	89.9 (89.5/93.2)
LRM-L	86.3 (90.5/82.4)	45.2 (57.1/40.7)	87.4 (88.3/88.5)
LRM-F	85.5 (86.9/84.2)	50.8 (56.0/52.8)	87.5 (87.8/89.7)
SVM	84.4 (95.1/75.9)	35.3 (51.4/28.7)	84.2 (85.9/84.1)
SVM-J	22.4 (12.7/93.7)	28.6 (22.8/77.8)	44.1 (38.0/96.0)
SVM-F	86.5 (91.5/82.0)	45.6 (57.8/40.3)	87.5 (88.5/88.4)
MML	**87.8** (92.6/83.4)	49.5 (62.6/43.8)	**91.2** (93.8/90.8)

(b) WIPO

Method	F_μ (P_μ/R_μ)	F_M (P_M/R_M)	F_L (P_L/R_M)
MC-F_μ	**51.4** (57.8/46.2)	30.7 (36.3/30.2)	46.6 (48.3/51.0)
MC-F_M	50.7 (50.1/51.3)	**31.3** (33.1/33.9)	48.1 (48.1/56.3)
MC-F_L	49.3 (46.9/52.0)	**31.3** (33.0/35.6)	47.0 (46.2/57.0)
LRM-L	40.9 (70.2/28.9)	22.4 (34.1/18.1)	33.0 (36.7/32.5)
LRM-F	43.3 (69.0/31.5)	25.4 (34.9/21.7)	35.4 (39.1/35.3)
SVM	22.0 (84.8/12.6)	10.7 (21.0/8.2)	15.7 (17.7/14.9)
SVM-J	21.4 (12.8/66.8)	24.9 (20.1/49.2)	21.1 (13.0/69.4)
SVM-F	40.7 (70.1/28.7)	22.8 (35.3/18.7)	33.2 (36.5/32.6)
MML	48.6 (54.9/43.6)	30.8 (36.5/29.7)	**49.4** (56.2/48.4)

(c) JPAT

Method	F_μ (P_μ/R_μ)	F_M (P_M/R_M)	F_L (P_L/R_M)
MC-F_μ	41.9 (43.5/40.4)	17.9 (21.5/18.0)	40.5 (44.7/43.9)
MC-F_M	40.8 (35.7/47.6)	**20.2** (20.8/23.0)	39.5 (37.5/51.3)
MC-F_L	**42.0** (45.3/39.1)	17.5 (21.7/17.2)	**40.7** (46.3/42.6)
LRM-L	33.8 (44.3/27.4)	15.2 (22.7/12.6)	32.2 (46.6/29.8)
LRM-F	38.2 (44.5/33.4)	17.7 (23.3/16.9)	36.3 (46.6/36.8)
SVM	25.3 (75.1/15.2)	5.3 (12.8/3.8)	25.5 (65.5/17.7)
SVM-J	24.6 (15.2/65.3)	19.5 (13.9/56.3)	24.3 (15.9/67.7)
MML	32.7 (42.1/26.8)	14.7 (19.4/12.9)	32.2 (51.8/30.5)

次に，提案法と LRM-F，SVM-F との比較結果を考察する．MC-F_μ は LRM-F，SVM-F より大きな F_μ 値を与え，MC-F_M の F_M 値は LRM-F，SVM-F より大きかった．さらに，MC-F_L の F_L 値は LRM-F，SVM-F より大きかった．これらの結果は，マイクロ平均，マクロ平均，ラベリング F_1 値の最大化に基づくモデル統合が，テストデータに対するそれぞれの F_1 値を向上させるのに効果があったことを示している．提案法によるモデル統合が LRM-F，SVM-F より高精度な多重ラベリングを行うのに有用であったことを確認した．

提案法 MC-F_μ，MC-F_M，MC-F_L は，JPAT では，MML より大きな F_μ 値，F_M 値，F_L 値を与えた．しかし，MML は，Reuters と WIPO では MC-F_L よりも大きな F_L 値を与え，Reuters では MC-F_μ よりも大きな F_μ 値を与えた．MML は特徴ベクトルからラベルの組合せへの写像を直接学習するモ

表 3: MC-F_μ, LRM-F_μ, SVM-F_μ のマイクロ平均 F_1 値 (%)

	Reuters	WIPO	JPAT
MC-F_μ	**87.1** (87.9/86.4)	**51.4** (57.8/46.2)	**41.9** (43.5/40.4)
LRM-F_μ	85.4 (84.4/86.4)	49.0 (59.8/41.5)	39.9 (41.9/38.1)
SVM-F_μ	86.9 (88.3/85.6)	48.4 (55.2/43.0)	41.8 (44.9/39.1)

デルである．しかし，表 1 より，JPAT では，ラベルの組合せの総数が Reuters，WIPO と比較して非常に多く，同じ組合せのラベルが付与されるデータの個数が非常に少なかった．それ故，JPAT では，MML の過学習が生じたと考えられる．逆に，提案法では，ラベル毎に学習させる 2 値分類器を用いることで，このような過学習の問題を抑制する．提案法は，JPAT のように多くのラベルの組合せが存在する複雑なデータセットに対して有用であるといえる．

4.3.2 モデル統合の効果

複数のモデルを統合する効果を調べるため，式 (6) で示した識別関数に LRM のみを適用して構築した分類器 LRM-F_μ の性能を調べた．同様に，SVM のみを適用して構築した分類器 SVM-F_μ の性能も調べた．表 3 に，MC-F_μ と LRM-F_μ，SVM-F_μ により得られた F_μ 値を示す．

MC-F_μ の F_μ 値は，すべてのデータセットで LMR-F_μ，SVM-F_μ より大きかった．LMR-F_μ と SVM-F_μ の F_μ 値の差が小さい WIPO で，MC-F_μ は両手法より明らかに大きな F_μ 値を与えた．提案法によるモデルの統合は，用いるモデルの性能が近い場合に特に有用であることを確認した．

5 まとめ

本稿では，多重ラベリング問題に対して，モデル統合に基づく分類器設計法を提案した．提案法は，マイクロ平均，マクロ平均，ラベリング F_1 値のような評価値を最大化するように重みを与えて複数のモデルを統合することを特徴とする．3 つの実テキストデータセットを用いた実験により，提案法は 2 値分類に基づく従来手法より大きな評価値を与え，多重ラベリングの高精度化を実現することを確認した．また，提案法は，多くのラベルの組合せが存在する複雑なデータセットに特に有用であることを確認した．今後の課題は，より高精度な多重ラベリングを実現するため，付与されるラベルが既知のラベルありデータと同時に，付与されるラベルが未知のラベルなしデータを学習に用いる半教師あり学習の手法を検討することである．

参考文献

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

[2] Fall, C. J., Törcsvári, A., Benzineb, K., and Karetka, G. Automated categorization in the international patent classification. *ACM SIGIR Forum*, **37 (1)**, 10–25, 2003.

[3] Iwayama, M., Fujii, A., and Kando, N. Overview of classification subtask at NTCIR-6 patent retrieval task. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-6)*, 366–372, 2007.

[4] Jansche, M. Maximum expected F-measure training of logistic regression models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP2005)*, 692–699, 2005.

[5] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML '98)*, 137–142, 1998.

[6] Joachims, T. A support vector method for multivariate performance measures. *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, 377–384, 2005.

[7] Kazawa, H., Izumitani, T., Taira, H., and Maeda, E. Maximal margin labeling for multi-topic text categorization. In *Advances in Neural Information Processing Systems 17* (pp. 649–656). Cambridge, MA: MIT Press, 2005.

[8] Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Programming*, **45 (3, (ser. B))**, 503–528, 1989.

[9] Morik, K., Brockhausen, P., and Joachims, T. Combining statistical learning with knowledge-based approach. In *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, 268–277, 1999.

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

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

[12] Ting, K. M. and Witten, I. H. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, **10**, 271–289, 1999.

[13] Ueda, N. and Saito, K. Parametric mixture models for multi-topic text. In *Advances in Neural Information Processing Systems 15* (pp. 737–744). Cambridge, MA: MIT Press, 2003.

[14] Wolpert, D. H. Stacked generalization. *Neural Networks*, **5 (2)**, 241–259, 1992.

[15] 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)*, 42–49, 1999.