Vol.2010-MPS-80 No.8
2010/9/28

情報処理学会研究報告
IPSJ SIG Technical Report

ラベルありデータの選択バイアスに頑健な半教師あり学習

藤野昭典[†1]　上田修功[†1]　永田昌明[†1]

本稿では，目標ドメインのテストデータと分布が大きく異なるラベルありデータから汎化性能が高い分類器を設計するための頑健な半教師あり学習法を提案する．半教師あり学習の枠組のひとつである JESS-CM 法は複数の自然言語処理タスクで最良の結果を達成したが，本稿のタスク設定ではラベルありデータに過適合する危険性がある．提案法では，分類器を構成する識別・生成モデルの双方の学習に目標ドメインのラベルなしデータをラベルありデータと同時に用いることで過適合の問題を解決することを期待する．3 つのテストコレクションを用いたテキスト分類実験により，本タスク設定のほとんどの場合で，提案法では JESS-CM 法よりも高い分類性能を得られることを確認した．

Robust Semi-supervised Learning for Labeled Data Selection Bias

Akinori Fujino,[†1] Naonori Ueda[†1]
and Masaaki Nagata[†1]

We propose a robust semi-supervised learning method for designing good classifiers with a high generalization ability from labeled data whose distribution differs largely from that of test data in a target domain. Although JESS-CM is one of the most successful semi-supervised learning methods that achieved the best published results in natural language processing tasks, it has an overfitting problem in our task setting. We expect the proposed method to solve the overfitting problem by utilizing unlabeled data in the target domain with the labeled data for both training of discriminative and generative models composing a classifier. Our experimental results for text classification using three test collections confirmed that the classification performance obtained with the proposed method was better than that with JESS-CM in most case of the task setting.

1. はじめに

機械学習に基づく自動分類では，属するクラスが既知のデータ (ラベルありデータ) を用いて分類器を学習させ，新規データ (テストデータ) の属するクラスを推定する．一般に，分類対象となるテストデータ集合と類似する特徴ベクトルの分布をもつラベルありデータ集合を訓練データとして用いることができる場合，高性能な分類器が得られる．しかし，実問題では，テストデータ集合と異なる分布をもつラベルありデータ集合から分類器を学習させる必要がある．例えば，ニュース記事の自動分類では，日々ニュースの話題が変わるため，過去に作成したラベルありデータ集合と分類対象となる最新のニュースデータ集合の単語の分布は大きく異なる可能性がある．また，Web ページの自動分類では，分類対象となる膨大なテストデータの分布を網羅するようにラベルありデータ集合を作成するのは容易ではない．このため，テストデータ集合と分布が異なるラベルありデータ集合と，テストデータ集合と類似した分布をもつラベルなしデータ集合から高性能な分類器を設計することは機械学習分野の重要な課題であり，これまで転移学習[6] 問題の一種である標本選択バイアス (sample selection bias)[12]，共変量シフト (covariate shift)[7]，教師なしドメイン適応 (unsupervised domain adaptation)[4] 問題として研究されてきた．

この課題に対して，従来研究では，訓練データの重み付け，あるいは特徴空間の変換に基づいて分類器の性能を向上させる手法が提案されてきた[6]．重み付けによる手法では，文献[7),8),12)] のように，テストデータ集合に対する期待損失を最小化させるように個々に重みを与えたラベルありデータ集合を用いて分類器を学習させる．また，ラベルなしデータにも重みを与えて，分類器の半教師あり学習に適用する手法も提案されている[4),11)]．特徴空間の変換に基づく手法では，文献[1)] のように，データ $\boldsymbol{x}$ の特徴空間 $\mathcal{X}$ を，ラベルありデータ集合の分布とテストデータ集合の分布の差を小さくするように変換した特徴空間 Z 上で機械学習法を適用することで分類器の性能を向上させる．

上述の訓練データの重み付けや特徴空間の変換に基づく手法は，従来の教師あり学習や半教師あり学習[13)] に基づく分類器をベースの分類器として用い，その分類器を適用する際の前処理などの手段を与える．これらの方法はテストデータ集合とラベルありデータ集合間の分布差が大きい場合に自動分類の精度を向上させる効果的な手段であるが，ベースとなる

†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

分類器自体を改良することもまた自動分類の精度を向上させるために重要であると考えられる．

本稿では，後者の立場から，テストデータ集合と分布が異なるラベルありデータ集合から高い汎化性能をもつ分類器を設計するための半教師あり学習法を提案する．提案法は，識別モデルと生成モデルの統合に基づいて分類器の条件付確率モデルを与える点で JESS-CM (Joint probability model Embedding style Semi-Supervised Conditional Model) 法[9),10)] と類似する．JESS-CM 法は複数の自然言語処理タスクで最良の精度を達成した半教師あり学習法の枠組であるが，テストデータ集合とラベルありデータ集合の分布が大きく異なる場合を想定して開発された手法ではない．提案法ではこのような分布の違いに対処する頑健な学習アルゴリズムを与える．具体的には，JESS-CM 法では分類器の識別学習を行うのにラベルありデータのみを用いるのに対して，提案法ではラベルあり・なしデータの両方を識別・生成モデルの双方の学習に用いることでラベルありデータへの過適合を抑制し，テストデータの分類精度が向上することを期待する．

生成・識別モデルとしてナイーブベイズ (NB) モデルと多項ロジスティック回帰 (MLR) モデルを用いて，提案法をテキスト分類問題に適用する．3 つのテストコレクションを用いた実験により，ラベルありデータ集合の分布がテストデータ集合の分布と大きく異なる問題において，提案法が JESS-CM 法よりも高い自動分類の精度を与えることを示す．

2. タスク設定

本稿では，各データに対して，事前に定義された K 個のクラスラベルの中から 1 つのクラスラベルを選択する多値分類問題に焦点を当てる．V 次元の特徴空間 $\mathcal{X}$ における各データの特徴ベクトルを $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_V)^T \in \mathcal{X}$ で表し，各データのクラスラベルを $y \in \{1, \ldots, k, \ldots, K\}$ で表す．

標本選択バイアス，共変量シフト，教師なしドメイン適応の問題では，一般的に，分類対象となるテストデータが属する領域を目標ドメイン (target domain) と呼び，テストデータ集合と分布が異なるラベルありデータ集合が属する領域を元ドメイン (source domain) と呼ぶ．これらの問題では，目標ドメインと元ドメインの間に以下の関係があると仮定される．

(1) 目標ドメインの特徴空間 $\mathcal{X}_t$ と元ドメインの特徴空間 $\mathcal{X}_s$ は同一: $\mathcal{X}_t = \mathcal{X}_s$
(2) 目標ドメインのデータの確率分布 $p_t(\boldsymbol{x})$ と元ドメインのデータの確率分布 $p_s(\boldsymbol{x})$ が大きく異なる: $p_t(\boldsymbol{x}) \neq p_s(\boldsymbol{x})$
(3) データがクラスに属する確率の分布には目標ドメインと元ドメインの間で類似性がある: $P_t(y|\boldsymbol{x}) \simeq P_s(y|\boldsymbol{x})$

$p_t(\boldsymbol{x}) \simeq p_s(\boldsymbol{x})$ を想定する従来の機械学習とは，(2) の仮定が異なる．本研究では，元ドメインのみから集められたラベルありデータ集合 $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ と目標ドメインから集められたラベルなしデータ集合 $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$ を用いた半教師あり学習により，目標ドメインのテストデータに適した分類器を設計するタスクを対象とする．

3. 提 案 法

3.1 基本的な枠組

本稿では，目標ドメインのラベルなしデータと元ドメインのラベルありデータから半教師あり学習に基づいて汎化性能が高い分類器を設計するための手法 MHLE (Maximum Hybrid Log-likelihood Expectation) を提案する．提案法では，識別モデル $P_d(y|\boldsymbol{x}; W)$ と生成モデル $p_g(\boldsymbol{x}, y; \Theta)$ の統合に基づく分類器の条件付確率モデルとパラメータ推定アルゴリズムを単一の目的関数を用いて定式化する．

提案法では，目標ドメインと元ドメインの両方のデータで共通する真の条件付確率分布 $P(y|\boldsymbol{x})$ が存在すると仮定する．理想的な $P(y|\boldsymbol{x})$ は，正解のクラスラベルが y_n であるデータ $\boldsymbol{x}_n$ に対して以下を満たす．

$$P(y|\boldsymbol{x}_n) = \begin{cases} 1 & y = y_n \text{ の場合} \\ 0 & \text{それ以外} \end{cases} \tag{1}$$

$P(y|\boldsymbol{x})$ が既知である状況下では，データ $\boldsymbol{x} \in D$ に対して，生成モデル $p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y)$ のパラメータ $\boldsymbol{\theta}_y$ の推定値を，対数尤度の期待値に基づく以下の目的関数を最大化させる $\Theta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K]$ を計算することで得られる．

$$J_g(\Theta|D) \equiv \sum_{\boldsymbol{x} \in D} \sum_{k=1}^{K} P(k|\boldsymbol{x}) \log p_g(\boldsymbol{x}, k; \boldsymbol{\theta}_k) + \log p(\Theta) \tag{2}$$

上式中の $p(\Theta)$ は Θ の事前確率分布を表す．また，識別モデル $P_d(y|\boldsymbol{x}; W)$ のパラメータ W の推定値を，$P(y|\boldsymbol{x})$ と $P_d(y|\boldsymbol{x}; W)$ の KL (Kullback-Leibler) ダイバージェンス最小化に基づく以下の目的関数を最大化させる W を計算することで得られる．

$$J_d(W|D) \equiv -\sum_{\boldsymbol{x} \in D} \sum_{k=1}^{K} P(k|\boldsymbol{x}) \log \frac{P(k|\boldsymbol{x})}{P_d(k|\boldsymbol{x}; W)} + \log p(W) \tag{3}$$

ただし，$p(W)$ は W の事前確率分布である．ここで，生成モデル $p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y)$ と識別モデル $P_d(y|\boldsymbol{x}; W)$ をラベルありデータ集合 $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ のみで学習させる場合，式 (1) を式 (2) と式 (3) に代入すると，

$$J_g(\Theta|D_l) = \sum_{n=1}^{N} \log p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n}) + \log p(\Theta)$$
$$J_d(W|D_l) = \sum_{n=1}^{N} \log P_d(y_n|\boldsymbol{x}_n; W) + \log p(W)$$

が得られる．したがって，$J_g(\Theta|D)$ と $J_d(W|D)$ は識別モデルと生成モデルの教師あり学習でよく用いられる MAP 推定の目的関数を単純に拡張したものであるといえる．

しかし，ラベルなしデータを含めたすべてのデータの条件付確率を与える真の分布 $P(y|\boldsymbol{x})$ は未知であり，W, Θ と同様に推定する必要がある．提案法では，$J_d(W|D)$ と $J_g(\Theta|D)$ の線形結合で定義される以下の目的関数の最大化によって $P(y|\boldsymbol{x})$ を推定する．

$$J(W, \Theta, P) \equiv J_d(W|D) + \beta J_g(\Theta|D) \tag{4}$$

上式中の $\beta(\geq 0)$ は，$J_d(W|D)$ と $J_g(\Theta|D)$ の統合の重みである．$\sum_{k=1}^K P_c(k|\boldsymbol{x}) = 1$ の制約条件の下でラグランジュ未定乗数法を用いると，$J(W, \Theta, P)$ を最大化させる $P(y|\boldsymbol{x})$ の解

$$P(y|\boldsymbol{x}; W, \Theta, \beta) = \frac{P_d(y|\boldsymbol{x}; W) p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y)^\beta}{\sum_{k=1}^K P_d(k|\boldsymbol{x}; W) p_g(\boldsymbol{x}, k; \boldsymbol{\theta}_k)^\beta} \tag{5}$$

を得ることができる．提案法では，式 (5) のように識別・生成モデルを統合して得られる $P(y|\boldsymbol{x}; W, \Theta, \beta)$ を，真の分布 $P(y|\boldsymbol{x})$ の良いモデルを与えると考え，分類器の条件付確率モデルとして用いる．

W と Θ の値の推定には，ラベルありデータ集合 $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ とラベルなしデータ集合 $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$ を用いる．式 (2), (3) で表される $J_g(\Theta|D)$, $J_d(W|D)$ に含まれるラベルありデータの条件付確率に式 (1) を代入し，ラベルなしデータの条件付確率に式 (5) を代入することで，式 (4) を以下のように変形できる．

$$J(W, \Theta) = \log p(W) + \beta \log p(\Theta) + \sum_{n=1}^{N} \log P_d(y_n|\boldsymbol{x}_n; W) p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n})^\beta + \sum_{m=1}^{M} \log \sum_{k=1}^{K} P_d(k|\boldsymbol{x}_m; W) p_g(\boldsymbol{x}_m, k; \boldsymbol{\theta}_k)^\beta \tag{6}$$

β の値を設定し，$J(W, \Theta)$ を最大化させる W と Θ を計算することで W と Θ の推定値を得ることができる．紙面の都合により，パラメータ推定アルゴリズムの詳細を省略する．

3.2 従来法との相違

従来法である JESS-CM (Joint probability model Embedding style Semi-Supervised Conditional Model)[9),10)] では，生成モデル $p_g(\boldsymbol{x}, \boldsymbol{y}; \Theta)$ と，パラメータベクトルと特徴ベクトルの内積で与えられる線形の識別関数 $f(\boldsymbol{x}, \boldsymbol{y}; W)$ を用いて，以下のように分類器の条件付確率モデルを定義する[⋆1]．

$$P(\boldsymbol{y}|\boldsymbol{x}; W, \beta, \Theta) = \frac{f(\boldsymbol{x}, \boldsymbol{y}; W) p_g(\boldsymbol{x}, \boldsymbol{y}; \Theta)^\beta}{\sum_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}'; W) p_g(\boldsymbol{x}, \boldsymbol{y}'; \Theta)^\beta}$$

W と Θ はそれぞれ識別関数と生成モデルのパラメータを表し，β は生成モデルの統合の重みを与えるパラメータである．

JESS-CM 法では，以下のように個々に定義される 2 つの目的関数を用いてパラメータ値を推定する．

$$\mathcal{L}^1(W, \beta|\Theta) \equiv \sum_{n=1}^{N} \log P(\boldsymbol{y}_n|\boldsymbol{x}_n; W, \beta, \Theta) + \log p(W, \beta)$$
$$\mathcal{L}^2(\Theta|W, \beta) \equiv \sum_{m=1}^{M} \log \sum_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}_m, \boldsymbol{y}; W) p_g(\boldsymbol{x}_m, \boldsymbol{y}; \Theta)^\beta + \log p(\Theta)$$

W と β の推定では Θ を固定した上で $\mathcal{L}^1(W, \beta|\Theta)$ を最大化させる W と β の値を計算し，Θ の推定では W と β を固定した上で $\mathcal{L}^2(\Theta|W, \beta)$ を最大化させる Θ の値を計算する．W, β の値と Θ の値には依存関係があるため，$\mathcal{L}^1(W, \beta|\Theta)$ の最大化による W, β の値の推定と $\mathcal{L}^2(\Theta|W, \beta)$ の最大化による Θ の値の推定を交互に繰り返すことでパラメータの推定値を求める．

図 1 に，JESS-CM 法と MHLE 法の違いを示す．JESS-CM 法では，生成モデルのパラメータ推定にはラベルなしデータのみを用い，識別関数のパラメータ値の識別学習にはラベルありデータのみを用いる．しかし，本研究では，目標ドメインと元ドメインのデータの分布が異なる上に，目標ドメインからラベルありデータを得られない状況で分類器を学習させるタスクを対象とする．このタスク設定では，JESS-CM 法のように識別関数のパラメータ値をラベルありデータのみに適合させることで，目標ドメインのデータには必ずしも適

⋆1 文献[9)] では複数の生成モデルを用いて分類器を設計しているが，識別関数と生成モデルの統合方法とパラメータ推定法の基本的な枠組は生成モデルの数 J によらず同一である．そこで，本論文では $J = 1$ としてモデルを単純化して議論する．

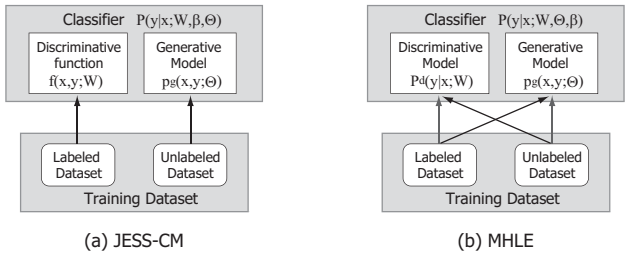

図 1 JESS-CM 法と MHLE 法の概要
Fig. 1 Outline of JESS-CM and MHLE methods

するとは限らない分類器を得る危険性があると考えられる．一方，提案法ではラベルあり・なしデータの両方を識別・生成モデルの双方の学習に用いる．目標ドメインでの分類精度を向上させるために，目標ドメインのラベルなしデータを識別モデルの学習に用いることで元ドメインへの分類器の過学習を抑制する．

3.3 テキスト分類への応用

提案法をテキスト分類に応用するため，生成モデル $p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y)$ にナイーブベイズ (NB) モデルを適用する．i 番目の単語の出現頻度 $\boldsymbol{x}_i$ を用いて文書の特徴ベクトルを $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_V)^T$ で表す．V は文書集合全体に含まれる語彙の総数を表す．NB モデルでは，クラス y での $\boldsymbol{x}$ の確率分布が多項分布 $p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y) \propto \pi_y \prod_{i=1}^{V} (\theta_{yi})^{x_i}$ に従うと仮定する．ここで，θ_{yi} (> 0) はクラス y での i 番目の単語の出現確率を表し，$|\boldsymbol{\theta}_y| = \sum_{i=1}^{V} \theta_{yi} = 1$ を満たす．$\boldsymbol{\theta}_y = (\theta_{y1}, \ldots, \theta_{yi}, \ldots, \theta_{yV})^T$ は訓練データから値を推定すべきパラメータである．π_y はクラス y の確率を表す．本応用では，$\pi_y = 1/K$, $\forall y$ に設定し，特徴ベクトルを $|\boldsymbol{x}| = \sum_{i=1}^{V} x_i = 1$ のように正規化して適用した．式 (6) 中の $p(\Theta)$ には，ディリクレ事前確率分布 $p(\Theta) \propto \prod_{k=1}^{K} \prod_{i=1}^{V} (\theta_{ki})^{\xi}$ を適用した．ξ (> 0) はハイパーパラメータである．

識別モデル $P_d(y|\boldsymbol{x}; W)$ には，多項ロジスティック回帰 (MLR) モデル $P_d(y|\boldsymbol{x}; W) = \exp\left(\boldsymbol{w}_y^T \boldsymbol{x}\right) / \sum_{k=1}^{K} \exp\left(\boldsymbol{w}_k^T \boldsymbol{x}\right)$ を適用する．$W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, \ldots, \boldsymbol{w}_K]$ は，訓練データから値を推定すべきパラメータである．式 (6) 中の $p(W)$ には，ガウス事前確率分布 $p(W) \propto \prod_{k=1}^{K} \exp\left(-\boldsymbol{w}_k^T \boldsymbol{w}_k / 2\sigma^2\right)$ を適用した．

4. 評価実験

4.1 テストコレクション

テキスト分類タスクでベンチマークテストによく利用される 3 つのテストコレクション *WebKB*, *SRAA*, *20 newsgroups (20news)* を用いて評価実験を行った．

WebKB[⋆1] は 4 つの大学から集められた web ページと雑多な情報源から集められた web ページから構成される．これらの web ページは 7 つのカテゴリに分類されている．文献[5] の設定に従い，4 つのカテゴリ *course, faculty, project, student* 含まれる 4199 の web ページを評価実験に利用した．4 つの大学から集められた web ページを元ドメインのデータとし，雑多な情報源から集められた web ページを目標ドメインのデータとした．各 web ページに含まれるタグ，リンク情報を除外し，停止語 (stop words) と 1 つの web ページのみに含まれる単語以外の 18525 語彙を用いて web ページの特徴ベクトルを与えた．

SRAA[⋆2] は 4 つのカテゴリ (*sim-auto, sim-aviation, real-auto, real-aviation*) に属する 73218 の UseNet 記事を集めたデータセットである．評価実験では，この 4 つのカテゴリに各記事を分類する問題に提案法を適用した．2 章で述べたように，本研究のタスク設定では，目標ドメインに属するデータの確率分布と元ドメインに属するデータの確率分布が異なることを仮定する．この設定で評価実験を行うため，球面 K-平均法[2] を用いて UseNet 記事を 2 つのサブセットに分割し，一方のサブセットを目標ドメイン，もう一方のサブセットを元ドメインとして用いた．各 UseNet 記事に含まれる *subject* 以外のヘッダテキストを除外し，停止語と 1 つの UseNet 記事のみに含まれる単語以外の 61526 語彙を用いて UseNet 記事の特徴ベクトルを与えた．

20news[⋆3] は 20 グループに属する UseNet 記事を集めたデータセットである．評価実験では，*comp* のグループに属する 4881 記事を用いて，5 つのサブグループに分類する問題に提案法を適用した．SRAA と同じ方法で，目標ドメインと元ドメインのサブセットを作成し，特徴ベクトルに用いる 19383 語彙を選択した．

4.2 実験設定

提案法 MHLE と従来の 3 つの半教師あり学習法 JESS-CM, NB/EM-λ, MLR/MER で得られるテキスト分類精度を比較することで，提案法の性能を評価する．実験では，ラベルありデータのみを用いて学習する 2 つの教師あり学習法 NB, MLR で得られる分類精度も比較した．

MHLE に基づくテキスト分類器の学習では， 3.3 節で述べたガウス事前確率とディリクレ事前確率のハイパーメータ σ, ξ の値を事前に設定する必要がある．本実験では，それぞ

⋆1 http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.gtar.gz
⋆2 http://www.cs.umass.edu/~mccallum/data/sraa.tar.gz
⋆3 http://people.csail.mit.edu/jrennie/20Newsgroups/20news-18828.tar.gz

れ $\sigma^2 \in \{10^n\}_{n=-1}^{6}, \xi \in \{10^n\}_{n=-4}^{-1}$ の候補値の中から選択して設定した．また，β の値を $\beta \in \{5 \times 10^n\}_{n=-1}^{2}$ の候補値の中から選択した．

JESS-CM では，識別関数と生成モデルにそれぞれ MLR モデルの識別関数と NB モデルを適用してテキスト分類器を実装した．モデルパラメータの事前確率分布には MHLE と同じガウス事前確率分布とディリクレ事前確率分布を用い，ハイパーパラメータにも同じ候補値を用いた．

MLR/MER では，最小エントロピー正則化項[3] を用いて MLR モデルの半教師あり学習を行う．本実験では，正則化項の重みパラメータの値を $\lambda \in \{\{1\times10^n, 2\times10^n, 5\times10^n\}_{n=-5}^{0}\}$ の候補値の中から選択して設定した．また，MLR/MER と MLR の学習では，MHLE と同じガウス事前確率分布を用い，ハイパーパラメータにも同じ候補値を用いた．

NB/EM-λ では，EM-λ アルゴリズム[5] を用いて NB モデルの半教師あり学習を行う．本実験では，ラベルなしデータの対数尤度項の重みパラメータの値を $\lambda \in \{0.01, 0.02, 0.05, \{n \times 0.1\}_{n=1}^{9}, 1, 2, 5\}$ の候補値の中から選択した．また，NB/EM-λ と NB の学習では，MHLE と同じディリクレ事前確率分布を用い，ハイパーパラメータの値を $\xi \in \{\{1\times10^n, 2\times10^n, 5\times10^n\}_{n=-3}^{-1}, 1\}$ の候補値の中から選択した．

上述のハイパーパラメータと重みパラメータの設定では，ラベルありデータの 5 分割交差検定で最良の結果を与える候補値を選択した．SRAA と 20news では，データの特徴ベクトルを単語頻度に基づく特徴量を用いて与えた．WebKB では，単語頻度よりも単語が含まれるか否かを表すバイナリ特徴量の方が良い自動分類精度を与えたため，バイナリ特徴量に基づいて特徴ベクトルを与えた．NB/EM-λ と NB を除く 4 つの分類器では，特徴ベルトルの大きさの分散による悪影響を抑えるため，L1 ノルムで正規化された特徴ベクトルを使用した．

学習に用いるラベルありデータとラベルなしデータを元ドメインと目標ドメインのサブセットからそれぞれ無作為に抽出した．SRAA と 20news では 2500 個のラベルなしデータを，WebKB では 2000 個のラベルなしデータを分類器の学習に使用した．また，サブセットに含まれる残りのデータから，800 個の目標ドメインのテストデータと 200 個の元ドメインのテストデータを無作為に抽出した．この無作為抽出を繰り返して，10 通りの評価用データセットを作成し，10 通りの実験で得られるテストデータの分類精度の平均値で分類器の性能比較を行った．

4.3 実験結果

表 1 に 6 つの分類器を用いて得られるテストデータの平均分類精度 (%) を示す．括弧内の数値は分類精度の分散である．また，r_l はラベルありデータ集合に含まれる語彙の数に対するラベルありデータ集合とラベルなしデータ集合の両方に含まれる語彙の数の比率 (%) であり，r_u はラベルなしデータ集合に含まれる語彙の数に対する両方の集合に含まれる語彙の数の比率 (%) である．r_l, r_u の値が小さいほどラベルあり集合の分布とラベルなし集合の分布の重なりが小さいことを示す．

表 1 目標ドメインと異なる分布から集めた N 個のラベルありデータを用いて得られる分類精度 (%)
Table 1 Classification accuracies (%) obtained by using N labeled data collected from a different distribution from target domain.

dataset	N	MHLE	JESS-CM	NB/EM-λ	MLR/MER	NB	MLR	r_l	r_u
WebKB	20	**69.4** (6.5)	50.8 (9.3)	65.9 (3.4)	46.6 (10.3)	60.5 (5.4)	46.9 (10.2)	94.1	7.5
	100	**72.3** (5.6)	66.8 (5.1)	69.7 (5.4)	63.2 (4.8)	69.2 (4.2)	63.3 (4.7)	90.6	21.6
	500	**88.1** (1.4)	82.5 (2.4)	73.0 (3.2)	81.2 (1.6)	72.8 (2.7)	81.3 (2.0)	85.5	46.2
SRAA	50	44.2 (5.5)	**46.1** (4.3)	36.1 (5.5)	38.8 (4.2)	34.2 (4.5)	38.0 (3.9)	79.8	8.4
	200	**48.0** (3.4)	47.4 (4.6)	34.6 (6.0)	38.1 (6.7)	40.0 (3.6)	39.7 (3.6)	72.4	20.3
	1000	**50.0** (3.1)	48.7 (2.9)	32.5 (4.0)	44.8 (3.8)	45.3 (3.3)	45.3 (2.4)	57.8	42.3
20news	50	**42.4** (13.0)	41.6 (12.2)	31.1 (15.2)	17.4 (5.6)	19.8 (5.9)	22.4 (3.2)	83.2	9.8
	200	**49.7** (13.6)	48.4 (14.7)	36.3 (14.7)	23.0 (6.9)	25.8 (8.3)	23.3 (6.2)	78.2	22.4
	1000	**71.6** (10.3)	52.8 (7.2)	49.9 (17.4)	37.3 (5.1)	36.8 (9.0)	37.6 (5.2)	74.9	45.8

(a) 目標ドメインのテストデータ (For target domain test data)

dataset	N	MHLE	JESS-CM	NB/EM-λ	MLR/MER	NB	MLR	r_l	r_u
WebKB	20	71.4 (7.3)	69.5 (6.4)	**77.0** (2.4)	66.2 (7.2)	72.0 (3.6)	66.2 (6.6)	94.1	7.5
	100	83.0 (1.5)	**83.8** (3.6)	80.3 (2.5)	83.5 (2.4)	81.5 (2.8)	83.5 (2.4)	90.6	21.6
	500	90.8 (2.4)	91.4 (2.3)	82.9 (3.5)	**91.6** (2.4)	83.7 (3.1)	91.5 (2.3)	85.5	46.2
SRAA	50	67.1 (6.8)	66.0 (4.5)	**69.6** (2.5)	62.9 (3.8)	64.7 (2.8)	63.3 (3.6)	79.8	8.4
	200	79.0 (2.9)	78.3 (4.0)	**81.6** (2.3)	75.5 (2.6)	78.0 (2.8)	75.5 (2.6)	72.4	20.3
	1000	**88.3** (3.2)	85.8 (2.3)	86.5 (2.7)	85.2 (2.3)	86.6 (2.7)	84.6 (2.4)	57.8	42.3
20news	50	62.5 (6.4)	**64.5** (7.1)	56.8 (7.0)	64.0 (5.6)	61.5 (4.7)	66.1 (4.6)	83.2	9.8
	200	**76.8** (3.0)	76.0 (3.4)	71.9 (6.8)	75.3 (3.6)	71.8 (7.3)	75.1 (4.1)	78.2	22.4
	1000	**86.8** (3.7)	85.8 (3.5)	82.2 (5.5)	84.9 (3.1)	83.2 (4.0)	84.9 (2.9)	74.9	45.8

(b) 元ドメインのテストデータ (For source domain test data)

表 1 (a) に示されるように，目標ドメインのテストデータの分類精度は，純粋な生成モデル，識別モデルである MLR/MER, NB/EM-λ よりも MHLE の方が高かった．MHLE と JESS-CM で得られた目標ドメインのテストデータの分類精度を比較すると，WebKB と 20news では MHLE の方が JESS-CM よりも高く，SRAA ではほぼ同等であった．また，元ドメインのテストデータの分類精度は，表 1 (b) に示されるように，MHLE は JESS-CM とほぼ同等であった．すなわち，MHLE を用いることで，元ドメインでの汎化性能を維持

表 2 単一ドメインの設定で N 個のラベルありデータを用いて得られる分類精度 (%)
Table 2 Classification accuracies (%) obtained by using N labeled data in a single domain setting.

dataset	N	MHLE	JESS-CM	NB/EM-λ	MLR/MER	NB	MLR	r_l	r_u
WebKB	20	70.3 (4.7)	62.3 (7.7)	**73.2** (4.8)	59.2 (7.5)	63.6 (4.0)	59.4 (7.4)	96.9	7.3
	100	81.3 (1.6)	**82.0** (1.6)	77.8 (2.4)	80.1 (1.5)	76.0 (1.9)	80.0 (1.5)	94.2	21.6
	500	89.5 (1.1)	**91.0** (0.9)	83.3 (1.1)	**91.0** (1.7)	83.0 (1.2)	90.4 (1.3)	89.9	51.0
SRAA	50	**58.0** (5.0)	57.1 (5.1)	56.6 (3.8)	49.0 (1.5)	49.4 (2.2)	48.9 (1.3)	90.3	7.7
	200	64.9 (1.1)	**65.8** (1.3)	61.5 (2.2)	57.2 (1.7)	59.4 (1.2)	57.2 (1.7)	84.6	20.3
	1000	71.3 (1.2)	**71.9** (1.2)	70.8 (1.7)	68.5 (1.0)	70.1 (1.6)	68.3 (0.9)	72.4	45.3
20news	50	71.2 (2.6)	**72.7** (1.5)	60.5 (6.0)	51.4 (5.7)	48.6 (4.7)	51.9 (4.6)	95.4	10.5
	200	78.0 (1.3)	**78.3** (1.4)	67.1 (4.8)	72.3 (1.8)	64.6 (3.8)	69.5 (2.0)	92.4	27.1
	1000	**84.7** (1.5)	84.2 (1.6)	77.3 (2.5)	81.8 (1.3)	76.5 (2.6)	81.4 (1.4)	88.3	63.8

し，かつ目標ドメインで高い汎化性能をもつテキスト分類器が得られた．

MHLE の汎用性を確認するため，データの確率分布が目標ドメインと元ドメインで類似する場合での性能評価も行った．表 2 に，目標ドメインと元ドメインのサブセットに分割せずに，単一のデータセットからラベルありデータとラベルなしデータ，テストデータを無作為に抽出して実験を行った結果を示す．WebLB では 2000 個のラベルなしデータを，SRAA と 20news では各 2500 個のラベルなしデータを分類器の学習に用いた．性能評価には，すべてのテストコレクションで各 1000 個のテストデータを用いた．表 2 より，MHLE と JESS-CM によって得られる分類精度はほぼ同等であり，これらの手法の分類精度は WebKB の一部の例外を除いたほとんどの場合で他の純粋な生成モデル，識別モデルの分類精度を上回った．以上の実験結果より，MHLE は目標ドメインと元ドメインのデータの確率分布が類似する場合でも高い汎化性能をもつ分類器を設計するのに有用であるとともに，ドメイン間のデータの分布の違いに頑健であることを確認した．

5. ま と め

自動分類の対象であるテストデータ集合と学習に用いるラベルありデータ集合の分布の違いに頑健な半教師あり学習法を提案した．提案法は，分類器を構成する識別・生成の両モデルの学習に，テストデータ集合と同じドメインから集めたラベルなしデータをラベルありデータと同時に用いることを特徴とする．3 つのテストコレクションを用いたテキスト分類実験により，テストデータ集合と異なる分布をもつラベルありデータ集合を用いて学習する問題設定では，ほとんどの場合で，従来の識別・生成モデルの統合に基づく半教師あり学習法 JESS-CM よりも提案法では高い汎化性能をもつ分類器を得られることを確認した．

参 考 文 献

1) Blitzer, J., Mcdonald, R. and Pereira, F.: Domain adaptation with structural correspondence learning, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pp.120–128 (2006).
2) Dhillon, I.S. and Modha, D.S.: Concept decompositions for large sparse text data using clustering, *Machine Learning*, Vol.42, pp.143–175 (2001).
3) Grandvalet, Y. and Bengio, Y.: Semi-supervised learning by entropy minimization, *Advances in Neural Information Processing Systems 17*, MIT Press, Cambridge, MA, pp.529–536 (2005).
4) Jiang, J. and Zhai, C.: Instance Weighting for Domain Adaptation in NLP, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, pp.264–271 (2007).
5) Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T.: Text classification from labeled and unlabeled documents using EM, *Machine Learning*, Vol.39, pp.103–134 (2000).
6) Pan, S.J. and Yang, Q.: A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering*, Vol.22, No.10, pp.1345–1359 (2010).
7) Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function, *Journal of Statistical Planning and Inference*, Vol.90, No.2, pp.227–244 (2000).
8) Sugiyama, M. and Müller, K.-R.: Input-dependent estimation of generalization error under covariate shift, *Statistics & Decisions*, Vol.23, No.4, pp.249–279 (2005).
9) Suzuki, J. and Isozaki, H.: Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data, *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL-08)*, pp.665–673 (2008).
10) Suzuki, J., Isozaki, H., Carreras, X. and Collins, M.: An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pp.551–560 (2009).
11) Wu, D., Lee, W.S., Ye, N. and Chieu, H.L.: Domain adaptive bootstrapping for named entity recognition, *Proceedings of 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pp.1523–1532 (2009).
12) Zadrozny, B.: Learning and evaluating classifiers under sample selection bias, *Proceedings of the 21st International Conference on Machine Learning (ICML 2004)*, pp.114–121 (2004).
13) Zhu, X.: Semi-supervised learning literature survey, Technical report, University of Wisconsin (2005).