LI-009

生成・識別ハイブリッドモデルに基づく半教師あり学習
Semi-supervised learning based on a hybrid of generative and discriminative models

藤野 昭典† 上田 修功† 斉藤 和巳†
Akinori Fujino Naonori Ueda Kazumi Saito

1. まえがき

統計的手法に基づく分類器は，クラスラベル y が既知の特徴ベクトル $\boldsymbol{x}$ (ラベルありデータ) を用いて学習される．一般に，多数のラベルありデータを学習に用いることで分類器の精度が向上することが知られている．しかし，ラベルありデータの作成は，分類対象に精通した専門家によるラベルの付与が必要であり，高コストである．一方，ラベルなしデータは，低コストで集めることができることが多い．例えば web page の分類問題では，インターネットから容易にデータを収集することができる．したがって，少数のラベルありデータと多数のラベルなしデータを用いて分類器の精度を向上させる半教師あり学習法は機械学習の重要な研究課題の1つであり，従来よりさまざまな分類器が提案されてきた (文献 [10] 参照)．

確率モデルに基づく半教師あり学習法は，生成モデル，あるいは識別モデルアプローチに基づいて提案されてきた．生成モデルアプローチでは，特徴ベクトルとクラスの同時確率 $P(\boldsymbol{x}, y)$ をモデル化して学習する．ベイズ則によりクラス事後確率 $P(y|\boldsymbol{x})$ を求め，その確率を最大にするクラス y を選択することで分類を行う．ラベルなしデータは，クラスラベルに関する不完全データとして，混合モデルで扱われる [7]．

識別モデルアプローチでは，特徴ベクトルのクラス事後確率 $P(y|\boldsymbol{x})$ を直接モデル化する．ただし，識別モデルアプローチでは $P(\boldsymbol{x})$ をモデル化しないため，ラベルなしデータを学習に用いる際に新たな仮定が必要となる．例えば，特徴ベクトル間の距離に基づいた仮定 [11]，クラス事後確率のエントロピーに基づく仮定 [3] などが用いられている．

一般的に，識別モデルによる分類器は生成モデルと比較して高い分類性能をもつことが知られているが，ラベルありデータが少数であるとき，生成モデルによる分類器が識別モデルよりしばしば高い分類性能をもつことが報告されている [5]．そこで本研究では，生成・識別モデル双方の利点を活かすためのハイブリッドアプローチを検討する．

教師あり学習の枠組では，生成・識別モデルのハイブリッド法が試みられている [8]．文献 [8] による方法では，特徴ベクトル空間を部分空間に分割して，部分空間ごとに生成モデルを学習する．そして，これらの生成モデルをクラス事後確率を最大化するように重み付け結合することで分類器を獲得する．文献 [8] では，テキスト分類問題で，文書を本文とタイトルの2つの部分空間に分割してハイブリッド法を適用することで純粋な生成・識別モデルより高い分類精度が得られることが報告されている．

本稿では，半教師あり学習の問題に対して，生成・識別モデルのハイブリッドに基づく新しい分類器設計法を提案する．文献 [8] では，特徴ベクトルのハイブリッド統合を行っているのに対し，提案法では，ラベルあり・なしデータの統合を行っているという点で異なる．具体的には，まず，ラベルありデータにより生成モデルを学習する．ラベルありデータが少数であるとき，学習された生成モデルは高いバイアスをもつ．そこで，そのバイアスを補正するための新たなモデル (バイアス補正モデル) を導入する．最大エントロピー原理 [1] に基づき，クラス事後確率を最大化するように，生成モデルとバイアス補正モデルを結合することで分類器の識別関数を与える．バイアス補正モデルはラベルなしデータを用いて学習する．

2節では，従来の生成・識別モデルアプローチに基づく半教師あり学習法について述べ，3節では，本研究で提案するハイブリッドアプローチについて述べる．4節では，提案法の生成モデルにナイーブベイズモデルを用いてテキスト分類問題に適用したときの実験結果を示し，従来法との比較により提案法の有用性を考察する．

2. 従来法

多クラス分類問題は，K 個のクラスの候補 $\{1, \ldots, k, \ldots, K\}$ から特徴ベクトル $\boldsymbol{x}$ が属するクラス y を1つ選択する問題である．半教師あり学習では，少数のラベルありデータ $D_l = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$ と多数のラベルなしデータ $D_u = \{\boldsymbol{x}_m\}_{m=1}^M$ を用いて分類器を学習する．本節では，生成モデル，識別モデルのそれぞれに基づく従来の半教師あり学習法について述べる．

2.1 生成モデルアプローチ

生成モデルでは，同時確率モデル $P(\boldsymbol{x}, y|\Theta)$ を仮定し，訓練データを用いてパラメータ Θ を学習する．$\boldsymbol{x}$ の属するクラスは，ベイズ則によりクラス事後確率 $P(y|\boldsymbol{x}, \Theta)$ を求め，その確率を最大化するクラス y を選択することで推定される．同時確率モデルは分類対象の特徴に合わせて，多項分布モデルやガウス分布モデルを仮定することができる．

生成モデルによる半教師あり学習では，混合モデルを仮定して，ラベルなしデータをクラスラベルに関する不完全データとして扱う [2]．訓練データ $D = \{D_l, D_u\}$ が与えられるとき，MAP 推定では，以下の目的関数を最大化する Θ をパラメータの推定値とする．

$$J(\Theta) = \sum_{n=1}^{N} \log P(\boldsymbol{x}_n, y_n|\Theta) + \sum_{m=1}^{M} \log \sum_{k=1}^{K} P(\boldsymbol{x}_m, k|\Theta) + \log P(\Theta). \quad (1)$$

上式の $P(\Theta)$ は Θ の事前確率である．$J(\Theta)$ によるパラ

†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

メータ推定はEMアルゴリズムを用いて行うことができる．

Θの推定は，NとMの大きさに影響される．$M >> N$のとき，Θの推定は，式(1)の第二項に強く影響され，ラベルなしデータによる教師なし学習に近い推定となる．生成モデルと混合モデルの仮定が分類対象に対してよく合っていない場合，ラベルなしデータをパラメータ学習に用いることで分類精度がむしろ悪化する危険性がある．この問題に対処するために，式(1)の第二項のパラメータ学習への寄与を下げる重みパラメータλを導入する方法(EM-λ)が提案されている[7]．この方法では，ラベルありデータのleave-one-out交差確認により，分類精度を最大にするパラメータΘを与えるλを探索する．探索されたλの下で，パラメータΘを推定する．

2.2 識別モデルアプローチ

識別モデルでは，クラス事後確率$P(y|\boldsymbol{x})$が直接モデル化される．多項ロジスティック回帰モデル(MLR)[4]では，クラス事後確率は未知パラメータ$W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K\}$を用いて以下のようにモデル化される．

$$P(k|\boldsymbol{x}, W) = \frac{\exp(\boldsymbol{w}_k \cdot \boldsymbol{x})}{\sum_{k'=1}^{K} \exp(\boldsymbol{w}_{k'} \cdot \boldsymbol{x})} \tag{2}$$

$\boldsymbol{w}_k \cdot \boldsymbol{x}$は$\boldsymbol{w}_k$と$\boldsymbol{x}$の内積を表す．

識別モデルでは，ラベルなしデータをパラメータ学習に用いるために，新たな仮定を必要とする．例えば，最小エントロピー正則項(MER)[3]に基づく方法では，各クラスによく分離されているラベルなしデータはパラメータ学習に有効であるとし，クラス事後確率のエントロピーを最小化することでラベルなしデータをよく分離するようにモデルを学習する．MERを用いたMLRの学習(MLR/MER)では，以下の目的関数を最大化するWをパラメータの推定値とする．

$$\begin{aligned} J(W) &= \sum_{n=1}^{N} \log P(y_n|\boldsymbol{x}_n, W) \\ &+ \lambda \sum_{m=1}^{M} \sum_{k=1}^{K} P(k|\boldsymbol{x}_m, W) \log P(k|\boldsymbol{x}_m, W) \\ &+ \log P(W) \end{aligned} \tag{3}$$

λは重みパラメータであり，$P(W)$はWの事前確率である．

3. ハイブリッドアプローチ

本研究では，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を提案する．図1にハイブリッド法に基づく分類器の構成を示す．提案法では，まず，クラスkにおける生成確率を表す生成モデル$P(\boldsymbol{x}|k, \theta_k)$を仮定し，教師あり学習と同様に，ラベルありデータを用いて学習する．ラベルありデータが少数であるとき，学習された生成モデル$P(\boldsymbol{x}|k, \hat{\theta}_k)$は統計的に高い偏りをもつ．そこで，生成モデルと同型の分布(パラメータは異なる)としてバイアス補正モデル$P(\boldsymbol{x}|k, \phi_k)$を新たに導入する．これらを，クラス事後確率を最大にするように，最大エントロピー(ME)原理[1]に基づいて結合する．すなわち，提案法では，データの生成過程を積極的に利用する生成モデルアプローチの長所と，識別率を最大化する識別モデルアプローチの長所を相補的に備えたハイブリッドアプローチとなっている．

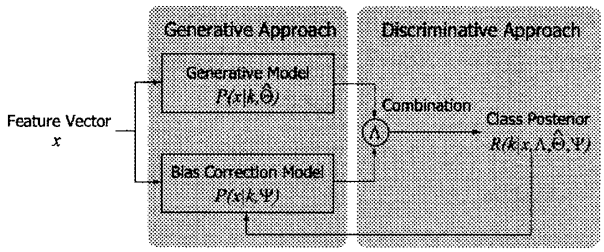

図1: ハイブリッド法に基づく分類器の構成

具体的には，分類器の識別関数は，生成モデルとバイアス補正モデルに対する制約の下で，ME原理を満たすクラス事後確率分布(MEモデル)$R(k|\boldsymbol{x})$として与える．ME原理は，与えられた制約を満たすモデルの中で最も一様な分布を獲得する枠組のひとつである．すなわち，提案法では，ME原理の下で生成モデルとバイアス補正モデルを反映する最も一様なクラス事後確率分布を分類器の識別関数として定義する．

MEモデルに生成モデルの特性を反映させるため，$R(k|\boldsymbol{x})$による生成モデルの対数尤度の期待値と，訓練データの経験分布$\tilde{P}(\boldsymbol{x}, k) = \frac{1}{N}\sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n, k - y_n)$による生成モデルの対数尤度の期待値が等しい，という制約を与える．この制約は，$\boldsymbol{x}$の経験分布$\tilde{P}(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n)$を用いて，以下の式で表される．

$$\begin{aligned} &\sum_{\boldsymbol{x},k} \tilde{P}(\boldsymbol{x}, k) \log P(\boldsymbol{x}|k, \hat{\theta}_k) \\ &= \sum_{\boldsymbol{x},k} \tilde{P}(\boldsymbol{x}) R(k|\boldsymbol{x}) \log P(\boldsymbol{x}|k, \hat{\theta}_k) \end{aligned} \tag{4}$$

バイアス補正モデルの対数尤度$\log P(\boldsymbol{x}|k, \psi_k)$に対する制約は，式(4)と同様の形で与えることができる．また，MEモデルにデータの帰属するクラスの偏りを反映させるために，$R(k|\boldsymbol{x})$によるデータがクラスk'に属する期待値と，訓練データの経験分布に基づくクラスk'に属する期待値が等しい，という制約を与える．この制約は，$z_{k'}$は$\boldsymbol{x}$がクラスk'に属するときに1，それ以外のときに0をとる変数を用いて，以下の式で表すことができる．

$$\sum_{\boldsymbol{x},k} \tilde{P}(\boldsymbol{x}, k) z_{k'} = \sum_{\boldsymbol{x},k} \tilde{P}(\boldsymbol{x}) R(k|\boldsymbol{x}) z_{k'}, \forall k' \tag{5}$$

以上の制約の下で，クラス事後確率分布のエントロピー$H(R) = -\sum_{\boldsymbol{x},k} \tilde{P}(\boldsymbol{x}) R(k|\boldsymbol{x}) \log R(k|\boldsymbol{x})$を最大化することにより，生成モデル，バイアス補正モデルとクラスの偏りを反映したMEモデルに基づくクラス事後確率:

$$\begin{aligned} &R(k|\boldsymbol{x}, \hat{\Theta}, \Psi, \Lambda) \\ &= \frac{P(\boldsymbol{x}|k, \hat{\theta}_k)^{\lambda_1} P(\boldsymbol{x}|k, \psi_k)^{\lambda_2} e^{\mu_k}}{\sum_{k'=1}^{K} P(\boldsymbol{x}|k', \hat{\theta}_{k'})^{\lambda_1} P(\boldsymbol{x}|k', \psi_{k'})^{\lambda_2} e^{\mu_{k'}}} \end{aligned} \tag{6}$$

が導出できる．ここで，$\Lambda = \{\lambda_1, \lambda_2, \{\mu_k\}_{k=1}^{K}\}$はラグランジュ乗数である．$\lambda_1$と$\lambda_2$は生成モデルとバイアス補

正モデルの結合の重みを，μ_k はクラス k の偏りを与える未知パラメータである．$R(k|\boldsymbol{x}, \hat{\Theta}, \Psi, \Lambda)$ は，生成モデルとバイアス補正モデルの結合により構成される分類器の識別関数とみなせる．

バイアス補正モデルは，分類器 $R(k|\boldsymbol{x}, \hat{\Theta}, \Psi, \Lambda)$ の訓練に際して，少数のラベルありデータから学習された生成モデルの統計的な偏りを補正するように，ラベルなしデータを用いて学習する．具体的には，式 (6) で表されるクラス事後確率分布を用いてラベルなしデータの属するクラスを推定して，バイアス補正モデルのパラメータ Ψ を学習する．ラベルなしデータの属するクラスを推定するには，Λ を与える必要がある．しかし，Λ は推定すべき未知パラメータであり，Ψ と互いに依存関係がある．このため，Ψ と Λ は反復的に学習する．パラメータ学習のアルゴリズムは誌面の都合により省略する．

4. 評価実験

4.1 テストコレクション

提案法の性能評価は，テキスト分類問題に適用することで行った．実験には，テキスト分類器のベンチマークテストによく用いられる Reuters-21578 (Reuters) [12] と WebKB, 20 Newsgroups (20news) [6] の 3 つのテストコレクションを用いた．

Reuters は，Reuters newswire の 135 のトピックカテゴリからなるデータセットである．本実験では，8 つのカテゴリ **acq, crude, earn, grain, interest, money-fx, ship, trade** に含まれる記事のみを抽出し，そのうち複数のカテゴリに属する記事を除外することで多クラス単一ラベルのデータセットを作成した．Reuters によるベンチマークテストでは，投稿時期の早い記事と遅い記事がそれぞれ訓練データ，テストデータとして用いられる．本実験でも同様に，投稿時期の遅い記事をテストデータとして用い，投稿時期の早い記事の中からラベルあり・なしデータを選択した．記事の特徴ベクトルを作成する際に，記事に含まれる停止語 [9] と 1 つの記事のみに出現する低頻度語彙を除去した．

WebKB は大学の web page を集めたものであり，7 つのカテゴリに分類されている．文献 [6] に従い，**student, faculty, course, project** の 4 つのカテゴリに含まれる web page を実験に用いた．web page に含まれるタグ，URL 等を除外して，テキストデータのみからページの特徴ベクトルを作成した．Reuters と同様に停止語と低頻度語彙を除去した．

20news は，UseNet の記事を集めたものであり，20 グループに分類されている．文献 [6] に従い，20 グループのうち，**comp.*** の 5 つのグループに属する記事を実験に用いた．Reuters と同様に停止語と低頻度語彙を除去した．

以上の処理により作成したデータセットに含まれる文書数 $|D|$，クラス数 K，語彙数 V を表 1 に示す．実験では，文書の特徴ベクトルとして語彙の出現頻度を用いた．

4.2 実験方法

テキスト分類に提案法を適用するために，生成モデルとバイアス補正モデルにナイーブベイズ (NB) モデル [7] を用いた．提案法を，2 節で述べた EM-λ および MLR/MER と比較評価した．EM-λ の生成モデルにも NB モデルを用いた．また，ラベルありデータのみで教師あり学習を行う NB モデルと MLR の分類性能も合わせて調べた．

表 1: テストコレクション

Name	$\lvert D\rvert$	K	V	$\lvert D_u\rvert$	$\lvert D_t\rvert$
Reuters	8162	8	11822	4500	2430
WebKB	4199	4	18525	2500	1000
20news	4881	5	19383	2500	1000

EM-λ の重みパラメータ λ は $\{0.01, 0.05, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 1.0\}$ の 14 候補から 2 節で述べた方法で推定した．MLR/MER の重みパラメータ λ の候補は $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.2, 0.5, 1\}$ の 8 つとした．λ は交差確認法などを用いて訓練データのみから推定すべきであるが，MLR/MER は計算コストが高いため，テストデータに対して最も高い分類精度を与える値を選択した．

各方法の評価には，正しく分類されたテストデータの割合 (分類精度) を用いた．各テストコレクションからランダムにデータを選択して 10 種類の学習・テストデータセットを作成し，10 回の実験で得られた分類精度の平均値をもとに各方法の分類性能を比較した．表 1 に，学習に用いたラベルなしデータ数 $|D_u|$ とテストデータ数 $|D_t|$ を示す．ラベルありデータは残りのデータから選択した．

4.3 実験結果と考察

表 2 (a)–(c) に Reuters, WebKB, 20news の各テストコレクションでの 10 回の実験による分類精度の平均値を示す．括弧内の数値は標準偏差を示す．表中の $|D_l|$ は学習に用いたラベルありデータ数を示す．

教師あり学習の場合では，文献 [5] で報告されているように，ラベルありデータが少数の場合は，生成モデルの分類性能が識別モデルよりも高く，逆に多数の場合は識別モデルの分類性能の方が高い傾向が本実験でもみられた．Reuters と WebKB では，WebKB の $|D_l| = 8$ 以外の場合で，NB と MLR の分類精度の結果が文献 [5] の報告と一致した．

しかし，20news では，ラベルありデータ数によらず，MLR の分類精度は NB よりも高かった．この結果を分析するため，以下の式で定義される，語彙数で正規化した NB モデルのテストパープレキシティ $\mathcal{P}$ をテストコレクションごとに調べた．

$$\mathcal{P} = \frac{1}{V} \exp\left(-\frac{\sum_{k=1}^{K} \sum_{s=1}^{S} z_{sk} \sum_{i=1}^{V} x_{si} \log \hat{\theta}_{ki}}{\sum_{s=1}^{S} \sum_{i=1}^{V} x_{si}} \right) \quad (7)$$

$\mathcal{P}$ は，学習に用いていないテストデータ $(\boldsymbol{x}_s, y_s)$ に対して分類器がどれだけ適合しているかを表す指標である．上式の $\hat{\theta}_{ki}$ は訓練データによって学習されたモデルパラメータである．z_{sk} は $y_s = k$ のときに 1，$y_s \neq k$ のときに 0 をとる変数である．$\mathcal{P}$ が小さいほど，モデルはテストデータに適合していることを示す．図 2 に示されるように，訓練データ数が 10^3 より少ないとき，20news の $\mathcal{P}$ は，Reuters, WebKB と比べて大きかった．これは，20news では訓練データが少数のときは学習された NB

表 2: 各テストコレクションでの分類精度 (%)

(a) Reuters

Training set		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Proposed	EM-λ	MLR/MER	NB	MLR
16	0.0036	83.6 (4.7)	**86.3** (3.6)	73.9 (6.3)	69.1 (9.1)	67.8 (7.5)
32	0.0071	**89.8** (1.6)	**89.8** (1.5)	82.2 (4.0)	81.9 (2.6)	80.9 (3.5)
64	0.014	**92.3** (0.8)	91.8 (1.2)	83.5 (5.0)	84.9 (2.8)	83.1 (4.6)
128	0.028	**92.9** (0.6)	92.6 (0.8)	88.5 (1.4)	89.1 (0.6)	87.8 (1.3)
256	0.057	**93.3** (0.7)	93.1 (0.8)	90.8 (0.8)	91.1 (1.1)	90.6 (0.8)
512	0.11	**94.1** (0.5)	93.7 (0.4)	93.3 (0.7)	93.0 (0.6)	93.2 (0.7)
1024	0.23	94.6 (0.3)	94.2 (0.3)	**94.7** (0.2)	93.8 (0.3)	94.5 (0.2)

(b) WebKB

Training set		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Proposed	EM-λ	MLR/MER	NB	MLR
8	0.0032	61.3 (6.0)	**65.4** (6.7)	53.9 (4.4)	52.1 (4.1)	53.6 (4.3)
16	0.0064	**66.7** (4.4)	64.6 (9.0)	54.0 (5.9)	54.9 (6.3)	54.0 (5.2)
32	0.013	**72.9** (2.7)	70.5 (5.0)	63.5 (4.6)	66.6 (5.3)	63.5 (4.3)
64	0.026	**76.8** (1.9)	73.2 (2.6)	72.0 (2.6)	73.2 (1.6)	71.9 (2.3)
128	0.051	**79.4** (1.4)	76.6 (2.4)	78.1 (2.4)	77.5 (1.8)	78.0 (2.2)
256	0.10	81.4 (1.6)	79.5 (1.6)	**83.6** (1.7)	80.0 (1.4)	83.4 (1.7)
512	0.20	83.2 (1.6)	81.8 (1.7)	**87.9** (1.3)	82.3 (1.4)	87.5 (1.1)

(c) 20news

Training set		Semi-supervised			Supervised	
$\|D_l\|$	$\frac{\|D_l\|}{\|D_u\|}$	Proposed	EM-λ	MLR/MER	NB	MLR
10	0.0040	**52.2** (14.2)	40.4 (8.1)	41.9 (8.4)	33.9 (5.9)	37.6 (5.5)
20	0.0080	**63.6** (5.6)	49.5 (7.2)	45.2 (5.1)	40.4 (4.7)	44.6 (4.2)
40	0.016	**68.5** (2.7)	52.3 (5.1)	52.4 (5.4)	45.9 (3.4)	50.9 (3.6)
80	0.032	**72.8** (2.4)	59.1 (4.8)	59.5 (2.7)	52.6 (3.5)	59.0 (2.3)
160	0.064	**76.1** (1.3)	65.8 (4.4)	67.7 (2.7)	61.5 (2.4)	66.6 (2.0)
320	0.13	**78.4** (0.9)	71.1 (2.7)	73.8 (1.3)	68.8 (1.9)	72.7 (1.2)
640	0.26	**81.3** (1.2)	75.2 (2.0)	79.1 (1.3)	74.7 (1.6)	77.7 (1.2)
1280	0.51	**83.8** (1.1)	78.2 (1.9)	82.3 (1.4)	78.1 (1.8)	81.1 (1.3)

モデルはテストデータにあまり適合しないことを示している．逆に，Reuter, WebKB のように，訓練データが少数のときに $\mathcal{P}$ が小さければ，NB は MLR の性能を上回る傾向があった．この結果，教師あり学習において，訓練データが少数で生成モデルのテストパープレキシティが十分小さいときに，生成モデルの分類性能は識別モデルを上回るといえる．

半教師あり学習の場合の，生成モデル，識別モデルと提案法の分類精度を比較する．まず，EM-λ は，すべてのテストコレクションにおいて，ラベルありデータが少数の場合は MLR/MER とほぼ同等，またはより高い分類精度を示し，逆に多数の場合は MLR/MER より低い分類精度を示した．このように，生成モデル，識別モデルに基づく方法では，教師あり学習での報告 [5] と同様の特徴が半教師あり学習の場合にもみられた．

つぎに，EM-λ と提案法の比較では，教師あり学習において MLR の分類性能が NB を上回るときに，提案法の分類性能が EM-λ を上回る傾向があった．この結果は，提案法では生成・識別モデルのハイブリッドにより，識別モデルの利点が活かされていることを示唆している．

最後に，MLR/MER と提案法の比較では，ラベルありデータが多数の場合を除いて，すべてのテストコレクションで提案法の分類性能が MLR/MER を上回った．これは，MLR/MER は少数のラベルありデータに対して過学習する傾向があるのに対し，提案法は過学習を抑制する生成モデルの特徴を備えているからと考えられる．過学習を引き起こさないように多数のラベルありデータが与えられるとき，識別モデルの分類性能は提案法を上回ると考えられる．

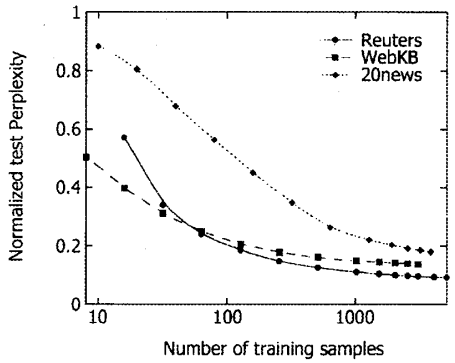

図 2: 各テストコレクションでのテストパープレキシティ

5. まとめ

本稿では，生成・識別モデルのハイブリッド法による，半教師あり学習のための，新しい分類器設計法を提案した．3 つの実データを用いたテキスト分類実験により，提案法を従来の生成モデルと識別モデルに基づく方法と比較した．その結果，識別モデルの分類性能が生成モデルとほぼ同等か，やや上回るときに，ハイブリッド法の分類性能が生成モデルと識別モデルを上回ることを確認した．

参考文献

[1] Berger, A., Della Pietra, S., and Della Pietra, V. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22(1)**, 39–71.

[2] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

[3] Grandvalet, Y. and Bengio, Y. (2005). Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17* (pp. 529–536). Cambridge, MA: MIT Press.

[4] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag New York Berlin Heidelberg.

[5] Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 14* (pp. 841–848). Cambridge, MA: MIT Press.

[6] Nigam, K., Lafferty, J., and McCallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, 61–67.

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

[8] Raina, R., Shen, Y., Ng, A. Y., and McCallum, A. (2004). Classification with hybrid generative/discriminative models. In *Advances in Neural Information Processing Systems 16.* Cambridge, MA: MIT Press.

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

[10] Seeger, M. (2001). Learning with labeled and unlabeled data, Technical Report, University of Edinburgh.

[11] Szummer, M. and Jaakkola, T. (2001). Kernel expansions with unlabeled examples. In *Advances in Neural Information Processing Systems 13* (pp. 626–632). Cambridge, MA: MIT Press.

[12] Yang, Y. and Liu, X. (1999). 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.