東海大学紀要情報通信学部
Vol.10,No.1,2017,pp.15-24

論文

トピックモデルおよび最小平均自乗推定を用いたテキスト文書のマルチラベル分類法について

藤野　巖[*1]，池田　悟[*2]，山本　宙[*3]

An Approach to Multi-label Classification of Text Documents by Combining Topic Model and Least Mean Square Estimation

by

Iwao FUJINO[*1], Satoru IKEDA[*2] and Hiroshi YAMAMOTO[*3]

(received on Mar. 31, 2017　&　accepted on Jul. 13, 2017)

あらまし

本論文は，テキスト文書のマルチラベル分類問題について，トピックモデルと最小平均自乗推定の2段階からなる実現手法について提案する．第1段階では，トピックモデルを用いて各文書のトピック分布を求める．第2段階では，最小平均自乗推定に基づき，文書のトピック分布から文書のラベルを推定する．本研究では，理想的な条件のもとでトピックモデルのトピックの間が独立であるという点に着目し，文書のラベルをトピック空間のベクトルであると仮定している．その上で最小平均自乗推定に基づき，トピックからラベル出現確率に変換する重み行列を算出し分類器を構築する．そして，与えられた文書について，この分類器を用いて，ラベルの出現確率を計算し，所定ルールに従ってラベルを決定し当該文書に付与する．提案手法を検証するために，朝日新聞の分類付きニュース記事データを用いてマルチラベル自動分類実験を行った．検証実験の結果により従来の手法と比べて提案手法の優位性が確認された．

Abstract

This paper presents a two-stage approach for multiple labels classification of text documents. The first stage of the approach applies topic model to obtain a topic proportion, while the second stage applies least mean square estimation to obtain the labels. We suppose that a topic should be independent from each other under ideal conditions and this leads us to express the label probability by a linear transform from topic proportion. Therefore we can calculate the optimum solution under the least mean square criterion. Furthermore we introduced a series decision rule for deciding the most appropriate labels from the LMS estimation results. In order to evaluate the proposed methods, we conducted a confirmation experiment using labeled news data from Asahi Shimbun. The results of our experiment show that we achieved remarkable improvements comparing with conventional method.

キーワード: マルチラベル分類, 最小平均自乗推定, トピックモデル, 新聞記事の自動分類
Keywords: *Multi-label Classification, Least Mean Square Estimation, Topic Model, Automatic Classification of News Articles*

1. はじめに

近年の高度情報化社会の発展に伴い，過去の時代では考えられないほど膨大な情報がインターネットを経由して流通している．これらの情報の多くは個人のブログやソーシャルネットワークサービス（SNS）からも発信されている文書情報で，最初からきちんと分類はされていないものがほとんどである．ユーザーに良質を情報サービス提供するために，このような一次的な情報について分類やタグ付けなどを行って整理加工する必要がある．もちろん，情報の量から見ると，人手による作業は現実的に不可能で，コンピュータシステムによる自動作業が求められている．

大量なデータを自動的に分類するには，機械学習分野の分類（classification）に基づく手法がしばしば用いられる[1)]．この手法では，与えられたラベル付き訓練データから自動的に学習して，未知の文書に付与するラベルを決めるように行っている．ここで言うラベルに関して，従来的には互いに独立したもので，それぞれの文書には訓練データから集めたラベル集合の要素の一つが付与される．つまり，従来の手法は一つ文書に一つラベルしか付与されないものである．しかし，実際には，情報の多面性から，一つ文書に対して，単一ラベルで十分なものが少ない，むしろ複数ラベルを付けるべきものが多い．例えば，「国会で政府提案の景気対策を議論」のようなニュースは政治ラベルと経

*1 情報通信学部通信ネットワーク工学科　教授
School of Information and Telecommunication Engineering, Department of Communication and Network Engineering, Professor
*2 情報通信学研究科情報通信学専攻　修士課程
Graduate School of Information and Telecommunication Engineering, Course of Information and Communication Engineering, Master's Program
*3 情報通信学部通信ネットワーク工学科　准教授
School of Information and Telecommunication Engineering, Department of Communication and Network Engineering, Associate Professor

済ラベルの両方が付けられる.「芸能人の○○が××マラソン大会で転倒」のようなツイートについて,芸能ラベル,スポーツラベル,速報ラベルが付けられる.またタグとしては,＃○○や＃××マラソンが考えられる.また,特別なケースとして,文書に特徴がないためラベルが付与されない場合もある.このように,文書データの整理加工をするために,ラベル数を固定しないマルチラベル分類器の実現が求められている.

マルチラベル分類問題について,近年多くの研究者から注目と関心が寄せられ,活発に研究されている.マルチラベル分類問題に直接的に取り組むアルゴリズムはいくつか提案されている.1999 年に McCallum らによってマルチラベルテキスト分類に取り組むために,Mixture Model[2]が提案された.その発展として,2003 年に Ueda らによる Paramatric Mixture Model が提案された[3].2000 年に Scapire と Singer によって提案された BoosTexter と呼ばれる手法[4],および 2001 年に Elisseeff と Weston によって導入された SVM に基づくランキングアルゴリズム[5]がある.さらにシングルラベル分類問題の k 近傍法（kNN: k Nearest Neighbors）の拡張として 2006 年に Zhang らによって MLKNN[6]が提案された.その後,2011 年に Wei らによって NaiveBayes 法を拡張したアルゴリズム[7]が提案された.2012 年に Chiang らによってランキングベースの kNN アルゴリズム[8]が提案された.早期の研究では,比較的小規模の非テキストデータセットを処理の対象として用いられているため,対応するラベルの数は比較的に少ない.確率的生成モデルを用いたアプローチでは,EM アルゴリズムに基づく計算手法がいくつか提案されたが,直接的に推定されるべきパラメータが多いため,十分な精度に達するには困難である.また,kNN 法のような怠惰学習（lazy learning）に属する手法では,すべての計算がインスタンスを分類するまでに行われないものである.訓練用事例が多くなると,一つインスタンスが与えられたら,毎回すべての訓練事例との距離を計算するので,大変計算時間の掛かる手法で,テキスト文書の分類には不向きである.

本研究では,トピックモデルと最小平均自乗推定との結合により,マルチラベル分類の手法を提案する.この手法では,まず大量な文書をまとめて,トピックモデルの LDA(Latent Dirichlet Allocation)法によって準備処理を行う.この準備処理によって,各文書のトピック分布が得られる.そして,このトピック分布から,訓練データのラベルセットにより,最小平均自乗推定を行い,分類器の構築を行う.従来の各手法と比較して,提案手法は以下のような優位性があると考える.第 1 には,提案手法では訓練データから学習して事前準備することができるようになるため,MLKNN のような怠惰学習手法と比べて,個別インスタンスのラベルの算出は非常に高速にできる.第 2 には,ベイズの定理に基づく手法では,ラベル間の独立であるとの仮定が必要である.提案手法では最小平均自乗推定を用いることによって,この仮定を回避することができるようになり,新聞記事のトピックや Twitter のタグのような現実社会で使われているものをラベルとして用いることができる.第 3 には,トピックモデルを事前処理に用いることで,テキスト文書にある潜在的な重要な要素（トピック）を自動的に抽出することができると同時に文書を表すデータを単語の次元からトピックの次元に大幅に低次元することができる.これによって,インスタンスのマルチラベルラベル推定時の計算時間の低減と計算精度の向上に貢献できる.

本論文は以下のように構成される.第 2 章では,マルチラベル分類,トピックモデルを用いたマルチラベル分類と新聞記事の自動分類の多方面から既存研究を確認する.第 3 章では,本研究の基本的な考えと処理手順を示し,各処理段階の実現手法について説明する.第 4 章では,第 3 章で説明した提案手法の検証実験を行い,実験の結果を報告するとともに,その考察を行う.最後に,第 5 章では,本研究で得られた結論をまとめ,近いうちに行うべき研究作業を示す.

2. 関連研究

2.1 マルチラベル分類に関する研究

まずは,Zhang らによって提案された MLKNN 法[6]について紹介する.この手法はラベル別に学習により 2 分類の分類器を構築するものである.従来シングルラベル分類問題の k 近傍法（kNN: k Nearest Neighbors）をマルチラベル分類問題に拡張したものである.その処理手順は以下のようなものである.未知のインスタンスに対して,まずは,訓練データセットからそのインスタンスの k 個の近傍インスタンスを見つけ出す.そして,これらの近傍インスタンスから各クラスに属するインスタンスの数をもとに,条件確率を算出する.さらにベイズの定理により各クラスに関する事後確率を算出し,最大事後確率（MAP）の法則に基づき,そのインスタンスのラベルを決定する.この論文では,3 つのマルチラベル分類問題データセット,すなわち Yeast データセット,Scene データセットおよび Yahoo データセットに対して,検証実験を行い,提案手法の方がいくつか定評のある手法よりも優れた性能があることを確認した.前述のように,この手法は怠惰学習に属するもので,計算のために時間が掛かるので,主に比較的小規模データセットに用いられる.

2.2 トピックモデルを用いたマルチラベル分類に関する研究

トピックモデルでは,文書は複数のトピックからそれぞれ一定の割合から合成されるものとして考える.トピックモデルを用いた文書をマルチラベルに分類する研究として,山本らによる「実生活Tweet[+1]に対する局面の階層的推定法」[9]が報告されている.この研究では,第1段階で大量のTweet[+1]からLDAを用いてトピックを抽出した.第2段階で手作業によってラベル

+1 ここの Tweet は Twitter における投稿を表す言葉である.

付けされた少量のラベルつきTweetを訓練データとして用い，Labeled-LDA法によってトピックと局面の関連度を算出する．局面毎に設定した閾値を越えた関連度を持つトピックと局面の対応関係を構築する．そして，第3段階で入力された未知のTweetから単語を抽出して，第1段階で得られたトピック中の各単語の生起確率と第2段階であられたトピックと局面の関連度を用いて，局面ごとにスコアを算出する．そして所定の閾値を超えたスコアを持つ局面をTweetに対して付与することによって，マルチラベル分類を実現している．

2.3 新聞記事の自動分類に関する研究

新聞記事の自動分類に関する早期の研究としては，森本らによる「新聞記事自動分類システムの構築の検討と評価」[10]の報告がある．この研究では，1990 年代日立製作所において開発されたテキスト分類支援ツール FLUTE を新聞記事に適用し，新聞記事自動分類システムの構築を検討した．検証実験として，新聞記事 1 年分を用いて，22 の大分類カテゴリに自動分類する実験を行った．また，村上らによる「新聞記事を対象にした，検索，分類，複数文書要約システム ELIOT システム」[11]の報告がある．この研究では，文書を単語の TF･IDF のベクトル空間のベクトルとして表し，文書の間の類似度に基づきクラスタリングを行ったものである．オンライン新聞記事検索システムに投入されたキーワードに関する検索結果の記事を複数のグループに分けて表示して，グループ内の複数の文書から一つ要約を作成する機能を実現している．

3. 提案手法

3.1 基本的な考え方

kNN に基づく処理手法は，怠惰学習に属する手法で，ラベルを付与したいインスタンスが到着してから，初めて訓練データセットの各インスタンスとの間の距離が計算される．大きな訓練用データセットになると大変時間が掛かりものである．このような考察に基づき，我々はテキスト文書のマルチラベル分類問題の実現手法として 2 段階構成法を提案する．第 1 段階でトピックモデルを用いて，文書のトピック分布を求める．第 2 段階では，文書のトピック分布からその文書に付与するラベルを推測する．以下の提案手法の処理手順について説明する．

（1）予備処理の段階として，教師なし学習のトピックモデルを導入する．これにより，各文書に関するトピックの分布が得られることのほかに，トピックモデルによって，従来の特徴量のベクトル空間を，単語空間（数万語）からトピック空間（数十から百）となるので，大きく次元を削減することができる．

（2）理想条件の下では，トピックモデルの各トピックの間では独立となるから，各トピックを軸とする線形空間が形成できると考える．そして，文書のラベルをこのトピック空間にあるベクトルとして考えることができる．このような理解に基づき，我々はラベルをトピックの線形結合として表すことができると仮定する．その上で，最小平均自乗推定を用いて，この線形結合の重み行列の最適解を求めることができる．

（3）さらに，(2)で得られた重み行列を用いて，与えられた文書のラベルの出現確率を計算し，最後に一定のルールに従って，当該文書に付与するラベルを決定する．ここでいうルールでは，訓練データセットにある文書に付随するラベル数の平均値と標準偏差，および当該文書の各ラベルの出現確率の最大値，平均値と標準偏差のような統計情報を用いて表している．

3.2 ラベル確率ベクトルの最適解

トピックモデルはテキスト文書の確率的生成モデルである．最も利用されるトピックもモデルの実現法は Blei 氏によって提案された LDA 法 [12]である．LDA 法のグラフィカルモデル表現は Fig.1 に示してある．

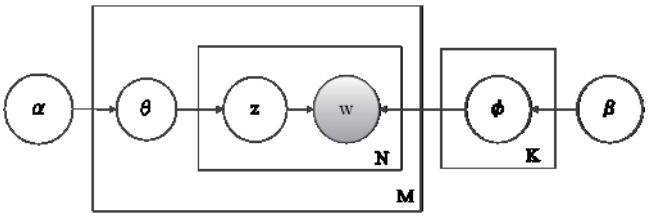

Fig. 1　Graphical model representation of LDA

ここでは，$\boldsymbol{\theta}$は文書のトピック分布を表すもので，φはトピックの単語分布である．αは$\boldsymbol{\theta}$を生成するためのハイパーパラメータで，βは$\boldsymbol{\varphi}$を生成するためのハイパーパラメータである．また，z はトピックマトリクスで，w はトピックごと単語ごとの単語数である．LDA 法を処理した結果の中からトピックの結果より，トピック分布を以下のようにベクトル形式表される．

$$\boldsymbol{\theta}_d = \begin{pmatrix} \theta_{d1} \\ \theta_{d2} \\ \vdots \\ \theta_{dK} \end{pmatrix} \ (d = 1,2,\dots,M) \tag{1}$$

ここで M は文書数を表し，K はトピック数を表す．前述のトピック分布からラベル確率の間の関係を下式のように表される．

$$\hat{\boldsymbol{L}}_d = \boldsymbol{W}\boldsymbol{\theta}_d \ (d = 1,2,\dots,M) \tag{2}$$

ただし，ラベル確率ベクトル $\hat{\boldsymbol{L}}_{\boldsymbol{d}}$ と重み行列$\boldsymbol{W}$は以下に示したものである．

$$\hat{\boldsymbol{L}}_{\boldsymbol{d}} = \begin{pmatrix} p_{d1} \\ p_{d2} \\ \vdots \\ p_{dL} \end{pmatrix} \ (d = 1,2,\dots,M) \tag{3}$$

$$\boldsymbol{W} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{L1} & w_{L2} & \cdots & w_{LK} \end{pmatrix} \tag{4}$$

ここで L はラベルの数である．この重み行列を決定す

るために，最小平均自乗の基準に従った，以下のように評価関数を定義する．

$$J = \mathrm{E}\left[\sum_{n=1}^{L} e_n^2\right] = E[\boldsymbol{e}^T\boldsymbol{e}] \tag{5}$$

ここで $\boldsymbol{e} = \boldsymbol{L}_d - \hat{\boldsymbol{L}}_d$ はラベル確率の正解値ベクトル$\boldsymbol{L}_d$と推定値ベクトル$\hat{\boldsymbol{L}}_d$との差である．この式に対して行列の代数演算を行い，以下のように変形することができる．

$$J = E[\boldsymbol{L}_d^T\boldsymbol{L}_d] - 2E[\boldsymbol{L}_d^T\boldsymbol{W}\boldsymbol{\theta}_d] + E[\boldsymbol{\theta}_d^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\theta}_d] \tag{6}$$

さらに，この式を重み行列$\boldsymbol{W}$に対して微分すれば，以下のような式が得られる．

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{W}} &= -2E[\boldsymbol{L}_d\boldsymbol{\theta}_d^T] + 2\boldsymbol{W}E[\boldsymbol{\theta}_d\boldsymbol{\theta}_d^T] \\ &= -2R_{L\theta} + 2\boldsymbol{W}R_{\theta\theta}\end{aligned} \tag{7}$$

そして$\frac{\partial J}{\partial \boldsymbol{W}} = 0$とおいて重み行列$\boldsymbol{W}$について解くと，下記のように重み行列の最適値$\boldsymbol{W}^*$が得られる．

$$\boldsymbol{W}^* = R_{L\theta}R_{\theta\theta}^{-1} \tag{8}$$

最後に，この最適な重み行列は時間によって変化しないと仮定するもとで，与えられた文書のトピック分布から以下のように当該文書のラベル確率ベクトルを求めることができる．

$$\hat{\boldsymbol{L}}_t = R_{L\theta}R_{\theta\theta}^{-1}\boldsymbol{\theta}_t \tag{9}$$

評価関数の微分計算時に行要素の合計値=1 のような確率にするための制約条件は付けていないが，(9)の計算のあとに，$\hat{\boldsymbol{L}}_t$の各行について，各要素をその行の合計値で割り算することで，行要素の合計値が 1 となるように処理行う．

3.3 ラベル決定のルール

2 分類の場合は，ラベルlの有無の決定は以下のルールに従って行われる．

$$y_l = \begin{cases} True & p(l=1|\boldsymbol{D}) > p(l=0|\boldsymbol{D}) \\ False & otherwise \end{cases} \tag{10}$$

このルールを用いると，すべてのラベルの決定結果はゼロとなる場合がある．しかしながら，実際問題ではよく最低でも一つラベルが付与される前提がよくある．このような問題に取り組むために，ラベル確率ベクトルの平均値を閾値とし，その閾値を超えたラベルが付与されるようなルールが報告されている [7]．我々はこのルールを拡張してより正確なルールを提案する．推定ラベル確率ベクトル$\hat{\boldsymbol{L}}_t$からその要素について降順並べ替えを行い，以下のように表せる．

$$p_{l_1} > p_{l_2} > \cdots > p_{l_f} > \cdots > p_{l_L}$$

もし以下の条件が成り立つならば，1 番目ラベルl_1からを f 番目ラベルl_fを当該文書の付与すべきラベルとする．

(1) 訓練セットを用いてラベル数の平均値μ_tと標準偏差σ_tを算出する．この時，テストインスタンスの推定ラベルの最大番号fは下式に基づいて決定される．

$$f \le int(\mu_t + 3\sigma_t) \tag{11}$$

(2) 当該文書のラベル確率$\hat{\boldsymbol{L}}_t$の要素の最大値をρ とし，平均値をμとし，標準偏差をσとする．この時，各ラベルについて下式に基づきラベルが決定される．

$$y_{l_i} = \begin{cases} True & p_{l_i} > \max(\rho - \lambda\mu_t\sigma, \mu) \\ False & otherwise \end{cases} \tag{12}$$

上式においてλは正解ラベルの範囲を調整するためのパラメータである．また，ここに訓練セットから得られたラベル数の平均値μ_tを導入することによって，訓練セットから得られたラベル数の平均値の情報を用いることができ，自動的に個別インスタンスのラベル確率の範囲を調整することができるようになる．

4. 検証実験

本研究で提案した手法の有効性を検証するために，新聞記事を用いた検証実験を行った．本章では，検証実験に用いた各種評価指標，検証実験に用いたデータの概要，システムのセットアップおよび実験結果を示し，その結果について考察する．

4.1 データセットのマルチラベル度合いを示す指標

以下の記述において，データセットを$D(x_i, Y(x_i))$と表すものとする．ここでは，x_i は特徴量を表す，$Y(x_i)$は特徴量x_iに対応する正解ラベル集合を表す．さらに T は文書数，L はラベル数を表す．データセットのマルチラベル度合いを示す指標として，以下のように LC と LD を定義する．

(1) LC(Label Cardinality): LCは文書あたりのラベル数を表す評価指標である．その定義式を以下に示す [13]．

$$LC = \frac{1}{T}\sum_{i=1}^{T}|Y(x_i)| \tag{13}$$

(2) LD(Label Density): LDは文書あたりの正解ラベル数のラベル数に対する割合を表す評価指標である．その定義式を以下に示す[13]．

$$LD = \frac{1}{T}\sum_{i=1}^{T}\frac{|Y(x_i)|}{L} \tag{14}$$

4.2 分類器の性能評価指標

(1) 分類器の結果はラベルの集合で表される場合，特徴量x_iに対して，その正解ラベル集合が$Y(x_i)$で与えられ，分類器の推定結果ラベル集合が$Z(x_i)$与えられるものとする．その分類器の性能を評価する

指標として，精度(Precision)[13]，再現率(Recall)[13]と F 値(F-measure)を以下のように定義する．

$$精度 = \frac{1}{T}\sum_{i=1}^{T}\frac{|Y(x_i) \cap Z(x_i)|}{|Z(x_i)|} \quad (15)$$

$$再現率 = \frac{1}{T}\sum_{i=1}^{T}\frac{|Y(x_i) \cap Z(x_i)|}{|Y(x_i)|} \quad (16)$$

$$F = \frac{2 \times 精度 \times 再現率}{精度 + 再現率} \quad (17)$$

ここで精度を計算するために，最低一つ推定ラベルが必要で，再現率を計算するために，最低一つ正解ラベルが必要である．つまり，F値を評価指標として用いる場合は，最低でも一つの推定ラベルと一つの正解ラベルが必要である．

(2) 分類器の結果はブール値のラベルベクトルで表される場合，特徴量x_iに対して，その正解ラベルベクトルが$y(x_i)$で与えられ，分類器の推定結果ラベルベクトル$z(x_i)$で与えられるものとする．その分類器hの性能を評価する指標として，文書あたりの不正解ラベル数のラベル総数に対する割合をHammingLossと呼ぶ．その定義式を以下に示す[13]．

$$HammingLoss(h) = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{L}|y(x_i)\Delta z(x_i)| \quad (18)$$

(3) 分類器の結果をあるスコアのランキングに基づき決定される場合はランキングベース分類器という．この分類器を評価するために，OneError指標を用いる．OneError指標は文書あたりの分類器の推定結果のトップにあるラベルが正解ラベル集合に存在しない回数である．その定義式を以下に示す[13]．

$$OneError(f) = \frac{1}{T}\sum_{i=1}^{T} I\left(argmax_{l \ni L} f(x_i, l) \notin Y(x_i)\right) \quad (19)$$

4.3 データの概要

本研究では，「朝日新聞記事データ（学術・研究用）2013 年」を用いて検証実験を行った．この新聞記事データは，新聞記事の見出しや本文だけでなく，Table4 に示すように，「掲載年月日」，「刊種」，「紙誌」，「面名」，「本紙・地方面」，「記事分類」，「文字数」，「見出し」と「本文」の 9 項目からなるフォーマットで記事のデータを記録している．つまり記事分類データは本文ごとに付与されており，これにより研究で行った分類の結果と実際の分類がどれだけ合っているか，検証，評価することができる．

検証実験を行うため，「文書」と「ラベル」のペアからなるデータセットを用意する必要がある．今回では，上記の朝日新聞データセットの「本文」項目の内容を「文書」とし，「記事分類」項目の内容の中，Table5 に示した記事テーマ分類表の小分類に該当するものを「ラベル」として用いる．検証実験の事前準備のために，全記事データにおける上記小分類の出現回数と文書ごとに付与されるラベルの数の出現頻度について調べ，その結果を Fig.2 と Fig.3 に示す．

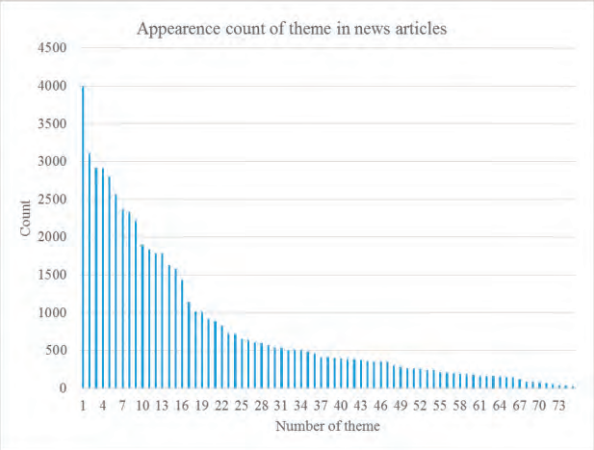

Fig. 2 Bar-chart of appearances of labels

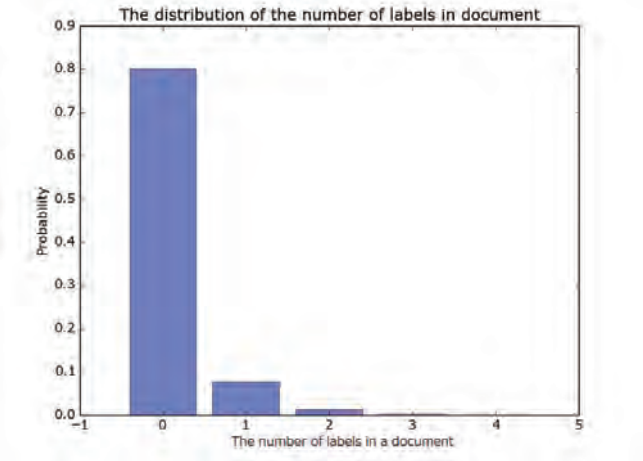

Fig. 3 The distribution of the number of labels in document

4.4 実験用データセット

前述の「朝日新聞記事データ（学術・研究用）2013 年」から記事全体におけるラベルの最小出現回数を 1500 回以上，800 回以上，400 回以上と 100 回以上別に，ラベルが一つ以上のテーマが付与された記事データから「本文」項目と「記事分類」項目を抽出してデータセットを作成した．得られたデータをランダムに順番付けしてから訓練データ 90%，テストデータ 10%の割合で訓練データセットとテストデータセットを作成した．各データセットの概要を Table6 に示す．

4.5 実験結果

4.5.1 トピック数をパラメータとする実験

前述の各データセットを用いて，提案手法について検証実験を行った．検証実験では，各データセットについて，LDA 法のトピック数を 10，30，50，70，90 と設定して，提案手法によりマルチラベル分類を行って，得られた推定ラベルの結果から精度，再現率，F 値，HammingLoss と OneError を算出した．実験の結果の数値を Table7～Table9 にまとめており，それぞれ

の表に対応するグラフを Fig.4～Fig.6 に示した．ただし，これらの実験においてラベル決定ルールのパラメータλを 2.0 とした．これらの実験結果から，提案手法が正常に動作することが確認できた．F 値の結果からみると，トピック数が大きくなると，F 値が大きくなって，性能が改善されるが，トピック数が小さいとき，改善幅が大きい，トピック数が大きいときは，改善幅が小さいことが分かる．また，Hamming Loss と One Error についても基本的に同様な傾向がみられるが，トピック数が 70 と 90 の時，値が大きくなり，性能が少し悪くなる事例もある．さらに，Hamming Loss と One Error の性能の良さに関して，事例の順位が逆になっているので，Hamming Loss にとって性能のよい事例が逆に One Error にとって性能の悪い事例となっていることが分かる．

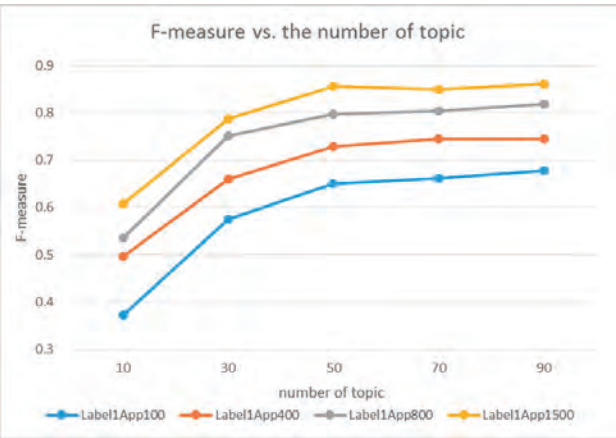

Fig.4　F-measure vs. the number of topic

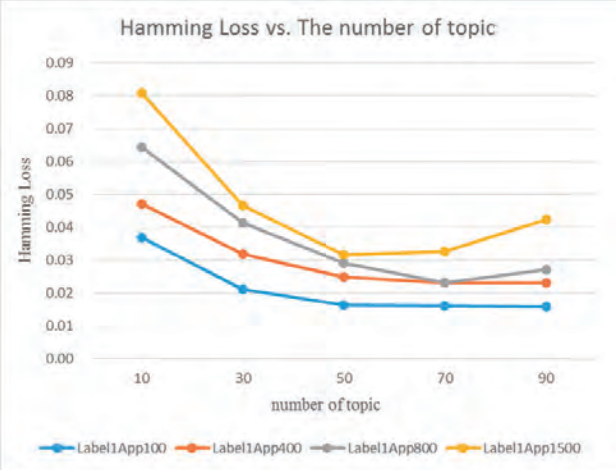

Fig.5　Hamming Loss vs. The number of topic

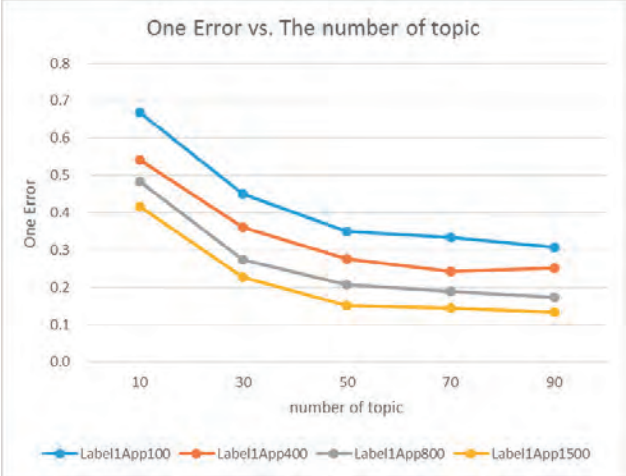

Fig.6　OneError vs. The number of topic

4.5.2 ラベル決定ルールパラメータλに関する実験

前述検証実験の中から，性能の良い事例と性能の悪い事例を一つずつ用いて，ラベルの出現確率からラベルを決定するルールの中にパラメータとして用いたλの値について検証実験を行った．検証実験では，パラメータλの値を 1.0, 1.5, 2.0, 2.5, 3.0 と設定して，マルチラベル分類を行って，得られた推定ラベルの結果から精度，再現率，F 値，HammingLoss と OneError を算出した．実験の結果の数値を Table10～Table12 にまとめており，それぞれの表に対応するグラフを Fig.7～Fig.9 に示した．これらの実験結果から，パラメータλは，1.0 から 3.0 まで範囲内では，分類器の性能にそれほど影響しないことが確認できた．これは決定ルールにラベル数の平均値や当該インスタンスのラベル確率の統計量導入したことによるものと考える．これにより，パラメータλの選択は比較的に容易である．

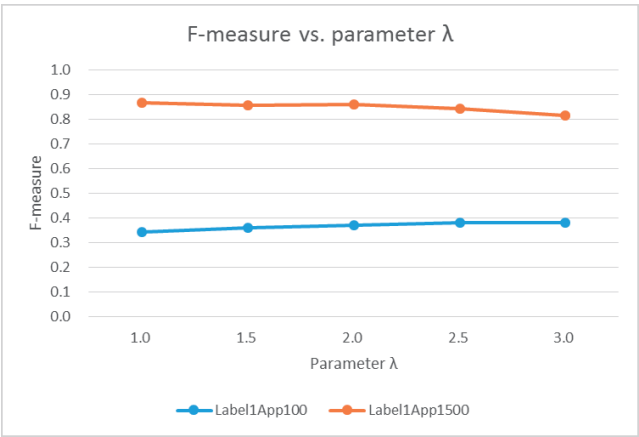

Fig.7　F-measure vs. Parameter　λ

4.6 既存手法との比較

4.6.1　MLKNN 法との比較

前述の検証実験で用いた 1 年間の新聞記事データから生成したデータセットに関しては，MLKNN 法は現実的な時間内にプログラムの処理を完了することができない．比較するため 1 か月の新聞記事データを用いて，データ量の異なる二つデータセットを作成し，MLKNN と提案手法の比較実験を行った．評価指標の F 値の計算結果および計算時間を Table1 に示した．

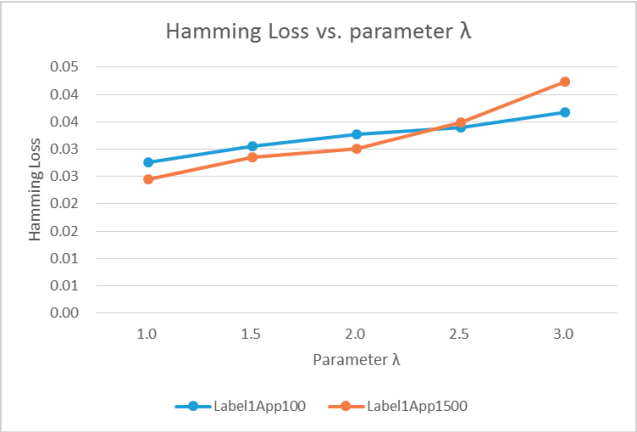

Fig.8 Hamming Loss vs. Parameter　λ

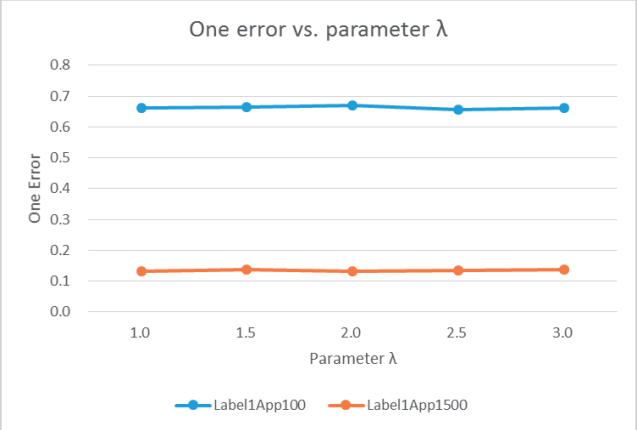

Fig.9 One Error vs. parameter　λ

Table1 Compare with MLKNN

Dataset	Training Instance	Test Instance	Labels	Method	F-measure	Time(s)	
						Step1(s)	Step2(s)
1	546	61	32	MLKNN	0.6562	813.6	
				Proposed	0.8062	113.7	1.5
2	976	109	6	MLKNN	0.798	4840.3	
				Proposed	0.9422	177.4	2.0

比較の結果，提案手法の方が性能は優れっているとともに，計算時間が大幅に短縮されたことが確認できる．また，データセット 1 からデータセット 2 にデータの量が 1.78 倍になったとき，MLKNN 法の計算時間が 5.95(>1.78^3=5.64)倍で，べき乗的に増加するが，提案手法の方が 1.56(<1.78)倍で，比例的増加量よりも少ないことが確認できる．これにより，MLKNN 法は大規模データの処理に不向きで，提案手法の方が大規模データの処理に適していることが確認できる．

4.6.2　文献 9 の実験結果との比較

文献 9 の検証実験では，著者等による手作業により Twitter データを 14 カテゴリに分けて生成したデータセットを用いて分類実験を行って，その結果を評価指標 F 値で評価した．同じデータを用いて比較実験を行うことができない．比較するために，朝日新聞記事データを用いてできるだけ近い指標のデータセットを作成した．それを用いた実験の結果と文献 9 から得られた結果を Table2 に示した．

Table2 Compare with the Ref.8

	Labels	LC	Precision	Recall	F-measure
Ref.8 (Rat; r=0)	14	3.4	0.57	0.67	0.59
Ref.8 (Rat; r=1)	14	3.4	0.66	0.54	0.58
Proposed Method	14	3.0	0.81	0.90	0.86

比較の結果，提案手法の方では，各性能指標とも比較対象より大幅に優れていることが確認できる．

4.6.3　文献 10 の実験結果との比較

文献 10 では，新聞記事の自動分類について報告した．この研究では，1 年分の新聞記事を 22 のカテゴリに分類する実験を行って，その性能を再現率で評価した．ただ，この研究でいう分類とは，シングルラベル分類と思われる．同じデータを用いて比較実験を行うことができないが，朝日新聞記事データを用いてできるだけ近い指標のデータセットを作成した．それを用いた実験の結果と文献 10 から得られた結果を Table3 に示した．

Table3 Compare with the Ref.9

	Labels	LC	Recall(maximum)
Ref.9	22	1.0	0.6974
Proposed Method	22	1.1	0.8640

比較の結果，提案手法の方では，性能評価指標の再現率とも比較対象より大幅に優れていることが確認できる．

5. おわりに

本論文では，トピックモデルと最小平均自乗推定との結合により，マルチラベル分類するための手法について提案し，その検証実験の結果について報告した．提案手法では，第 1 段階として，トピックモデルの LDA 法によって文書からトピックを抽出する準備処理を行う．そして，第 2 段階では，訓練データのラベルセットを用いて最小平均自乗推定を行い，トピック分布からラベル確率を推定した上で，最終的に提案ルールに従ってラベル決定する．朝日新聞記事データを用いた検証実験により，先行研究の 3 つ報告と比較して，提案手法の方が，分類性能や計算時間など面で顕著な改善があることが確認できた．

今後はラベル側の準備処理として，マトリクス分解の手法を導入し，ラベル間の相関を除去することを検討している．特徴量側とラベル側の両方で，独立化や無相関化の準備処理を行った上でラベルを推定することで，より高性能なマルチラベル分類器を実現できると考える．

Table4 Format of Asahi Shimbun article data

掲載年月日	刊種	紙誌	面名	本紙・地方面	記事分類	文字数	見出し	本文

Table5 News theme categories in Asahi Shimbun article data

No.	テーマ	出現回数	No.	テーマ	出現回数	No.	テーマ	出現回数
1	教育	4006	26	福祉	642	51	言論報道	262
2	裁判	3106	27	大戦	604	52	地域おこし	258
3	医療	2917	28	防衛・安保	598	53	建築	242
4	経済事件	2910	29	スポーツ事件	575	54	自然保護	239
5	食生活	2802	30	紛争	541	55	航空事故	213
6	交通事故	2566	31	高齢社会	540	56	倒産	206
7	音楽	2375	32	衣生活	508	57	麻薬事件	199
8	殺人傷害事件	2334	33	障害者	505	58	欠陥商品	191
9	労働・雇用	2215	34	住民運動	504	59	在日外国人	189
10	美術	1906	35	原子力事故	483	60	健康	177
11	火災	1832	36	政治倫理	461	61	被爆	162
12	災害	1784	37	鉄道事故	409	62	汚職事件	162
13	強盗窃盗事件	1782	38	領土領海	408	63	選挙違反	160
14	動物	1629	39	自殺	401	64	人権	153
15	学校の事件	1575	40	皇室王室	392	65	知的所有権	147
16	商品	1434	41	水の事故	382	66	脱税事件	142
17	公務員の不祥事	1143	42	人質事件	381	67	地球環境	120
18	植物	1013	43	宇宙	377	68	住生活	90
19	文化財	1009	44	公害	355	69	国際軍事	83
20	風俗・性犯罪	914	45	ネットワーク事件	353	70	民族	77
21	選挙結果	887	46	医療事件	348	71	えん罪	67
22	外交政策	830	47	テロ事件	347	72	密出入国事件	49
23	食品衛生	734	48	山の事故	300	73	情報公開	37
24	家庭の事件	719	49	核問題	286	74	平和運動	35
25	青少年犯罪	645	50	ごみ問題	264	75	行政改革	30

Table6 Summary of experimental Data set

Data Set Condition		Summary				
Least Label	Least Appearance	Training Instance	Test Instance	Category	Label Cardinality	Label Density
1	100	41948	4661	67	1.2394	0.0184
1	400	36021	4003	39	1.2123	0.0311
1	800	28627	3181	22	1.1219	0.0510
1	1500	23205	2579	15	1.0857	0.0724

Table7 Experimental result: F-measure

Data Set Condition		F-measure				
Least Label number	Least Label Appearance	LDA+LMS topic=10	LDA+LMS topic=30	LDA+LMS topic=50	LDA+LMS topic=70	LDA+LMS topic=90
1	100	0.3715	0.5757	0.6502	0.6618	0.6792
1	400	0.4967	0.6605	0.7295	0.7463	0.7458
1	800	0.5362	0.7516	0.7980	0.8049	0.8194
1	1500	0.6074	0.7874	0.8572	0.8502	0.8620

Table8 Experimental result: Hamming Error

Data Set Condition		Hamming Loss				
Least Label number	Least Label Appearance	LDA+LMS topic=10	LDA+LMS topic=30	LDA+LMS topic=50	LDA+LMS topic=70	LDA+LMS topic=90
1	100	0.03675	0.02105	0.01643	0.01615	0.01578
1	400	0.04705	0.03191	0.02475	0.02314	0.02320
1	800	0.06428	0.04131	0.02913	0.02314	0.02704
1	1500	0.08070	0.04645	0.03153	0.03264	0.04242

Table9 Experimental result: One Error

Data Set Condition		One Error				
Least Label number	Least Label Appearance	LDA+LMS topic=10	LDA+LMS topic=30	LDA+LMS topic=50	LDA+LMS topic=70	LDA+LMS topic=90
1	100	0.6695	0.4499	0.3492	0.3353	0.3070
1	400	0.5405	0.3602	0.2752	0.2425	0.2518
1	800	0.4838	0.2741	0.2065	0.1902	0.1735
1	1500	0.4176	0.2279	0.1504	0.1454	0.1329

Table10 Experimental result: F-measure (parameter λ)

Data Set Condition		Topic	F-measure				
Least Label number	Least Label Appearance		LDA+LMS ramda=1.0	LDA+LMS ramda=1.5	LDA+LMS ramda=2.0	LDA+LMS ramda=2.5	LDA+LMS ramda=3.0
1	100	10	0.3438	0.3621	0.3715	0.3823	0.3808
1	1500	90	0.8678	0.8587	0.8620	0.8453	0.8149

Table11 Experimental result: Hamming Loss (parameter λ)

Data Set Condition		Topic	Hamming loss				
Least Label number	Least Label Appearance		LDA+LMS λ=1.0	LDA+LMS λ=1.5	LDA+LMS λ=2.0	LDA+LMS λ=2.5	LDA+LMS λ=3.0
1	100	10	0.02764	0.03053	0.03267	0.03402	0.03675
1	1500	90	0.02453	0.02848	0.03001	0.03487	0.04241

Table12 Experimental result: One Error (parameter λ)

Data Set Condition		Topic	One error				
Least Label number	Least Label Appearance		LDA+LMS λ=1.0	LDA+LMS λ=1.5	LDA+LMS λ=2.0	LDA+LMS λ=2.5	LDA+LMS λ=3.0
1	100	10	0.6623	0.6655	0.6695	0.6575	0.6623
1	1500	90	0.1310	0.1380	0.1328	0.1337	0.1376

参考文献

1) T. Mitchell, “Machine Learning”, McGraw-Hill, Boston, MA, 1997.
2) A. K. McCallum, “Multi-label Text Classification with a Mixture Model Trained by EM,” Proceedings of AAAI’99 Workshop on Text Learning, Orlando FL, 1999.
3) N. Ueda and K. Saito, “Parametric Mixture Models for Multi-label Text,” in: S. Becker, S. Thrun and K. Obermayer (eds.) Advances in Neural Information Processing System 15, MIT Press, Cambridge, MA, pp.721-728, 2003.
4) R. E. Schapire and Y. Singer, “Bootexter: A Boosting-based System for Text Categorization,” Machine Learning, vol.39(2-3), pp.135-168, 2000.
5) A. Elisseeff and J. Weston, “A Kernel Method for Multi-labelled Classification,” In T. G. Dietterich, S. Becker and Z. Ghahramani (eds) Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, pp.681-687, 2001.
6) M. Zhang and Z. Zhou, “ML-KNN: A Lazy Learn Approach to Multi-label Learning,” Pattern Recognition, vol.40(7), pp.2038-2048, 2007.
7) Z. Wei, H. Zhang, Z. Li, W. Zhang and D. Miao, “A Naive Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Results,” International Journal of Advanced Intelligence, vol.3(2), pp. 173-188, 2011.
8) T. Lo H. Chiang and S. Lin, “A Ranking-based KNN Approach for Multi-label Classification,” JMLR: Workshop and Conference Proceedings, vol.25, pp.81-96, 2012.
9) 山本修平, 佐藤哲司, “実生活 Tweet に対する局面の階層的推定法,” Proceedings DEIM Forum, C4-1,2014.
10) 森本由起子, 間瀬久雄, 辻洋, 絹川博之, 新聞記事の自動分類システム構築の検討と評価, 情報処理学会第 53 回全国大会, 3-205, 1998.
11) 村上浩司, 野畑周, 関根聡, 井佐原均, 新聞記事を対象にした検索, 分類, 複数文書要約システム ELIOT システム. 言語処理学会第 10 回年次大会発表論文集, pp.143-146, 2004.
12) D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, vol.3, pp.993-1022, 2003.
13) G. Tsoumakas and I. Katakis, “Multi-Label Classification: An Overview,” International Journal of Data Warehousing and Mining, vol.3 (3), pp.1-13, 2007.