情報処理学会論文誌　数理モデル化と応用　Vol. 4　No. 2　31–42 (Mar. 2011)

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]

本論文では，自動分類の対象となるテストデータ集合と分布が大きく異なるラベルありデータ集合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みの 1 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，本論文で扱う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テストデータ集合と分布が類似するラベルなしデータ集合をラベルありデータ集合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．また，提案法の学習アルゴリズムと条件付き確率モデルを単一の目的関数を用いて定式化する．3 つの代表的な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．また，ラベルなしデータの選択に基づく手法と提案法を組み合わせることの効果を実験的に確認した．

Robust Semi-supervised Learning for Labeled Data Selection Bias

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

This paper presents a robust semi-supervised learning method for designing good classifiers with a high generalization ability from a labeled dataset whose distribution differs largely from that of a target test dataset. 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 the task setting we consider in this paper. We expect the proposed method to solve the overfitting problem by utilizing an unlabeled dataset, whose distribution is similar to that of the target test dataset, with the labeled data set for both training of discriminative and generative models composing a classifier. We formulate the training algorithm and conditional probability model by defining a single objecive function. Our experimental results for text classification using three typical test collections confirmed that the classification performance obtained with the proposed method was better than that of JESS-CM in most cases of the task setting. We also confirmed experimentally the effect of combining the proposed method with an unlabeled data selecting approach.

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持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持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ことが多い．たとえば，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記事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持つラベルなしデータ集合を用いて汎化性能が高い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る．

本論文では，多値分類問題を対象として，上記のような状況で汎化性能が高い分類器を得るための学習法に焦点を当てる．多値分類は，特徴ベクトル $\boldsymbol{x} \in \mathcal{X}$ で表されるデータに対して，事前に定義された K 個のクラスラベルの集合 $\mathcal{Y} = \{1, \ldots, k, \ldots, K\}$ の中から 1 つのクラスラベル $y \in \mathcal{Y}$ を選択する問題である．$\mathcal{X}$ は特徴空間を表す．先行研究[20] にならって，分類対象となるテストデータ集合の領域を目標ドメイン（target domain）と呼び，目標ドメインのデータと関連性があり，分類器の学習に利用されるデータ集合の領域を元ドメイン（source domain）と呼ぶことにする．目標ドメインと元ドメインの間には以下の関係があると仮定する．

(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：$\mathcal{X}_t = \mathcal{X}_s$

†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

(2)　目標ドメインのデータの分布 $p_t(\boldsymbol{x})$ と元ドメインのデータの分布 $p_s(\boldsymbol{x})$ が大きく異なる：$p_t(\boldsymbol{x}) \neq p_s(\boldsymbol{x})$

(3)　目標ドメインのクラスラベル集合 $\mathcal{Y}_t$ と元ドメインのクラスラベル集合 $\mathcal{Y}_s$ は同一：$\mathcal{Y}_t = \mathcal{Y}_s$

(4)　データがクラスに属する条件付き確率の分布には目標ドメインと元ドメインの間で高い類似性がある：$P_t(y|\boldsymbol{x}) \simeq P_s(y|\boldsymbol{x})$

上に述べたニュース記事の分類問題の例では，目標ドメインのデータは新しい記事に相当し，元ドメインのデータは過去の記事に相当する．特徴空間 $\mathcal{X}$ を単語集合で与え，各ニュース記事の特徴ベクトル $\boldsymbol{x}$ を単語出現頻度で与える場合，(1) の仮定は同一の集合に含まれる単語を用いて過去の記事と新しい記事が作成されることを意味し，(2) の仮定は過去の記事と新しい記事で単語出現頻度に違いがあることを意味する．(3) の仮定は記事を分類するのに同じクラス（カテゴリ）の候補を用いることを意味し，(4) の仮定は記事の新しさによらず同じ内容（同じ単語出現頻度）の記事であれば同じクラス（カテゴリ）に属することを意味する．本研究では，目標ドメインのラベルありデータが得られない問題を想定し，元ドメインのみから集められたラベルありデータ集合と目標ドメインから集められたラベルなしデータ集合を用いて目標ドメインの新規データに適した分類器を設計することを課題とする．

近年，ラベルありデータとラベルなしデータから分類器を設計する問題に対し，生成モデルと識別モデルのハイブリッドに基づく半教師あり学習法が提案され，実問題での有用性が確認されてきた[1),9),10),15),24),25)]．一般に，ラベルありデータを用いて学習させる場合に識別モデルでは生成モデルよりも高い自動分類の精度が得られる[17)]一方，ラベルなしデータの分布を分析するのに生成モデルが有効である[21)]ことが知られている．ハイブリッド法では，両モデルの利点を活かすことでラベルありデータとラベルなしデータから効果的に分類器を学習させる．

本論文では，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異なる場合でも，テストデータ集合と分布が類似するラベルなしデータ集合を用いて高い汎化性能を持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き確率モデルを与える点で JESS-CM（Joint probability model Embedding style Semi-Supervised Conditional Model）法[24),25)]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みであるが，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．また，提案法では，学習アルゴリズムと分類器の条件付き確率モデルを同一の目的関数を用いて定式化する．提案法の基本的なアイデアは文献 11) で述べたが，本論文では，提案法の導出過程を詳述するとともに，ラベルなしデータの選択による手法と提案法を併用することによる効果を実験的に確認する．

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. 関 連 研 究

近年，目標ドメインと関連があるが何らかの相違がある元ドメインのラベルありデータを活用して目標ドメインでの自動分類の精度を向上させる転移学習[20)]の研究がさかんに行われている．文献 20) に，ドメイン間の相違がどの点にあるか，各ドメインで学習に利用できるラベルありデータがあるかないか，の 2 つの観点でこれまで研究されてきた転移学習の問題設定が分類されている．1 章で述べた課題は，標本選択バイアス（sample selection bias）[27)]，共変量シフト（covariate shift）[22)]，教師なしドメイン適応（unsupervised domain adaptation）[13)]の研究で扱われてきた問題設定と類似する．

従来研究では，元ドメインのラベルありデータと目標ドメインのラベルなしデータのみから目標ドメインのデータに適した分類器を設計する課題に対して，訓練データの重み付け，あるいは特徴空間の変換に基づく手法が提案されてきた[20)]．重み付けによる手法では，目標ドメインのラベルなし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[3),22),23),27)]．具体的には，ラベルなしデータ集合を用いてラベルありデータの重みを推定し，重み付けされたラベルありデータを従来の教師あり分類器の学習に適用する．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[14),26)]．特徴空間の変換に基づく手法では，データ $\boldsymbol{x}$ の特徴空間 $\mathcal{X}$ を，元ドメインのラベルありデータ集合の分布と目標ドメインのラベルなし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 $\mathcal{Z}$ 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ている．代表的な方法として，両分布間で関連性がある特徴量をもとに特徴空間を変換する手法が提案されている[2),4),19)]．

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[5),28)]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．これらの方法は目標ドメインのラベルなしデータ集合と元ドメインの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．そこで，本研究では，後者の立場で，ベースの分類器の性能を向上させるための半教師あり学習法を提案する．

3. 提　案　法

3.1 基本的な枠組

本論文では，目標ドメインのラベルなしデータ集合 $D_u = \{\boldsymbol{x}_m\}_{m=1}^{M}$ と元ドメインのラベルありデータ集合 $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ 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，ラベルあり・なしデータの両方を用いて学習させた識別モデル $P_d(y|\boldsymbol{x}; W)$ と生成モデル $p_g(\boldsymbol{x}, y; \Theta)$ の統合に基づいて分類器を設計する．また，分類器の条件付き確率モデルとパラメータ推定アルゴリズムを同一の目的関数を用いて定式化する．

提案法では，教師あり学習でよく用いられる MAP 推定を応用して，生成モデルを学習させる．仮に，ラベルなしデータ $\boldsymbol{x}_m$ のクラスラベル y_m が観測されるならば，MAP 推定に基づく以下の目的関数の最大化により生成モデル $p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y)$ のパラメータ $\Theta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K]$ の推定値を得ることができる．

$$G(\Theta) \equiv \sum_{n=1}^{N} \log p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n}) + \sum_{m=1}^{M} \sum_{k=1}^{K} I_{y_m}(k) \log p_g(\boldsymbol{x}_m, k; \boldsymbol{\theta}_k) + \log p(\Theta)$$

ただし，$I_{y_m}(k)$ は $I_{y_m}(k = y_m) = 1$，$I_{y_m}(k \neq y_m) = 0$ を満たす指示関数であり，$p(\Theta)$ は Θ の事前確率分布を表す．しかし，ラベルなしデータ $\boldsymbol{x}_m$ のクラスラベル y_m は観測されず，$I_{y_m}(k)$ の値は未知である．そこで，提案法では，ラベルなしデータ $\boldsymbol{x}_m$ のクラスラベルが k である確率 $P(k|\boldsymbol{x}_m)$ を導入し，$P(k|\boldsymbol{x}_m)$ による対数尤度の重み付き和に基づく以下の目的関数の最大化で生成モデルを学習させる．

$$\begin{aligned} J_g(\Theta) &\equiv \sum_{n=1}^{N} \log p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n}) + \sum_{m=1}^{M} \sum_{k'=1}^{K} P(k'|\boldsymbol{x}_m) \sum_{k=1}^{K} I_{k'}(k) \log p_g(\boldsymbol{x}_m, k; \boldsymbol{\theta}_k) \\ &\quad + \log p(\Theta) \\ &= \sum_{n=1}^{N} \log p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n}) + \sum_{m=1}^{M} \sum_{k=1}^{K} P(k|\boldsymbol{x}_m) \log p_g(\boldsymbol{x}_m, k; \boldsymbol{\theta}_k) + \log p(\Theta) \end{aligned} \tag{1}$$

ここで，ラベルなしデータのクラスラベル y_m が既知であると考え，$P(k = y_m|\boldsymbol{x}_m) = 1$，$P(k \neq y_m|\boldsymbol{x}_m) = 0$ を代入すると，式 (1) の右辺は

$$\sum_{n=1}^{N} \log p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n}) + \sum_{m=1}^{M} \log p_g(\boldsymbol{x}_m, y_m; \boldsymbol{\theta}_{y_m}) + \log p(\Theta)$$

となる．したがって，$J_g(\Theta)$ は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 $P(k|\boldsymbol{x}_m)$ を用いて単純に拡張したものであるといえる．

識別モデル $P_d(y|\boldsymbol{x}; W)$ の学習では，生成モデルの学習のために導入した $P(k|\boldsymbol{x}_m)$ を用いてパラメータ W の値を推定する．ラベルなしデータ $\boldsymbol{x}_m$ は生成モデルと識別モデルの違いに依存しないデータであるため，両モデルの学習に共通の $P(k|\boldsymbol{x}_m)$ を導入するのは妥当といえる．$P(k|\boldsymbol{x}_m)$ の値が既知であるとき，ラベルありデータの条件付き対数尤度の最大化と，$P(y|\boldsymbol{x}_m)$ と $P_d(y|\boldsymbol{x}_m; W)$ 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，ラベルあり・なしデータの両方に適合する識別モデルのパラメータ W の推定値を得ることができる．

$$J_d(W) \equiv \sum_{n=1}^{N} \log P_d(y_n|\boldsymbol{x}_n; W) - \sum_{m=1}^{M} \sum_{k=1}^{K} P(k|\boldsymbol{x}_m) \log \frac{P(k|\boldsymbol{x}_m)}{P_d(k|\boldsymbol{x}_m; W)} + \log p(W) \tag{2}$$

ただし，$p(W)$ は W の事前確率分布である．式 (2) の右辺で $P(k = y_m|\boldsymbol{x}_m) \to 1, P(k \neq y_m|\boldsymbol{x}_m) \to 0$ とすると，以下の式が得られる．

$$\sum_{n=1}^{N} \log P_d(y_n|\boldsymbol{x}_n; W) + \sum_{m=1}^{M} \log P_d(y_m|\boldsymbol{x}_m; W) + \log p(W)$$

したがって，$J_d(W)$ もまた識別モデルの教師あり学習でよく用いられる目的関数を単純に

拡張したものであるといえる．

しかし，生成・識別モデルの学習のために導入した上記の $P(k|\boldsymbol{x}_m)$ の値は未知であり，W，Θ と同様に推定する必要がある．提案法では，識別モデルとして推定される条件付き確率 $P_d(y|\boldsymbol{x}_m;W)$ ではなく，式 (1) と式 (2) で与えた $J_d(W)$ と $J_g(\Theta)$ を同時に最大化させる $P=[P(k|\boldsymbol{x}_m)]_{m,k}$ が，生成・識別両モデルをラベルなしデータによく適合させる条件付き確率を与えると考え，$J_d(W)$ と $J_g(\Theta)$ の線形結合で定義される以下の目的関数を最大化させる P をラベルなしデータの条件付き確率の推定値として求める．

$$J(W,\Theta,P) \equiv J_d(W) + \beta J_g(\Theta) \tag{3}$$

上式中の β（≥ 0）は，$J_d(W)$ と $J_g(\Theta)$ の統合の重みである．$\sum_{k=1}^{K} P(k|\boldsymbol{x}_m)=1, \forall m$ 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 の解

$$P(y|\boldsymbol{x}_m;W,\Theta,\beta) = \frac{P_d(y|\boldsymbol{x}_m;W)p_g(\boldsymbol{x}_m,y;\boldsymbol{\theta}_y)^\beta}{\sum_{k=1}^{K} P_d(k|\boldsymbol{x}_m;W)p_g(\boldsymbol{x}_m,k;\boldsymbol{\theta}_k)^\beta}, \forall m, \forall y \in \{1,\ldots,k,\ldots,K\} \tag{4}$$

が得られる（導出方法は付録 A.1 を参照）．すなわち，提案法では，式 (4) のように識別・生成モデルを重み β で統合して得られる $P(y|\boldsymbol{x}_m;W,\Theta,\beta)$ をラベルなしデータの条件付き確率 $P(y|\boldsymbol{x}_m)$ の推定値とする．また，新規データ $\boldsymbol{x}$ に対しても $P(y|\boldsymbol{x};W,\Theta,\beta)$ が最も良い条件付き確率の推定値を与えると考え，分類器の条件付き確率モデルとして用いる．

ラベルなしデータの条件付き確率 $P(k|\boldsymbol{x}_m)$ が与えられた状況下では，$J_g(\Theta)$ を最大化させる Θ の値と $J_d(W)$ を最大化させる W の値は，$J_g(\Theta)$ と $J_d(W)$ の（重み付き）和を最大化させる Θ と W の値と同一である．そこで，式 (3) の $J_d(W)$ と $J_g(\Theta)$ に含まれる $P(k|\boldsymbol{x}_m)$ に式 (4) で与えられる $P(k|\boldsymbol{x}_m;W,\Theta,\beta)$ を代入して得られる以下の目的関数の最大化により，W と Θ の推定値を求める．

$$\begin{aligned} J(W,\Theta) = {} & \log p(W) + \beta \log p(\Theta) + \sum_{n=1}^{N} \log P_d(y_n|\boldsymbol{x}_n;W)p_g(\boldsymbol{x}_n,y_n;\boldsymbol{\theta}_{y_n})^\beta \\ & + \sum_{m=1}^{M} \log \sum_{k=1}^{K} P_d(k|\boldsymbol{x}_m;W)p_g(\boldsymbol{x}_m,k;\boldsymbol{\theta}_k)^\beta \end{aligned} \tag{5}$$

3.2　パラメータ推定アルゴリズム

$J(W,\Theta)$ によるパラメータ推定では，β の値を設定し，EM アルゴリズム[7] のような繰返し計算を行うことで，初期値[⋆1]周辺での $\Psi=\{W,\Theta\}$ の局所最適解を得ることができる．繰返し計算の (t) ステップでの Ψ の推定値 $\Psi^{(t)}$ を用いて，Q 関数を以下のように定義する．

$$Q(\Psi,\Psi^{(t)}) \equiv q_d(W,\Psi^{(t)}) + \beta\, q_g(\Theta,\Psi^{(t)})$$

ただし

$$\begin{aligned} q_d(W,\Psi^{(t)}) = {} & \log p(W) + \sum_{n=1}^{N} \log P_d(y_n|\boldsymbol{x}_n;W) \\ & + \sum_{m=1}^{M}\sum_{k=1}^{K} P(k|\boldsymbol{x}_m;\Psi^{(t)},\beta) \log P_d(k|\boldsymbol{x}_m;W) \end{aligned} \tag{6}$$

$$\begin{aligned} q_g(\Theta,\Psi^{(t)}) = {} & \log p(\Theta) + \sum_{n=1}^{N} \log p_g(\boldsymbol{x}_n,y_n;\boldsymbol{\theta}_{y_n}) \\ & + \sum_{m=1}^{M}\sum_{k=1}^{K} P(k|\boldsymbol{x}_m;\Psi^{(t)},\beta) \log p_g(\boldsymbol{x}_m,k;\boldsymbol{\theta}_k) \end{aligned} \tag{7}$$

である．一般に，$\log b \leq b-1$ より，$J(\Psi)-J(\Psi^{(t)}) \geq Q(\Psi,\Psi^{(t)})-Q(\Psi^{(t)},\Psi^{(t)})$ の関係があるため，$(t+1)$ ステップの推定値を

$$\Psi^{(t+1)} = \arg\max_{\Psi} Q(\Psi,\Psi^{(t)}),\ t=0,1,2,\ldots \tag{8}$$

のように計算することで，$J(\Psi)$ を単調に増大させる Ψ の推定値を得ることができる．式 (8) の解 $\Psi^{(t+1)}=\{W^{(t+1)},\Theta^{(t+1)}\}$ は，以下のように $W^{(t+1)}$ と $\Theta^{(t+1)}$ を独立に計算することで得られる．

$$W^{(t+1)} = \arg\max_{W} q_d(W,\Psi^{(t)}) \tag{9}$$

$$\Theta^{(t+1)} = \arg\max_{\Theta} q_g(\Theta,\Psi^{(t)}) \tag{10}$$

以上のパラメータ推定アルゴリズムを**図 1** にまとめる．

3.3　従来法との相違

生成・識別モデルのハイブリッドに基づく従来の半教師あり学習法 JESS-CM（Joint probability model Embedding style Semi-Supervised Conditional Model）[24),25)] は，複数の自然言語処理タスクで最良の結果を達成した手法である．JESS-CM 法では，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に基づく線形の関数 $f(\boldsymbol{x},\boldsymbol{y};W)$

⋆1 本論文では，4.2 節で述べる方法で得られる値を初期値として用いる．

1. Input training set, $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$.
2. Set β, $\Psi^{(0)} = \{W^{(0)}, \Theta^{(0)}\}$, and $t \leftarrow 0$.
3. Perform the following until convergence.
   [E-step]
   - Set $Q(\Psi, \Psi^{(t)})$ by computing $P(y|\boldsymbol{x}_m; \Psi^{(t)}, \beta)$, $\forall m$ and $\forall y$ using Eq.(4).

   [M-step]
   - Compute $W^{(t+1)}$ using Eq. (9).
   - Compute $\Theta^{(t+1)}$ using Eq. (10).
   - $t \leftarrow t + 1$.
4. Output parameter estimates $\hat{\Psi} = \{\hat{W}, \hat{\Theta}\}$ after setting as $\{\hat{W}, \hat{\Theta}\} \leftarrow \{W^{(t)}, \Theta^{(t)}\}$.

図 1　パラメータ推定アルゴリズム
Fig. 1　Algorithm for parameter estimation.

を用いて，以下のように分類器の条件付き確率モデルを定義する[⋆1]．

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

W と Θ はそれぞれ関数 $f(\boldsymbol{x}, \boldsymbol{y}; W)$ 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．

$$\mathcal{L}^1(W, \beta|\Theta) \equiv \sum_{n=1}^N \log P(\boldsymbol{y}_n|\boldsymbol{x}_n; W, \beta, \Theta) + \log p(W, \beta)$$

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

W と β の推定では Θ を固定したうえで $\mathcal{L}^1(W, \beta|\Theta)$ 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 を固定したうえで $\mathcal{L}^2(\Theta|W, \beta)$ 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．

⋆1 文献 24) 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，関数 $f(\boldsymbol{x}, \boldsymbol{y}; W)$ 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．

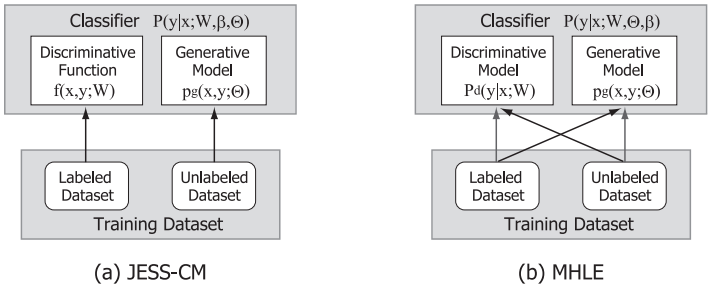

図 2　JESS-CM 法と MHLE 法の概要
Fig. 2　Outline of JESS-CM and MHLE methods.

図 2 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，関数 $f(\boldsymbol{x}, \boldsymbol{y}; W)$ のパラメータ値を推定するのにラベルありデータのみを用いた識別学習を行う．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うえに，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 法のように関数 $f(\boldsymbol{x}, \boldsymbol{y}; W)$ 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し，目標ドメインでの分類精度が向上することを期待する．4 章に示すように，本研究のタスク設定では，提案法により目標ドメインでの分類精度が向上することを実験的に確認した．

3.4　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 $p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y)$，$\forall y$ 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 x_i（≥ 0）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し，$\boldsymbol{a}^T$ は $\boldsymbol{a}$ の転置ベクトルを表す．NB モデルでは，クラス y での $\boldsymbol{x}$ の確率分布 $p(\boldsymbol{x}|y)$ が多項分布に従うと仮定し，

$$p_g(\boldsymbol{x}, y; \boldsymbol{\theta}_y) \propto \pi_y \prod_{i=1}^V (\theta_{yi})^{x_i}$$

のように $\boldsymbol{x}$ と y の同時確率密度をモデル化する．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，$\boldsymbol{\theta}_y$ の L1 ノルムは $||\boldsymbol{\theta}_y||_1 = \sum_{i=1}^{V} |\theta_{yi}| = 1$ 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．また，式 (7) 中の $p(\Theta)$ には，ディリクレ事前確率分布 $p(\Theta) \propto \prod_{k=1}^{K} \prod_{i=1}^{V} (\theta_{ki})^{\eta-1}$ を適用した．$\xi = \eta - 1$（> 0）はハイパーパラメータである．以上の設定の下では，式 (10) を満たす $\Theta^{(t+1)} = [\boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_k^{(t+1)}, \ldots, \boldsymbol{\theta}_K^{(t+1)}]$ の解を以下の式で計算できる．

$$\boldsymbol{\theta}_y^{(t+1)} = \frac{\sum_{n=1}^{N} I_{y_n}(y)\boldsymbol{x}_n + \sum_{m=1}^{M} P(y|\boldsymbol{x}_m; \Psi^{(t)}, \beta)\, \boldsymbol{x}_m + \xi\mathbf{1}}{\left|\left|\sum_{n=1}^{N} I_{y_n}(y)\boldsymbol{x}_n + \sum_{m=1}^{M} P(y|\boldsymbol{x}_m; \Psi^{(t)}, \beta)\, \boldsymbol{x}_m + \xi\mathbf{1}\right|\right|_1}, \ \forall y$$

$I_{y_n}(y)$ は指示関数であり，$y = y_n$ のとき $I_{y_n}(y) = 1$ で，それ以外のとき $I_{y_n}(y) = 0$ である．$\mathbf{1}$ は各要素の値が 1 の V 次元ベクトルを表す．

識別モデル $P_d(y|\boldsymbol{x}; W)$ には，以下に示す多項ロジスティック回帰（MLR）モデルを適用する．

$$P_d(y|\boldsymbol{x}; W) = \frac{\exp\left(\boldsymbol{w}_y^T \boldsymbol{x}\right)}{\sum_{k=1}^{K} \exp\left(\boldsymbol{w}_k^T \boldsymbol{x}\right)}$$

$W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k, \ldots, \boldsymbol{w}_K]$ 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．このとき，式 (9) 中の $g_d(W, \Psi^{(t)})$ の勾配 $\partial g_d(W, \Psi^{(t)})/\partial \boldsymbol{w}_y$ を

$$\frac{\partial}{\partial \boldsymbol{w}_y} g_d(W, \Psi^{(t)}) = -\frac{\boldsymbol{w}_y}{\sigma^2} + \sum_{n=1}^{N} \left\{ I_{y_n}(y) - P_d(y|\boldsymbol{x}_n; W) \right\} \boldsymbol{x}_n + \sum_{m=1}^{M} \left\{ P(y|\boldsymbol{x}_m; \Psi^{(t)}, \beta) - P_d(y|\boldsymbol{x}_m; W) \right\} \boldsymbol{x}_m, \ \forall y$$

で計算できる．そこで，準ニュートン法の一種である L-BFGS アルゴリズム[16] を用いて式 (9) の解を算出した．$g_d(W, \Psi^{(t)})$ は W に関する上に凸な関数であるため，$W^{(t+1)}$ の大域的最適解を推定できる．以上より，提案法に NB モデルと MLR モデルを適用したテキスト分類器では，繰返し計算の各ステップで式 (8) の大域的最適解を推定できる．

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*[⋆1]，*SRAA*[⋆2]，*20 newsgroups* (*20news*)[⋆3] を用いて評価実験を行った．

WebKB 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 18) の設定に従い，4 つのカテゴリ *course*，*faculty*，*project*，*student* に含まれる 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 18,525 語彙を用いて web ページの特徴ベクトルを与えた．

SRAA は 4 つのカテゴリ（*sim-auto*，*sim-aviation*，*real-auto*，*real-aviation*）に属する 73,218 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．1 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分布と元ドメインに属するデータの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[8] 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 61,526 語彙を用いて UseNet 記事の特徴ベクトルを与えた．

20news 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 4,881 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 19,383 語彙を選択した．

評価実験では，各テストコレクションのデータセットから排他的に選択したテストデータ集合と訓練データ集合を用いて，提案法と比較手法に基づく各分類器の性能を評価した．まず，テストコレクションごとに，800 個の目標ドメインのテストデータと 200 個の元ドメイ

⋆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

ンのテストデータを無作為に抽出した．次に，テストデータとは排他的に，元ドメインのサブセットに含まれる残りのデータからラベルありデータを無作為に抽出し，目標ドメインのサブセットに含まれる残りのデータからラベルなしデータを無作為に抽出することで訓練データ集合を作成した．SRAA と 20news では 2,500 個のラベルなしデータを，WebKB では 2,000 個のラベルなしデータを分類器の学習に利用した．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を評価した．

4.2 実験設定

提案法 MHLE と従来の 4 つの半教師あり学習法 JESS-CM，NB/EM-λ，MLR/MER TSVM 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した．実験では，ラベルありデータのみを用いて学習する 3 つの教師あり学習法 NB，MLR，SVM で得られる分類精度とも比較した．

MHLE に基づくテキスト分類器の学習では，3.4 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パラメータ 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．W の初期値 $W^{(0)}$ にはラベルありデータのみを用いて MLR モデルを学習させたときに得られる W の推定値を設定し，$\Theta^{(0)}$ の各クラスの要素を $\boldsymbol{\theta}_y^{(0)} = \bar{\boldsymbol{\theta}}$，$\forall y$ のように設定した．$\bar{\boldsymbol{\theta}}$ は，ラベルありデータとラベルなしデータの単語頻度の平均値をもとに見積ったデータセット全体の単語の確率ベクトルである．MHLE では，3.4 節で述べた NB モデルの π_y を $\pi_y = 1/K$，$\forall y$ に設定した．

JESS-CM では，関数 $f(\boldsymbol{x}, \boldsymbol{y}; W)$ と生成モデルにそれぞれ MLR モデルの識別関数 $\boldsymbol{w}_y^T \boldsymbol{x}$ 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[12] 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 $\lambda \in \{1 \times 10^n, 2 \times 10^n, 5 \times 10^n\}_{n=-5}^{0}$ 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[18] 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 $\lambda \in \{0.01, 0.02, 0.05\} \cup \{n \times 0.1\}_{n=1}^{9} \cup \{1, 2, 5\}$ 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 $\xi \in \{1 \times 10^n, 2 \times 10^n, 5 \times 10^n\}_{n=-3}^{-1} \cup \{1\}$ の候補値の中から選択した．

TSVM と SVM には，Universvm[6] ⋆1 を使用した．TSVM と SVM は 2 値分類器であるが，Universvm の実装では one-against-rest 法を用いて多値分類に拡張している．本実験では，線形カーネルを用い，文献 6) の実験結果をもとに TSVM でラベルなしデータに用いる Symmetric Ramp loss のパラメータ値を $s \in \{-0.5, -0.4, -0.3, -0.2\}$ の候補値の中から選択した．ラベルありデータとラベルなしデータに対するコストパラメータの値をそれぞれ $C \in \{10^n\}_{n=-2}^{5}$ と $C^* \in \{C \times 10^n\}_{n=-3}^{2}$ の候補値の中から選択した．

上記のように，提案法と比較手法に基づく各分類器の学習では，事前にハイパーパラメータと重みパラメータの値を設定する必要がある．公正かつ実問題の状況を想定した性能評価を行うため，ハイパー・重みパラメータの設定には，テストデータをまったく用いず，4.1 節で述べた訓練用のラベルありデータを用いた．ただし，分類器の学習とハイパー・重みパラメータの設定に同一のラベルありデータ集合を利用することで過学習の問題を引き起こす危険性がある．そこで，文献 18) と同様に，ラベルありデータの交差検定（cross-validation）に基づいてハイパー・重みパラメータの値を設定した．本実験では，ラベルありデータ集合を排他的に 5 つのサブ集合に分けて 5 分割交差検定を行った．ハイパー・重みパラメータの各候補値の組合せに対して，4 つのラベルありデータのサブ集合とラベルなしデータ集合を用いて分類器を学習させ，その分類器によって得られる残りのサブ集合のラベルありデータの分類精度を調べる．訓練データから除外するラベルありデータのサブ集合を入れ換えてこの実験を 5 回繰り返し，候補値の組合せごとに分類精度の平均値を求める．最高の平均値を与える候補値の組合せを選択することでハイパー・重みパラメータの値を設定した．

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で与えた．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，バイナリ特徴量を用いて特徴ベクトルを与えた．NB/EM-λ と NB を除く 6 つの分類器では，特徴ベクトルの大きさの分散による悪影響を抑えるため，$||\boldsymbol{x}||_1 = \sum_{i=1}^{V} |x_i| = 1$ のように L1 ノルムで正規化された特徴ベクトルを適用した．

4.3 実験結果と考察

表 1 に 8 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，**表 2** (a) にラベルありデータ集合とラベルなしデー

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

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 (%)：(a) 目標ドメインのテストデータの分類精度，(b) 元ドメインのテストデータの分類精度

Table 1　Classification accuracies (%) obtained by using N labeled data with a distribution different from that of target domain: (a) classification accuracies of target domain test data, and (b) classification accuracies of source domain test data.

	N	MHLE	JESS-CM	NB/EM-λ	MLR/MER	TSVM	NB	MLR	SVM
(i)	20	**69.4** (6.5)	50.8 (9.3)	65.9 (3.4)	46.6 (10.3)	59.9 (13.7)	60.5 (5.4)	46.9 (10.2)	52.9 (9.9)
	100	72.3 (5.6)	66.8 (5.1)	69.7 (5.4)	63.2 (4.8)	**78.8** (4.5)	69.2 (4.2)	63.3 (4.7)	69.3 (4.8)
	500	**88.1** (1.4)	82.5 (2.4)	73.0 (3.2)	81.2 (1.6)	86.5 (1.2)	72.8 (2.7)	81.3 (2.0)	82.8 (1.9)
(ii)	50	44.2 (5.5)	**46.1** (4.3)	36.1 (5.5)	38.8 (4.2)	35.0 (5.0)	34.2 (4.5)	38.0 (3.9)	37.1 (4.5)
	200	**48.0** (3.4)	47.4 (4.6)	34.6 (6.0)	38.1 (6.7)	44.1 (1.3)	40.0 (3.6)	39.7 (3.6)	39.3 (3.3)
	1000	**50.0** (3.1)	48.7 (2.9)	32.5 (4.0)	44.8 (3.8)	47.8 (1.3)	45.3 (3.3)	45.3 (2.4)	44.7 (2.5)
(iii)	50	**42.4** (13.0)	41.6 (12.2)	31.1 (15.2)	17.4 (5.6)	31.5 (5.4)	19.8 (5.9)	22.4 (3.2)	21.7 (4.0)
	200	**49.7** (13.6)	48.4 (14.7)	36.3 (14.7)	23.0 (6.9)	40.7 (4.5)	25.8 (8.3)	23.3 (6.2)	27.0 (6.2)
	1000	**71.6** (10.3)	52.8 (7.2)	49.9 (17.4)	37.3 (5.1)	46.7 (5.3)	36.8 (9.0)	37.6 (5.2)	38.8 (5.5)

(i) WebKB, (ii) SRAA, (iii) 20news

(a)

	N	MHLE	JESS-CM	NB/EM-λ	MLR/MER	TSVM	NB	MLR	SVM
(i)	20	71.4 (7.3)	69.5 (6.4)	**77.0** (2.4)	66.2 (7.2)	66.0 (16.4)	72.0 (3.6)	66.2 (6.6)	69.8 (5.3)
	100	83.0 (1.5)	83.8 (3.6)	80.3 (2.5)	83.5 (2.4)	**86.2** (1.7)	81.5 (2.8)	83.5 (2.4)	85.2 (2.4)
	500	90.8 (2.4)	91.4 (2.3)	82.9 (3.5)	91.6 (2.4)	**93.3** (1.9)	83.7 (3.1)	91.5 (2.3)	92.0 (2.6)
(ii)	50	67.1 (6.8)	66.0 (4.5)	**69.6** (2.5)	62.9 (3.8)	64.8 (4.6)	64.7 (2.8)	63.3 (3.6)	63.6 (3.9)
	200	79.0 (2.9)	78.3 (4.0)	**81.6** (2.3)	75.5 (2.6)	76.7 (3.1)	78.0 (2.8)	75.5 (2.6)	76.2 (2.4)
	1000	**88.3** (3.2)	85.8 (2.3)	86.5 (2.7)	85.2 (2.3)	84.9 (2.3)	86.6 (2.7)	84.6 (2.4)	84.9 (2.1)
(iii)	50	62.5 (6.4)	**64.5** (7.1)	56.8 (7.0)	64.0 (5.6)	64.0 (7.2)	61.5 (4.7)	66.1 (4.6)	64.9 (5.4)
	200	**76.8** (3.0)	76.0 (3.4)	71.9 (6.8)	75.3 (3.6)	75.6 (3.5)	71.8 (7.3)	75.1 (4.1)	75.6 (3.6)
	1000	**86.8** (3.7)	85.8 (3.5)	82.2 (5.5)	84.9 (3.1)	85.5 (3.3)	83.2 (4.0)	84.9 (2.9)	85.6 (3.0)

(i) WebKB, (ii) SRAA, (iii) 20news

(b)

タ集合の分布の重なり具合を示す．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_a は少なくともいづれか一方の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 (%) を表す．r_l，r_u，r_a の値が小さいほど分布間の重なりが小さいことを示す．

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，識別モデルの各アプローチに基づく MLR/MER，NB/EM-λ，TSVM よりも MHLE の方がほとんどの場合で高かった．WebKB で $N = 100$ の場合のみ，TSVM の分類精度が MHLE よりも高かった．この場合，SVM の分類精度は MLR よりも高かった．本実験では，MHLE では MLR モデルと NB モデルを組み合わせて分類器を設計した．それゆえ，MHLE の分類精度が TSVM を下回った要因は MLR の汎化性能の低さにあったと考えられる．

表 2　ラベルありデータ集合とラベルなしデータ集合の間の語彙の重なり度合い (%)：(a) 異なるドメイン設定，(b) 単一ドメイン設定

Table 2　Word overlap ratio (%) between labeled and unlabeled datasets in (a) different domain setting and (b) single domain setting.

dataset	N	r_l	r_u	r_a
WebKB	20	94.1	7.5	7.5
	100	90.6	21.6	21.2
	500	85.5	46.2	42.8
SRAA	50	79.8	8.4	8.2
	200	72.4	20.3	18.9
	1000	57.8	42.3	32.3
20news	50	83.2	9.8	9.5
	200	78.2	22.4	20.9
	1000	74.9	45.8	39.4

(a)

dataset	N	r_l	r_u	r_a
WebKB	20	96.9	7.3	7.3
	100	94.2	21.6	21.3
	500	89.9	51.0	48.2
SRAA	50	90.3	7.7	7.6
	200	84.6	20.3	19.6
	1000	72.4	45.3	38.6
20news	50	95.4	10.5	10.4
	200	92.4	27.1	26.4
	1000	88.3	63.8	58.8

(b)

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し，かつ目標ドメインで高い汎化性能を持つテキスト分類器が得られた．

MHLE の学習法の効果を確認するため，パラメータ推定時に行う繰返し計算の各ステップで得られる分類精度を調べた．**図 3** は，(t) ステップ時のパラメータ値を用いて得られる MHLE と JESS-CM によるテストデータの分類精度の例を示す．図中の $CA(C)$ は MHLE または JESS-CM による目標ドメインのテストデータの分類精度を示す．また，$CA(D)$ と $CA(G)$ は分類器を構成する $P_d(y|\boldsymbol{x}; W)$ と $p_g(\boldsymbol{x}, y; \Theta)$ をそれぞれ用いて得られる目標ドメインのテストデータの分類精度を示す．

図 3 (a)，(c) の例では，20news と WebKB の両方で，パラメータ推定の繰返し計算ごとに MHLE の識別モデルによる目標ドメインの分類精度 $CA(D)$ が徐々に向上し，MHLE 分類器自体による分類精度 $CA(C)$ も向上するのが観察された．一方，JESS-CM の分類精度

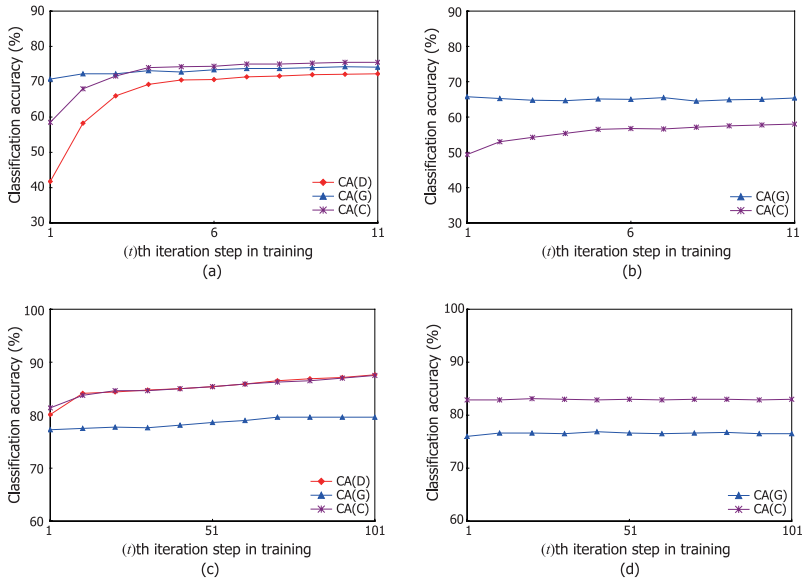

図 3　パラメータ推定の (t) ステップでの分類精度の例：(a) MHLE（20news，$N = 1000$，$M = 2500$），(b) JESS-CM（20news，$N = 1000$，$M = 2500$），(c) MHLE（WebKB，$N = 500$，$M = 2000$），(d) JESS-CM（WebKB，$N = 500$，$M = 2000$）

Fig. 3　Examples of the classification accuracies at the (t)th iteration step in training (a) an MHLE classifier for 20news ($N = 1000$, $M = 2500$), (b) a JESS-CM classifier for 20news ($N = 1000$, $M = 2500$), (c) an MHLE classifier for WebKB ($N = 500$, $M = 2000$), and (d) a JESS-CM classifier for WebKB ($N = 500$, $M = 2000$).

$CA(C)$ は，図 3 (b) の 20news の例では上昇する傾向がみられたが，図 3 (d) の WebKB の例では上昇する傾向がみられなかった．すなわち，MHLE では JESS-CM よりも繰返し計算の効果がみられた．MHLE では，繰返し計算の各ステップで，前ステップで得られたパラメータ推定値をもとに目標ドメインに属するラベルなしデータの条件付き確率 $P(y|\boldsymbol{x})$ を予測し，識別モデルの学習にラベルなしデータを利用する．この各ステップでの学習により，識別モデルをラベルなしデータに徐々に適合させる．それに対して，JESS-CM では，各ステップで元ドメインのラベルありデータのみを用いて分類器を識別学習させる．図 3 の例より，ラベルなしデータを用いた識別モデルの学習が目標ドメインにおける MHLE の汎化性能を向上させるのに有効であったと考えられる．

表 3　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 3　Classification accuracies (%) obtained by using N labeled data in a single domain setting.

	N	MHLE	JESS-CM	NB/EM-λ	MLR/MER	TSVM	NB	MLR	SVM
(i)	20	70.3 (4.7)	62.3 (7.7)	**73.2** (4.8)	59.2 (7.5)	64.9 (5.0)	63.6 (4.0)	59.4 (7.4)	61.7 (7.0)
	100	81.3 (1.6)	**82.0** (1.6)	77.8 (2.4)	80.1 (1.5)	81.0 (3.4)	76.0 (1.9)	80.0 (1.5)	81.2 (1.2)
	500	89.5 (1.1)	**91.0** (0.9)	83.3 (1.1)	**91.0** (1.7)	89.4 (1.2)	83.0 (1.2)	90.4 (1.3)	90.2 (1.1)
(ii)	50	**58.0** (5.0)	57.1 (5.1)	56.6 (3.8)	49.0 (1.5)	47.8 (5.4)	49.4 (2.2)	48.9 (1.3)	49.4 (1.6)
	200	64.9 (1.1)	**65.8** (1.3)	61.5 (2.2)	57.2 (1.7)	58.5 (1.3)	59.4 (1.2)	57.2 (1.7)	57.6 (2.1)
	1000	71.3 (1.2)	**71.9** (1.2)	70.8 (1.7)	68.5 (1.0)	67.6 (1.2)	70.1 (1.6)	68.3 (0.9)	67.6 (1.3)
(iii)	50	71.2 (2.6)	**72.7** (1.5)	60.5 (6.0)	51.4 (5.7)	63.2 (6.7)	48.6 (4.7)	51.9 (4.6)	50.5 (5.4)
	200	78.0 (1.3)	**78.3** (1.4)	67.1 (4.8)	72.3 (1.8)	74.3 (1.5)	64.6 (3.8)	69.5 (2.0)	68.7 (2.3)
	1000	**84.7** (1.5)	84.2 (1.6)	77.3 (2.5)	81.8 (1.3)	82.4 (1.7)	76.5 (2.6)	81.4 (1.4)	80.6 (1.8)

(i) WebKB, (ii) SRAA, (iii) 20news

MHLE の汎用性を確認するため，データの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 3** に，各テストコレクションのデータをサブセットに分けずに，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 2,500 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 1,000 個のテストデータを用いた．表 2 (b) に，この実験のために抽出したラベルありデータ集合とラベルなしデータ集合の分布の重なり度合いを示す．表 3 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．MHLE では，分類器を構成する生成・識別両モデルの学習に目標ドメインのラベルなしデータを利用することで，目標ドメインのデータに適した分類器が得られることを期待する．このため，MHLE では，目標ドメインのラベルなし・テストデータと元ドメインのラベルありデータの分布が類似する場合でも，目標ドメインのラベルなしデータに分類器を適合させることで高いテストデータの分類精度が得られたと考えられる．MHLE と JESS-CM でほぼ同等の分類精度が得られたのは，目標・元ドメインのデータの分布差がほとんどない場合には元ドメインよりも目標ドメインに分類器を適合させることが不可欠ではなかったためと考えられる．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分布が類似する場合でも高い汎化性能を持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．

4.4　データの重み付けとの併用による効果

2 章で述べたように，目標ドメインと元ドメインでデータの分布が異なる分類問題に対し，従来のアプローチでは教師あり学習や半教師あり学習に基づく分類器をベースの分類器として利用する．ベースの分類器として提案法を利用することの効果を確認するため，各々を 1 または 0 に重み付けしたラベルなしデータを用いて MHLE 分類器を学習させた場合の分類精度を調べた．

文献 26) では，学習に用いる目標ドメインのラベルなしデータを選択する手法を提案している．この手法では，各データに含まれる情報にはドメイン独立な情報とドメイン従属な情報があり，各データをドメイン独立な情報のみを含むデータと，ドメイン従属な情報のみを含むデータ，ドメイン独立・従属の両方の情報を含むデータの 3 つのタイプに分けて考える．ドメイン独立な情報は目標ドメインと元ドメインの両方のデータに含まれる情報であり，ドメイン従属な情報はいずれか一方のドメインのデータのみに含まれる情報である．文献 26) では，ドメイン独立・従属の両方の情報を含むラベルなしデータを学習に利用し，ドメイン従属な情報のみを含むラベルなしデータを学習用データから除外することで分類精度が向上することを確認している．そこで，このドメイン独立・従属の情報の考え方をもとに，本論文では，含まれる語彙の t_r%以上が元ドメインのラベルありデータのいずれかに含まれるラベルなしデータを選択して半教師あり学習を行った場合の分類精度を調べた．$t_r = 0$ の場合はすべてのラベルなしデータを学習に用い，$t_r = 100$ の場合はドメイン従属の情報をまったく含まないラベルなしデータのみを学習に用いることを意味する．

実験には，表 2 (a) でラベルありデータとラベルなしデータの分布の重なりが小さい傾向がみられた SRAA と 20news を用いた．**表 4** に，$N = 1000$ の場合に t_r の値を 0，75，80，85，90 に設定した場合の目標ドメインのテストデータの平均分類精度を示す．表中の $\bar{M}_s$ は実際に学習に用いたラベルなしデータの個数の平均値を示す．表 4 より，SRAA と 20news の両方で，MHLE による分類精度がその他の手法よりも高かった．20news では，JESS-CM 以外の手法で，$t_r = 75$ に設定した場合に高い分類精度が得られる傾向があった．また，SRAA では，HMLE で $t_r = 85$ に設定した場合に最も高い分類精度が得られた．

学習に用いるラベルなしデータを選択することで分類精度が向上した要因を確認するため，SRAA では $t_r = 85$，20news では $t_r = 75$ に設定したときに選択されたラベルなしデータ（SUD）の分類精度と除外されたラベルなしデータ（RUD）の分類精度を調べた．**表 5** 中の MHLE-A はすべてのラベルなしデータを MHLE の学習に用いた場合（$t_r = 0$）に得られる SUD と RUD の分類精度を，MHLE-S は SUD のみを学習に用いた場合の SUD と RUD の分類精度を表す．表 5 より，MHLE-A で得られる RUD の分類精度が SUD の分類精度よりも低かった．この結果から，MHLE-A では，クラスラベルを誤って予測した多くの RUD をパラメータ学習の繰返し計算に用いることで高い汎化性能を得られなかったと考えられる．一方，MHLE-S では，RUD を学習に用いずにそれらの影響を除外することで結果的に SUD と RUD の両ラベルなしデータの予測精度が向上した．以上より，MHLE-S では，クラスラベルを誤って予測する危険性の高いラベルなしデータを訓練データから除外することでテストデータの分類精度が向上したと考えられる．

表 4　1000 個のラベルありデータと選択した M_s 個のラベルなしデータを用いて得られる目標ドメインのテストデータの分類精度 (%)
Table 4　Classification accuracies (%) of target domain test data obtained by using 1000 labeled data and M_s selected unlabeled data.

dataset	t_r	$\bar{M}_s$	MHLE	JESS-CM	MLR/MER	NB/EM-λ	TSVM
SRAA	0	2500	50.0 (3.1)	48.7 (2.9)	44.8 (3.8)	32.5 (4.0)	47.8 (1.3)
	75	2191	50.2 (2.2)	48.4 (3.8)	45.2 (3.3)	33.8 (4.0)	48.1 (1.4)
	80	1792	50.0 (2.9)	47.6 (2.2)	44.9 (3.1)	33.3 (4.1)	48.4 (1.2)
	85	1140	**51.7** (2.6)	47.3 (4.9)	43.1 (4.8)	35.6 (4.8)	49.1 (1.7)
	90	536	49.8 (4.7)	46.2 (3.7)	43.6 (3.6)	38.2 (3.1)	48.2 (2.2)
20news	0	2500	71.6 (10.3)	52.8 (7.2)	37.3 (5.1)	49.9 (17.4)	46.7 (5.3)
	75	2095	**74.3** (3.9)	47.4 (7.4)	37.8 (5.2)	50.0 (16.2)	46.9 (5.5)
	80	1752	70.5 (10.4)	47.0 (6.8)	38.0 (4.8)	46.7 (17.3)	45.8 (6.5)
	85	1229	69.1 (6.0)	43.5 (6.6)	37.7 (5.3)	42.5 (14.9)	46.0 (6.1)
	90	690	60.8 (6.5)	42.4 (6.2)	37.5 (5.4)	39.2 (11.6)	43.7 (5.9)

表 5　MHLE で得られるラベルなしデータの分類精度 (%)
Table 5　Classification accuracies (%) of unlabeled samples obtained with MHLE.

dataset	data type	MHLE-A	MHLE-S
SRAA	SUD	52.6 (2.3)	54.2 (3.5)
	RUD	44.6 (1.7)	47.4 (2.1)
20news	SUD	68.6 (9.6)	70.9 (5.7)
	RUD	67.2 (11.4)	75.0 (4.6)

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・生成の両モデルを，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

と同時に用いて学習させ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持つラベルありデータ集合を用いて分類器を学習させる問題設定では，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持つ分類器が得られることを確認した．また，学習に用いるラベルなしデータを選択した場合に提案法で得られる分類精度を確認することで，従来のデータの重み付けに基づくアプローチに提案法を適用することの有用性を示した．今後の課題は，画像データや，シーケンスデータ，木構造を持つデータなど，異なる識別・生成モデルを用いる問題に本手法を適用することである．

参 考 文 献

1) Agarwal, A. and Daumé III, H.: Exponential family hybrid semi-supervised learning, *Proc. 21st International Joint Conference on Artifical Intelligence* (*IJCAI-09*), pp.974–979 (2009).
2) Ando, R.K. and Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data, *The Journal of Machine Learning Research*, Vol.6, pp.1817–1853 (2005).
3) Bickel, S., Brückner, M. and Scheffer, T.: Discriminative learning for differing training and test distributions, *Proc. 24th International Conference on Machine Learning* (*ICML 2007*), pp.81–88 (2007).
4) Blitzer, J., McDonald, R. and Pereira, F.: Domain adaptation with structural correspondence learning, *Proc. 2006 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2006*), pp.120–128 (2006).
5) Chapelle, O., Schölkopf, B. and Zien, A. (Eds.): *Semi-supervised Learning*, MIT Press, Cambridge, MA (2006).
6) Collobert, R., Sinz, F., Weston, J. and Bottou, L.: Large scale transductive SVMs, *Journal of Machine Learning Research*, Vol.7, pp.1687–1712 (2006).
7) 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).
8) Dhillon, I.S. and Modha, D.S.: Concept decompositions for large sparse text data using clustering, *Machine Learning*, Vol.42, pp.143–175 (2001).
9) 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).
10) Fujino, A., Ueda, N. and Saito, K.: A hybrid generative/discriminative approach to semi-supervised classifier design, *Proc. 20th National Conference on Artificial Intelligence* (*AAAI-05*), pp.764–769 (2005).
11) Fujino, A., Ueda, N. and Nagata, M.: A robust semi-supervised classification method for transfer learning, *Proc. 19th ACM International Conference on Information and Knowledge Management* (*CIKM'10*), pp.379–388 (2010).
12) 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).
13) Jiang, J.: A literature survey on domain adaptation of statistical classifiers (2007). http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/
14) Jiang, J. and Zhai, C.: Instance weighting for domain adaptation in NLP, *Proc. 45th Annual Meeting of the Association of Computational Linguistics* (*ACL-07*), pp.264–271 (2007).
15) Lasserre, J.A., Bishop, C.M. and Minka, T.P.: Principled hybrids of generative and discriminative models, *Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR'06*), pp.87–94 (2006).
16) 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).
17) Ng, A.Y. and Jordan, M.I.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, *Advances in Neural Information Processing Systems 14*, pp.841–848, MIT Press, Cambridge, MA (2002).
18) 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).
19) Pan, S.J., Tsang, I.W., Kwok, J.T. and Yang, Q.: Domain adaptation via transfer component analysis, *Proc. 21st International Joint Conference on Artifical Intelligence* (*IJCAI-09*), pp.1187–1192 (2009).
20) Pan, S.J. and Yang, Q.: A survey on transfer learning, *IEEE Trans. Knowledge and Data Engineering*, Vol.22, No.10, pp.1345–1359 (2010).
21) Seeger, M.: Learning with labeled and unlabeled data, Technical report, University of Edinburgh (2001).
22) 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).
23) 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).
24) Suzuki, J. and Isozaki, H.: Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data, *Proc. 46th Annual Meeting of the Association*

of Computational Linguistics (*ACL-08*), pp.665–673 (2008).
25) Suzuki, J., Isozaki, H., Carreras, X. and Collins, M.: An empirical study of semi-supervised structured conditional models for dependency parsing, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2009*), pp.551–560 (2009).
26) Wu, D., Lee, W.S., Ye, N. and Chieu, H.L.: Domain adaptive bootstrapping for named entity recognition, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2009*), pp.1523–1532 (2009).
27) Zadrozny, B.: Learning and evaluating classifiers under sample selection bias, *Proc. 21st International Conference on Machine Learning* (*ICML 2004*), pp.114–121 (2004).
28) Zhu, X.: Semi-supervised learning literature survey, Technical report, University of Wisconsin (2005).

付　　録

A.1　条件付き確率の導出

式 (3) の目的関数を最大化させる $P = [P(k|\boldsymbol{x}_m)]_{m,k}$ の解をラグランジュ未定乗数法を用いて導出する手順を述べる．各々のラベルなしデータ $\boldsymbol{x}_m$ の制約条件 $\sum_{k=1}^{K} P(k|\boldsymbol{x}_m) = 1$ に対してラグランジュ乗数 λ_m を導入すると，ラグランジュ関数は

$$L \equiv J(W, \Theta, P) + \sum_{m=1}^{M} \lambda_m \left\{ \sum_{k=1}^{K} P(k|\boldsymbol{x}_m) - 1 \right\}$$

で与えられる．$P = [P(k|\boldsymbol{x}_m)]_{m,k}$ の解は，$\partial L / \partial P(k|\boldsymbol{x}_m) = 0$, $\forall m$, $\forall k$ と $\partial L / \partial \lambda_m = 0$, $\forall m$ で得られる連立方程式を解くことで求めることができる．

$$\frac{\partial L}{\partial P(k|\boldsymbol{x}_m)} = -\log \frac{P(k|\boldsymbol{x}_m)}{P_d(k|\boldsymbol{x}_m; W)} - 1 + \beta \log p_g(\boldsymbol{x}_m, k; \boldsymbol{\theta}_k) + \lambda_m, \ \forall m, \forall k$$

$$\frac{\partial L}{\partial \lambda_m} = \sum_{k=1}^{K} P(k|\boldsymbol{x}_m) - 1, \ \forall m$$

であるため，各々の m ごとに連立方程式を解くことで

$$P(k|\boldsymbol{x}_m) = \frac{P_d(k|\boldsymbol{x}_m; W) p_g(\boldsymbol{x}_m, k; \boldsymbol{\theta}_k)^{\beta}}{\sum_{k'=1}^{K} P_d(k'|\boldsymbol{x}_m; W) p_g(\boldsymbol{x}_m, k'; \boldsymbol{\theta}_{k'})^{\beta}}, \ \forall m, \forall k$$

が得られる．

(平成 22 年 8 月 31 日受付)
(平成 22 年 10 月 19 日再受付)
(平成 22 年 12 月 7 日再受付 (2))
(平成 22 年 12 月 17 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 年同大学大学院修士課程修了．2009 年同大学院博士課程修了．博士（情報学）．1997 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所長，奈良先端科学技術大学院大学客員教授，国立情報学研究所客員教授．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，IEEE 各会員．

永田　昌明（正会員）
1987 年京都大学大学院工学研究科修士課程修了．工学博士．同年 NTT 入社．1989 年から 4 年間 ATR 自動翻訳電話研究所へ出向．1999 年から 1 年間 AT&T 研究所客員研究員．統計的自然言語処理の研究に従事．現在，NTT コミュニケーション科学基礎研究所主幹研究員．情報処理学会奨励賞（1991 年），情報処理学会論文賞（1995 年），人工知能学会研究奨励賞（1995 年）等受賞．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．