Vol. 47 No. 10 情報処理学会論文誌 Oct. 2006

最大エントロピー原理に基づく付加情報の効果的な利用によるテキスト分類

藤野昭典† 上田修功† 斉藤和巳†

Webページのリンク情報など，本文の他に付加情報を含むテキストデータの分類問題のために，付加情報を同時に用いて高精度な多クラス分類器を設計する手法を提案する．この問題に対して，従来の確率的アプローチでは，生成，識別の各アプローチと，生成，識別アプローチのハイブリッドに基づく分類器が提案されてきた．従来のハイブリッド分類器が2クラス問題を対象とするのに対して，提案法では多クラス問題を直接扱うハイブリッド分類器を与える．具体的には，データに含まれる構成要素ごとに設計した生成モデルを最大エントロピー原理に基づいて結合することで分類器を構築する．文書やWebページに含まれるテキスト，リンクの各構成要素の生成モデルとして，ナイーブベイズモデルを用いる．3つの実データを用いた分類実験により，付加情報をテキスト分類に用いる効果を確認するとともに，生成，識別アプローチでの分類精度の差が小さいほど提案法による分類精度が両アプローチを大きく上回ることを確認した．

Text Classification by Effectively Using Additional Information Based on Maximum Entropy Principle

Akinori Fujino,† Naonori Ueda† and Kazumi Saito†

We propose a multi-class text classifier that can handle both main text and additional information such as link information in web pages and thus improve classification performance. Existing probabilistic approaches to classifier design with main text and additional components are generative, discriminative, or a hybrid of the two. As the conventional hybrid classifier was designed for binary classification, we present a hybrid classifier for dealing directly with multi-class classification, which is constructed by combining component generative models based on the maximum entropy principle. We use naive Bayes models as component generative models designed for text and link information contained in documents and web pages. Our experimental results for three test collections confirmed the effectiveness of using additional information for text classification. The results also revealed that our hybrid classifier greatly outperformed both the generative and discriminative classifiers when there was little difference in their performance.

1. はじめに

近年，コンピュータやインターネットの普及によりWebページやBlog，Eメールなどのテキストデータが飛躍的に増大している．これらのデータは複数の構成要素からなるものが多い．たとえば，Webページは本文以外にタイトル，ハイパーリンク，アンカーテキストなどの付加情報を含む．このようなデータの分類では，本文が最も重要な役割を果たすとともに，その他の構成要素についても分類精度の向上に寄与する情報を含む可能性がある．このため，Webページの本文とハイパーリンク[4),6),11),16)]，文書の本文と引用情報[6),11)]，楽曲のテキストと音情報[3)]など，複数の構成要素を扱う分類器が提案されている．本論文では，文献3), 11)のように，任意の付加情報を扱うことが可能なテキスト分類器を，教師あり学習の枠組みで確率的アプローチにより設計する手法に焦点を当てる．

確率的アプローチに基づく複数の構成要素を扱う分類器は，生成，識別アプローチと，ハイブリッド法の3つに大別される．生成アプローチでは，データの特徴ベクトル $\boldsymbol{x}$ とクラスラベル y の同時確率分布 $p(\boldsymbol{x}, y)$☆をモデル化し，ベイズ則に基づいてクラス事後確率 $P(y|\boldsymbol{x})$ を計算することでデータのクラスラベ

† 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

☆ 本論文では，確率と確率密度関数をそれぞれ $P(y)$ と $p(\boldsymbol{x})$ のように大文字・小文字で区別して表記する．また，確率変数に対する確率と確率密度関数の分布をともに確率分布と呼ぶ．

ルを推定する．同時確率分布は，しばしば，構成要素の独立性を仮定し，構成要素の生成確率分布の積に分解してモデル化される[3]．

識別アプローチでは，クラス事後確率 $P(y|\boldsymbol{x})$ を直接モデル化する．この方法では，構成要素を意識せずに，単一のクラス事後確率モデルを与えることができる．また，構成要素を考慮して，構成要素ごとにクラス事後確率モデルを設計し，その積からクラスラベルを推定する方法[11]も提案されている．

ハイブリッド法では，生成アプローチによる構成要素のモデル化（構成要素モデル）と，識別アプローチによる構成要素モデルの重み付け結合に基づいてクラス事後確率分布を与える．文献 14) では，本文とタイトルの 2 つの構成要素を含む文書に対して，ハイブリッド法に基づく 2 クラス分類器を提案している．

本論文では，ハイブリッド法に基づく分類器の多クラス問題（3 クラス以上）への拡張を与え，複数の付加情報を含むテキストデータのための分類器設計法を提案する．提案法では，最大エントロピー（ME）原理[1]に基づいてクラス事後確率分布を定式化することで，文献 14) の 2 クラス分類器を特殊な場合として含む多クラス分類器を与える．また，分類器の学習に際して，交差検定法に基づいて構成要素モデルのパラメータを最適化することで，汎化性能の向上を図る．

提案法を文書，Web ページ分類問題に適用して，その有用性を実験的に確認する．実験では，テキスト情報とリンク情報による構成要素を用いて分類器を設計する．各構成要素は，ナイーブベイズ（NB）モデル[17]を用いてモデル化する．3 つの実データを用いた実験により，生成，識別の各アプローチに対する提案法の優位性を示す．

2. 従 来 法

本論文では，複数の構成要素からなるテキストデータの多クラス単一ラベル分類問題に対して，汎化性能の高い分類器を設計することを課題とする．

多クラス単一ラベル分類問題とは，K 個の候補の中からデータ $\boldsymbol{x}$ のクラスラベル $y \in \{1,\ldots,k,\ldots,K\}$ を 1 つに決定する問題である．J 個の構成要素からなるデータは，構成要素 j の特徴ベクトル $\boldsymbol{x}^j$ を用いて，$\boldsymbol{x}=\{\boldsymbol{x}^1,\ldots,\boldsymbol{x}^j,\ldots,\boldsymbol{x}^J\}$ で表される．分類器の学習は，訓練データ集合 $D=\{\boldsymbol{x}_n,y_n\}_{n=1}^N$ を用いて行う．以下に，確率的アプローチに基づく従来法について述べる．

2.1 生成アプローチ

生成アプローチでは，データ $\boldsymbol{x}$ とクラスラベル y の同時確率分布 $p(\boldsymbol{x},y)$ をモデル化する．しかし，異なる特性を持つ構成要素を含むデータを直接モデル化することは容易ではない．そこで，各クラスで各構成要素は “独立” に生成されると仮定し，構成要素ごとに仮定する生成モデル（構成要素モデル）$p(\boldsymbol{x}^j|y;\boldsymbol{\theta}_y^j)$ を用いて同時確率分布を

$$p(\boldsymbol{x},y;\boldsymbol{\theta}_y)=P(y)\prod_{j=1}^{J}p(\boldsymbol{x}^j|y;\boldsymbol{\theta}_y^j) \tag{1}$$

のようにモデル化する[3]．ここで，$\boldsymbol{\theta}_y^j$ は y における構成要素 j のモデルパラメータを表す．

モデルパラメータ $\Theta=\{\boldsymbol{\theta}_k^j\}_{j,k}$ は，訓練データ集合 D を用いて最大事後確率（MAP）推定を行うとき，以下の目的関数の最大化により推定される．

$$G(\Theta)=\sum_{n=1}^{N}\left\{\log P(y_n)+\sum_{j=1}^{J}\log p(\boldsymbol{x}_n^j|y_n;\boldsymbol{\theta}_{y_n}^j)\right\}+\sum_{j=1}^{J}\sum_{k=1}^{K}\log p(\boldsymbol{\theta}_k^j) \tag{2}$$

ここで，$p(\boldsymbol{\theta}_k^j)$ はパラメータ $\boldsymbol{\theta}_k^j$ の事前確率分布を表す．明らかに構成要素 j のモデルパラメータは他の構成要素と独立に最適化される．

データのクラス事後確率は，ベイズ則により

$$P(y=k|\boldsymbol{x};\Theta)=\frac{P(k)\prod_{j=1}^{J}p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)}{\sum_{k'=1}^{K}P(k')\prod_{j=1}^{J}p(\boldsymbol{x}^j|k';\boldsymbol{\theta}_{k'}^j)} \tag{3}$$

として与えられる（ベイズ事後確率）．データのクラスラベルは，クラス事後確率を最大にする k として推定される．これは，データのクラスラベルを $p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)$ の単純な積による比較から予測することを意味する．

2.2 識別アプローチ

識別アプローチでは，特徴ベクトル $\boldsymbol{x}$ のクラス事後確率 $P(y|\boldsymbol{x})$ を直接モデル化できる．多項ロジスティック回帰（MLR）モデル[9]では，クラス事後確率分布はモデルパラメータ $W=\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_k,\ldots,\boldsymbol{w}_K\}$ を用いて以下のように書ける．

$$P(y=k|\boldsymbol{x};W)=\frac{\exp(\boldsymbol{w}_k\cdot\boldsymbol{x})}{\sum_{k'=1}^{K}\exp(\boldsymbol{w}_{k'}\cdot\boldsymbol{x})} \tag{4}$$

ここで，$\boldsymbol{w}_k\cdot\boldsymbol{x}$ は $\boldsymbol{w}_k$ と $\boldsymbol{x}$ の内積を表す．文献 12) では，特徴ベクトル $\boldsymbol{x}$ に対して直接的に ME 原理を適用してクラス事後確率をモデル化している．このモデルによるクラス事後確率分布も式 (4) のように書ける．

モデルパラメータ W は，訓練データのクラス事後確率の対数尤度と W の事前確率分布 $p(W)$ の対数の和で表される以下の目的関数の最大化により推定さ

れる．

$$F(W)=\sum_{n=1}^{N}\log P(y_n|\boldsymbol{x}_n;W)+\log p(W) \quad (5)$$

文献 11) では，分類器に構成要素の特徴を反映させるために，構成要素ごとに MLR モデル $P(k|\boldsymbol{x}^j;W^j)$ を設計し，個別に学習する手法を提案している．この方法では，MLR モデルの積 $Q=\prod_{j=1}^{J}P(k|\boldsymbol{x}^j;W^j)$ が最大になる k をデータ $\boldsymbol{x}$ のクラスラベルとして推定する．Q は，$Z=\sum_{k=1}^{K}\prod_{j=1}^{J}P(k|\boldsymbol{x}^j;W^j)$ を用いて正規化し，$\boldsymbol{w}_k=\{\boldsymbol{w}_k^1,\ldots,\boldsymbol{w}_k^j,\ldots,\boldsymbol{w}_k^J\}$ とすることで，式 (4) の右辺と一致する．すなわち，構成要素ごとに MLR モデルを設計する方法は，データを直接 MLR でモデル化する方法と分布の形状が一致する．文献 11) の方法は，パラメータ学習を個別に行うことで構成要素の特徴を分類器に反映させている．

2.3 2 クラス問題に対するハイブリッド分類器

文献 14) では，2 クラス分類問題 $y\in\{1,2\}$ に対し，生成アプローチによる構成要素のモデル化と，識別アプローチであるロジスティック回帰モデルとのハイブリッドに基づく方法（ハイブリッド法）を提案している．この方法では，式 (3) を

$$P(y=1|\boldsymbol{x};\Theta)=\frac{1}{1+\exp\left\{\sum_{j=1}^{J}\log\frac{p(\boldsymbol{x}^j|2;\boldsymbol{\theta}_2^j)}{p(\boldsymbol{x}^j|1;\boldsymbol{\theta}_1^j)}+\log\frac{P(2)}{P(1)}\right\}} \quad (6)$$

のように変形した分布に，構成要素 j の重み b_j と $b_0=\log\{P(2)/P(1)\}$ を新たに導入した

$$R(y=1|\boldsymbol{x};\Theta,B)\equiv\frac{1}{1+\exp\left\{\sum_{j=1}^{J}b_j\log\frac{p(\boldsymbol{x}^j|2;\boldsymbol{\theta}_2^j)}{p(\boldsymbol{x}^j|1;\boldsymbol{\theta}_1^j)}+b_0\right\}} \quad (7)$$

を分類器のクラス事後確率分布として定義する．重みパラメータ $B=\{b_j\}_{j=0}^{J}$ の値は，$R(y=1|\boldsymbol{x};\Theta,B)$，$R(y=2|\boldsymbol{x};\Theta,B)=1-R(y=1|\boldsymbol{x};\Theta,B)$ を対数尤度比 $\log\{p(\boldsymbol{x}^j|2;\boldsymbol{\theta}_2^j)/p(\boldsymbol{x}^j|1;\boldsymbol{\theta}_1^j)\}$ に対するロジスティック回帰モデルと見なして推定される．すなわち，ハイブリッド法では，構成要素モデル $p(\boldsymbol{x}^j|y;\boldsymbol{\theta}_y^j)$ を識別アプローチより重み付け結合することで分類器を構築する．**図 1** に，ハイブリッド法の概要をまとめる．

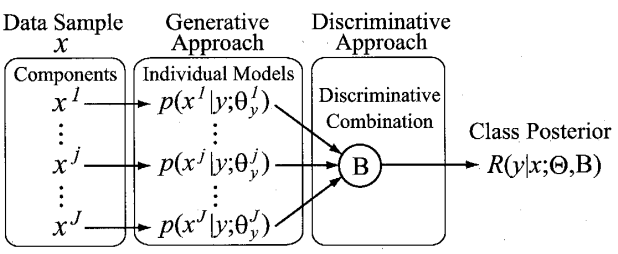


図 1 生成・識別アプローチのハイブリッド法の概要
Fig. 1 The outline of a hybrid of generative and discriminative approaches.

3. 提 案 法

本論文では，複数の付加情報を含むテキストデータの多クラス分類問題に対して，生成，識別アプローチのハイブリッドに基づく分類器設計法を提案する．提案法では，図 1 に示したハイブリッド法と同様に，生成アプローチによる構成要素のモデル化と，識別アプローチによる構成要素モデルの結合により分類器を構築する．式 (7) によるハイブリッド分類器が 2 クラス問題を対象としているのに対し，提案法では，最大エントロピー（ME）原理に基づく構成要素モデルの結合により，多クラス（3 クラス以上）問題を直接扱うハイブリッド分類器を与える．以下に，提案法の詳細と，文書，Web ページ分類への適用方法について述べる．

3.1 多クラス問題を扱うハイブリッド分類器

3.1.1 構成要素モデル

提案法ではまず，クラスラベル y が k であるときのデータ $\boldsymbol{x}$ の構成要素 j の生成モデル（構成要素モデル）$p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)$ を個別に設計する．ここで，$\boldsymbol{\theta}_k^j$ はモデルパラメータを表す．2.1 節で述べた生成アプローチと同様の MAP 推定により，訓練データ集合 D に対して以下の目的関数を最大化するように構成要素モデルのパラメータ $\Theta^j=\{\boldsymbol{\theta}_k^j\}_k$ を学習する．

$$G_j(\Theta^j)=\sum_{n=1}^{N}\log p(\boldsymbol{x}_n^j|y_n;\boldsymbol{\theta}_{y_n}^j)+\sum_{k=1}^{K}\log p(\boldsymbol{\theta}_k^j) \quad (8)$$

3.1.2 クラス事後確率モデル

次に，分類器の識別関数を，すべての構成要素モデルの特性を反映する，エントロピー基準の下で最も一様なクラス事後確率分布として定義する．この分布は，構成要素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 $R(k|\boldsymbol{x})$ として与えられる．

$R(k|\boldsymbol{x})$ に構成要素モデルの特性を反映させるため，$R(k|\boldsymbol{x})$ による構成要素モデルの対数尤度の期待値と，訓練データの経験分布 $\tilde{p}(\boldsymbol{x},k)=\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n,k-y_n)/N$ による構成要素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 $\tilde{p}(\boldsymbol{x})=\sum_{n=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_n)/N$ を用いて，以下の式で表される．

$$\sum_{\boldsymbol{x},k} \tilde{p}(\boldsymbol{x},k) \log p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j)$$
$$= \sum_{\boldsymbol{x},k} \tilde{p}(\boldsymbol{x}) R(k|\boldsymbol{x}) \log p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j),\ \forall j \qquad (9)$$

ここで，$\hat{\Theta} = \{\hat{\boldsymbol{\theta}}_k^j\}_{j,k}$ は構成要素モデルのパラメータの推定値を表す．また，$R(k|\boldsymbol{x})$ にクラスに対するデータの帰属の偏りを反映させるために，$R(k|\boldsymbol{x})$ によるクラス確率の推定値と，訓練データの経験分布によるクラス確率の推定値が等しい，という制約を与える．この制約は以下の式で表すことができる．

$$\sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x},k) = \sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x}) R(k|\boldsymbol{x}),\ \forall k \qquad (10)$$

以上の制約の下で，確率分布 $R(k|\boldsymbol{x})$ 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，ME 原理に基づく構成要素モデルとクラスの偏りを反映したクラス事後確率分布：

$$R(k|\boldsymbol{x};\hat{\Theta},\Lambda) = \frac{e^{\mu_k} \prod_{j=1}^{J} p(\boldsymbol{x}^j|k;\hat{\boldsymbol{\theta}}_k^j)^{\lambda_j}}{\sum_{k'=1}^{K} e^{\mu_{k'}} \prod_{j=1}^{J} p(\boldsymbol{x}^j|k';\hat{\boldsymbol{\theta}}_{k'}^j)^{\lambda_j}} \qquad (11)$$

が得られる．$\Lambda = \{\{\lambda_j\}_{j=1}^J, \{\mu_k\}_{k=1}^K\}$ はラグランジュ乗数であり，λ_j は j 番目の構成要素モデルの結合の重みを，μ_k はクラス k の出現の偏りを与える．提案法では，$R(k|\boldsymbol{x};\hat{\Theta},\Lambda)$ を分類器の識別関数として与える．

ラグランジュ乗数 Λ の値は，$R(k|\boldsymbol{x};\hat{\Theta},\Lambda)$ による訓練データの対数尤度を最大化する Λ と一致することが知られている[1),12)]．しかし，Λ の推定に $\hat{\Theta}$ の計算と同一の訓練データ集合 D を用いると過学習を引き起こす危険性がある．そこで，訓練データの leave-one-out 交差検定法に基づき Λ を推定する[14)]．具体的には，訓練データサンプル $(\boldsymbol{x}_n, y_n)$ のクラス事後確率を，そのサンプルを除外して学習した構成要素モデルのパラメータ値 $\hat{\Theta}^{(-n)}$ を用いて見積もることで得られる，以下の目的関数 $CV(\Lambda)$ の最大化により Λ を推定する．

$$CV(\Lambda) = \sum_{n=1}^{N} \log R(y_n|\boldsymbol{x}_n;\hat{\Theta}^{(-n)},\Lambda) + \log p(\Lambda) \qquad (12)$$

提案法では，Λ の事前確率分布 $p(\Lambda)$ にガウス事前確率分布[5)] を適用し，$p(\Lambda) \propto \prod_{j=1}^{J} \exp\{-(\lambda_j - 1)^2/2\sigma_j^2\} \prod_{k=1}^{K} \exp(-\mu_k^2/2\rho_k^2)$ とする．$CV(\Lambda)$ は上に凸な関数であるため，Λ の学習では大域的最適性が保証される．Λ の推定値は，準ニュートン法の 1 種である L-BFGS アルゴリズム[10)] を用いて計算する．

訓練データ集合 $D = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
1. 式 (8) による $\hat{\Theta}$ の計算．
2. $(\boldsymbol{x}_n, y_n)$ を除外した部分訓練データ集合の式 (8) への適用による $\hat{\Theta}^{(-n)}, \forall n$ の計算．
3. 式 (12) による Λ の計算．
4. $R(k|\boldsymbol{x};\hat{\Theta},\hat{\Lambda})$ の出力．

図 2 モデルパラメータの学習アルゴリズム
Fig. 2 Algorithm for learning model parameters.

図 2 にモデルパラメータの学習アルゴリズムをまとめる．

3.1.3 クラス事後確率モデルの考察

ME 原理から誘導される $R(k|\boldsymbol{x};\hat{\Theta},\Lambda)$ は，生成アプローチのクラス事後確率 $P(k|\boldsymbol{x};\Theta)$ と 2 クラスのハイブリッド分類器のクラス事後確率 $R(k|\boldsymbol{x};\Theta,B)$ を特殊な場合として含む分布となっている．仮に，ラグランジュ乗数の値が $\lambda_j = 1, \forall j$，$\mu_k = P(k), \forall k$ であるとすると，式 (11) の右辺は，式 (3) の右辺と一致する．また，$K = 2$ のとき，$\lambda_j = b_j, \forall j$，$\mu_2 - \mu_1 = b_0$ とすれば，式 (11) の右辺は式 (7) の右辺と一致する．したがって，ME 原理から誘導されるハイブリッド分類器は生成アプローチによる分類器と 2 クラスのハイブリッド分類器の自然な拡張になっているといえる．

3.2 文書，Web ページ分類への適用

提案法を文書，Web ページ分類に適用するために，NB モデル[17)] を用いて構成要素をモデル化する．文書分類では本文やタイトルなどの単語列からなる構成要素を用いて，また Web ページ分類では単語列からなる構成要素とリンク情報からなる構成要素を用いて分類器を設計する．

NB モデルでは，構成要素に含まれる素性（単語，リンク）が独立に生起すると見なし，構成要素の特徴ベクトルを素性の頻度ベクトル $\boldsymbol{x}^j = (x_1^j, \ldots, x_i^j, \ldots, x_{V_j}^j)$ で表す．ここで，x_i^j は素性 i が構成要素中で出現した頻度を表し，V_j は構成要素 j で出現する素性の種類の総数を表す．そして，クラスラベルが k であるデータの構成要素 j の特徴ベクトル $\boldsymbol{x}^j$ の生成確率が，

$$P(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j) \propto \prod_{i=1}^{V_j} \left(\theta_{ki}^j\right)^{x_i^j} \qquad (13)$$

で表される多項分布に従うと仮定する．ここで，$\theta_{ki}^j > 0$ はクラスラベルが k のデータでの構成要素 j の素性 i の生起確率を表し，$\sum_{i=1}^{V_j} \theta_{ki}^j = 1$ を満たす未知パラメータである．

提案法では，構成要素モデル $p(\boldsymbol{x}^j|k;\boldsymbol{\theta}_k^j)$ に NB モデルを適用するのに際して，構成要素の特徴ベクトル $\boldsymbol{x}^j$ を $|\boldsymbol{x}^j| = \sum_{i=1}^{V_j} x_i^j = 1$ となるように正規化する．

構成要素モデルのパラメータ学習は，式 (8) の目的関数を用いて MAP 推定により行う．その際，パラメータの事前確率分布 $p(\boldsymbol{\theta}_k^j)$ には，NB モデルでしばしば仮定されるディリクレ分布[17] $p(\boldsymbol{\theta}_k^j) \propto \prod_{i=1}^{V_j} (\theta_{ki}^j)^{\xi_k^j - 1}$ を用いる．ξ_k^j はハイパーパラメータである．提案法では，ξ_k^j の値を，訓練データサンプル $(\boldsymbol{x}_n, y_n) \in D$ の leave-one-out 交差検定法により推定される構成要素モデルの対数尤度 $\log p(\boldsymbol{x}_n^j | y_n; \boldsymbol{\theta}_{y_n}^{j,(-n)})$ を最大化するように調節する．この調節は，期待値最大化 (EM) アルゴリズム[7] を援用して効率的に行う．

4. 評価実験

4.1 テストコレクション

文書と Web ページの分類問題に対して，ベンチマークテストによく用いられる 20 Newsgroups (20news) と WebKB の 2 つのテストコレクション[12] と，NIPS☆のデータを用いて，評価実験を行った．

20news は，UseNet の 18,828 記事を集めたものであり，20 グループに分類されている．実験では，これらの記事の分類器を設計するのに，“Subject:” のあとに続くタイトル (T) とそれ以外の本文 (M) の 2 つ構成要素を用いた．各記事から M と T の構成要素をそれぞれ抽出し，各語彙の出現頻度を表す特徴ベクトルを作成した．ただし，冠詞などの文書を特徴づける効果を持たない停止語 (stop words)[15] と各構成要素で 1 つの記事にしか出現しない低頻度語彙を除去した．構成要素の語彙数 (特徴ベクトルの次元) はそれぞれ 19,273 (M) と 1,775 (T) であった．

NIPS は，Yann が第 1–12 回の国際会議 NIPS で発表された 1,740 論文を OCR で電子化して作成したデータセットである．実験では，第 5–12 回の会議で発表された 1,164 論文を用いた．これらの論文は 9 分野に分類される．構成要素として，各論文に含まれる本文 (M)，タイトル (T)，アブストラクト (A)，参考文献 (R) の 4 つを用いた．これらの構成要素に対して，20news と同様の方法で停止語と低頻度語彙を除去して特徴ベクトルを作成した．特徴ベクトルの次元は，それぞれ 20,485 (M)，904 (T)，5,021 (A)，8,303 (R) であった．

WebKB は大学の Web ページ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 12) の設定に従い，**student**，**faculty**，**course**，**project** の 4 つのカテゴリに含まれる 4,199 の Web ページを実験に用いた．実験では，本文 (M)，タイトル (T)，他のページからのリンク (IL)，他のページへのリンク (OL)，画像などのファイルへのリンク (FL)，アンカーテキスト (AT) の 6 つの構成要素を用いて分類器を設計した．IL はこのテストコレクションに含まれる Web ページからのリンクのみを抽出し，AT として当該ページをリンクしている他のページが参照に用いているアンカーテキストを抽出した．M，T，AT は単語を素性として含むのに対し，IL，OL，FL の素性は URL である．これらの構成要素に対して，20news と同様の方法で停止語と低頻度語彙 (URL) を除去して特徴ベクトルを作成した．特徴ベクトルの次元は，それぞれ 18,471 (M)，995 (T)，496 (AT) 500 (IL)，4,131 (OL)，484 (FL) であった．

☆ http://www.cs.toronto.edu/~roweis/data.html

4.2 評価方法

提案法の分類性能を生成，識別アプローチに基づく分類器と比較することで評価する．比較のための尺度として，テストデータに対する分類精度を用いる．実験では，各データセットで，訓練データとテストデータをランダムに選択して評価セットを作成した．評価セット作成のために選択したテストデータの数は，20news，NIPS，WebKB でそれぞれ 8,000，500，2,000 であり，訓練データはテストデータ以外から選択した．この評価セットを 10 通り作成し，10 回の実験で得られる分類精度の平均値で提案法と他手法を比較した．

4.3 実験結果

4.3.1 付加情報の効果

付加情報を同時に扱う効果を確かめるために，提案法を，本文 (M) のみを用いた NB モデル (NB-M) と MLR モデル (MLR-M) による分類器と比較した．**図 3** に，提案法と NB-M，MLR-M の分類精度の比較結果を示す．提案法では，構成要素 M と 4.1 節で述べた付加的な構成要素をすべて用いている．図中の各グラフは，訓練データ数を変えて分類器を学習したときの分類精度を示す．図 3 より，20news，NIPS，WebKB のすべてのテストコレクションで，訓練データ数によらず，提案法の分類精度が NB-M と MLR-M より明らかに高い結果が得られた．これは，付加情報を同時に扱うことが分類精度の向上に大きく寄与することを示している．

4.3.2 複数の構成要素を扱う分類器の比較

次に，提案法を，2 章で述べた生成，識別の各アプローチによる複数の構成要素を扱う分類器と比較した．比較に用いた分類器は以下の 4 つである．

- PNB (生成)：式 (3) に基づき構成要素ごとに生成モデル $p(\boldsymbol{x}^j | y; \boldsymbol{\theta}_y^j)$ を仮定して分類器を設計．テキスト，リンクの各構成要素の生成モデルとし

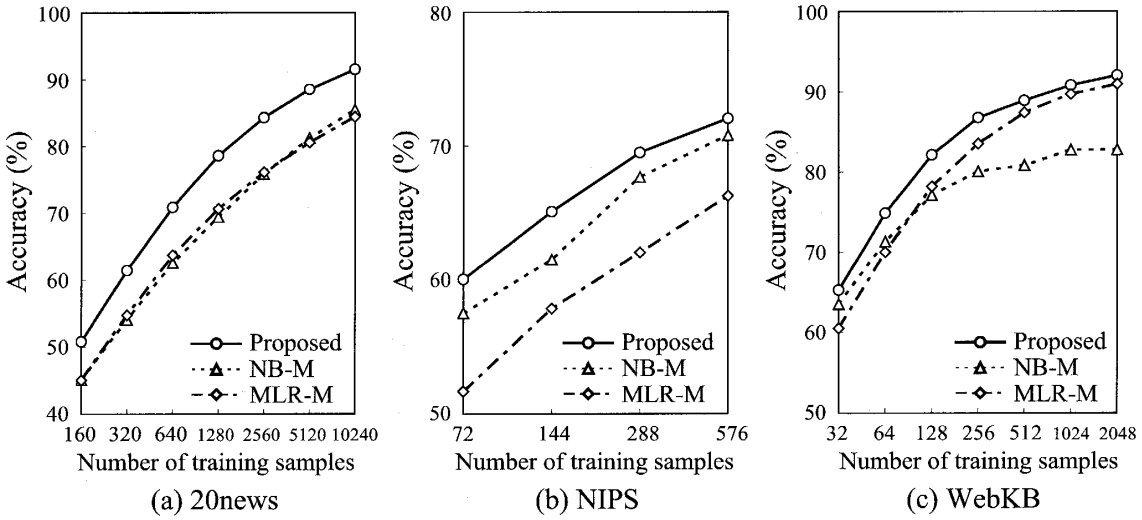


図 3 提案法と NB-M, MLR-M の分類精度 (%)
Fig. 3 Classification accuracies (%) with the proposed method, NB-M and MLR-M.

て NB モデルを使用.

- PMLR（識別）：構成要素ごとに MLR モデル $P(k|\boldsymbol{x}^j)$ を設計. MLR モデルの積を最大化する k をデータ $\boldsymbol{x}$ のクラス y として選択（2.2 節参照）.
- SNB（生成）：構成要素ごとではなくデータ $\boldsymbol{x}$ に単一のモデル $p(\boldsymbol{x}|y;\boldsymbol{\theta}_y)$ を仮定. モデルとして NB モデルを使用☆.
- SMLR（識別）：MLR を用いてデータ $\boldsymbol{x}$ のクラス事後確率 $P(k|\boldsymbol{x};W)$ を直接モデル化（式 (4)）.

表 1 に，提案法と 4 つの比較手法の分類精度を示す. 表中の数値は，10 通りの評価セットを用いた実験で得られた分類精度の平均値（括弧内は標準偏差）である. $|D|$ は訓練データ数を表す.

20news と NIPS では，すべての場合で提案法の分類精度が 4 つの比較手法を上回った. WebKB では，全体的に提案法の分類精度が最も高い傾向があったが，$|D| = 2{,}048$ のときは SMLR の分類精度が提案法を上回った. このとき，SMLR の分類精度は SNB, PNB よりも約 6%大きかった. これは，生成，識別アプローチの分類精度の差が大きい場合，一方のアプローチの短所の影響を受ける危険性があることを示唆している. しかし，実験では，生成，識別アプローチの分類精度の差が小さいほど，提案法の分類精度が両アプローチを大きく上回る傾向がみられた. この結果より，提案法では，生成，識別の両アプローチの長所を効果的に取り込んだハイブリッド分類器を実現しているといえる.

☆ 単一の NB モデルは生成モデルとして適切とはいえないが，分類器としての有用性を確認するために比較対象に加えた.

表 1 提案法と PNB, PMLR, SNB, SMLR の分類精度 (%)
Table 1 Classification accuracies (%) with the proposed method, PNB, PMLR, SNB and SMLR.

(a) 20news

$\|D\|$	Proposed	PNB	PMLR	SNB	SMLR
160	**50.8** (2.0)	47.4 (1.8)	47.8 (1.4)	47.5 (1.8)	46.7 (1.8)
320	**61.4** (1.4)	56.9 (1.3)	57.4 (1.6)	56.9 (1.3)	56.8 (1.8)
640	**70.9** (0.9)	66.0 (1.2)	66.6 (0.7)	66.0 (1.1)	65.9 (1.1)
1280	**78.6** (0.5)	73.3 (0.8)	74.7 (0.4)	73.3 (0.9)	73.3 (0.5)
2560	**84.3** (0.5)	79.8 (0.6)	80.9 (0.4)	79.8 (0.6)	78.8 (0.6)
5120	**88.5** (0.2)	85.0 (0.4)	85.6 (0.3)	85.0 (0.4)	83.3 (0.4)
10240	**91.5** (0.2)	88.9 (0.3)	89.5 (0.2)	88.9 (0.3)	87.3 (0.2)

(b) NIPS

$\|D\|$	Proposed	PNB	PMLR	SNB	SMLR
72	**60.0** (2.8)	58.4 (2.8)	51.5 (2.4)	57.7 (2.5)	51.5 (2.9)
144	**65.1** (1.6)	63.0 (1.8)	57.5 (2.4)	62.6 (2.2)	58.2 (2.5)
288	**69.5** (1.8)	68.5 (1.6)	62.6 (1.8)	68.6 (1.5)	62.6 (1.9)
576	**72.0** (1.9)	71.5 (1.7)	65.8 (1.6)	71.6 (1.6)	66.6 (1.9)

(c) WebKB

$\|D\|$	Proposed	PNB	PMLR	SNB	SMLR
32	**65.3** (2.9)	64.5 (4.6)	61.7 (6.1)	63.5 (4.6)	60.4 (5.9)
64	**74.9** (2.7)	72.4 (4.2)	70.9 (3.3)	71.8 (4.5)	69.9 (4.2)
128	**82.1** (1.6)	78.7 (2.4)	78.1 (1.3)	78.3 (2.6)	78.5 (2.3)
256	**86.8** (1.1)	82.0 (1.6)	82.6 (0.4)	81.9 (1.7)	84.2 (1.3)
512	**88.9** (0.7)	83.6 (2.8)	85.8 (0.8)	83.6 (2.7)	88.1 (0.9)
1024	**90.8** (0.7)	85.7 (1.6)	88.1 (0.6)	85.8 (1.5)	**90.8** (0.5)
2048	92.1 (0.4)	86.6 (0.8)	89.2 (0.4)	86.5 (0.8)	**92.6** (0.4)

4.4 考　　察

提案法による構成要素の結合の有効性を確認するため，構成要素 j の結合の重み λ_j の推定値を調べた. **図 4** に，訓練データ数を 10,240 (20news)，576 (NIPS)，2,048 (WebKB) としたときに提案法で推定された構成要素 j の結合の重み λ_j と，構成要素ごとに NB モデルで分類器を設計したときのテストデータに対する分類精度を示す. ただし，各分類精度は構

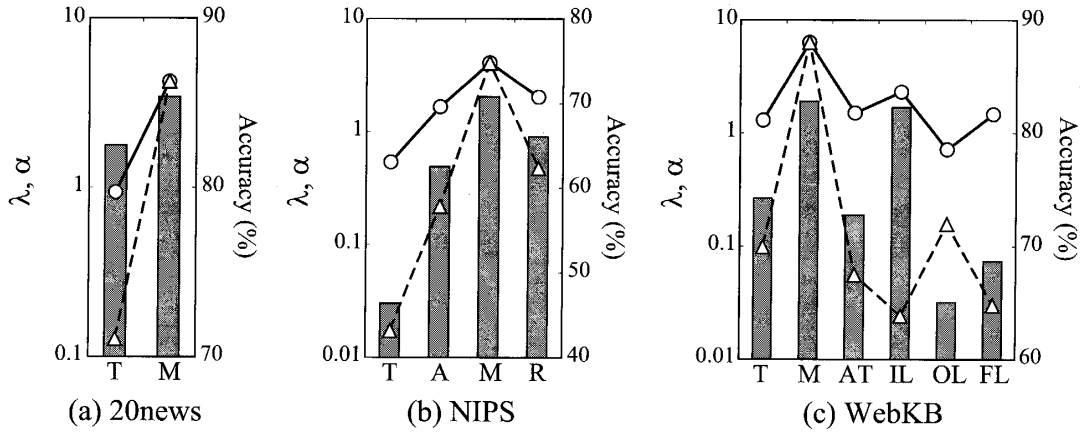


図 4 提案法で推定された構成要素の結合の重み λ (○), 構成要素の大きさ α (△) と各構成要素の NB モデルの分類精度 (%, 棒グラフ)

Fig. 4 Combination weights of components estimated with the proposed method, λ (○), component size α (△) and classification accuracies (%) of component NB models (bars).

表 2 各構成要素に含まれる素性の総数 $|F|$

Table 2 The total number of features contained in each component, $|F|$.

(a) 20news

Component	T	M
$\|F\|$	59261	1960166

(b) NIPS

Component	T	A	M	R
$\|F\|$	5903	74385	1386484	161992

(c) WebKB

Component	T	M	AT	IL	OL	FL
$\|F\|$	10403	668192	5882	2547	16535	3165

成要素が空のテストデータを除外して算出している. また, 構成要素の大きさを比較するために, 構成要素 j に含まれる素性 (単語, URL) の総数 $|F|_j$ (**表 2** 参照) を, $\alpha_j = \lambda_M |F|_j / |F|_M$ で規格化してプロットしている. ただし, λ_M と $|F|_M$ は構成要素 M の λ と $|F|$ を表す.

図 4 より, すべてのテストコレクションで, 構成要素 M の NB モデルの分類精度が他の構成要素よりも高く, λ が最も大きく推定されていたことが分かる. また, NB モデルの分類精度の高い構成要素ほど λ が大きい傾向がみられた. とくに, WebKB の IL は α が小さいにもかかわらず, λ が大きく推定されていた. これは, 提案法による構成要素モデルの結合が, 構成要素に含まれる素性の数ではなく, 分類精度を基準に与えられていることを示唆している. 提案法では, 分類能力が優れる構成要素ほど大きな重みで結合して高い分類精度を実現していると考えられる.

5. 関連研究

機械翻訳や音声認識の分野では, 複数のモデルを識別的な基準により結合して分類器を設計する方法が提案されている. 文献 13) では, 翻訳器に含まれる翻訳モデルと言語モデルを, ME 原理に基づいて結合することで重みの調節を行っている. 音声認識の分野では, 音響や言語などの様々なモデルの識別的な結合に基づく認識器が提案されている[2),8)]. それに対して, 提案法では, 文書と Web ページの分類のために, それらに含まれる構成要素を個別に NB モデルを用いてモデル化し, ME 原理に基づいて構成要素モデルを結合することで分類器を構築している.

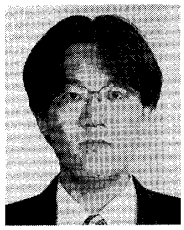
6. まとめ

本論文では, 複数の付加情報を含むテキストデータの多クラス分類問題に対して, 生成, 識別アプローチのハイブリッドに基づく分類器設計法を提案した. 従来のハイブリッド分類器が 2 クラス問題を対象とするのに対して, 提案法では, 最大エントロピー (ME) 原理に基づいて本文と付加情報の生成モデルを結合することにより, 多クラス (3 クラス以上) 問題を直接扱うハイブリッド分類器を構築する. 文書と Web ページの分類実験により, 本文と付加情報を同時に用いて分類器を構築する効果を確かめるとともに, 提案法が, 生成, 識別アプローチでの分類精度の差が小さい場合に, とくに有用であることを確認した. また, 提案法では, 構成要素モデルを個々の分類能力に応じた重みで結合して, 高い分類精度を得ていることを確認した.

参考文献

1) Berger, A.L., Della Pietra, S.A. and Della Pietra, V.J.: A maximum entropy approach to natural language processing, *Computational Linguistics*, Vol.22, No.1, pp.39–71 (1996).
2) Beyerlein, P.: Discriminative model combination, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP1998)*, pp.481–484 (1998).
3) Brochu, E. and Freitas, N.: "Name that song!": A probabilistic approach to querying on music and text, *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, MA, pp.1505–1512 (2003).
4) Chakrabarti, S., Dom, B. and Indyk, P.: Enhanced hypertext categorization using hyperlinks, *Proc. ACM International Conference on Management of Data (SIGMOD-98)*, pp.307–318 (1998).
5) Chen, S.F. and Rosenfeld, R.: A Gaussian prior for smoothing maximum entropy models, Technical report, Carnegie Mellon University (1999).
6) Cohn, D. and Hofmann, T.: The missing link - A probabilistic model of document content and hypertext connectivity, *Advances in Neural Information Processing Systems 13*, MIT Press, Cambridge, MA, pp.430–436 (2001).
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) Glotin, H., Vergyri, D., Neti, C., Potamianos, G. and Luettin, J.: Weighting schemes for audio-visual fusion in speech recognition, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2001)*, pp.165–168 (2001).
9) Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York Berlin Heidelberg (2001).
10) 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).
11) Lu, Q. and Getoor, L.: Link-based text classification, *IJCAI Workshop on Text-Mining & Link-Analysis (TextLink 2003)* (2003).
12) Nigam, K., Lafferty, J. and McCallum, A.: Using maximum entropy for text classification, *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp.61–67 (1999).
13) Och, F.J. and Ney, H.: Discriminative training and maximum entropy models for statistical machine translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pp.295–302 (2002).
14) Raina, R., Shen, Y., Ng, A.Y. and McCallum, A.: Classification with hybrid generative/discriminative models, *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA (2004).
15) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, New York (1983).
16) Sun, A., Lim, E.P. and Ng, W.K.: Web classification using support vector machine, *Proc. 4th Int. Workshop on Web Information and Data Management (WIDM 2002) held in conj. with CIKM 2002*, pp.96–99 (2002).
17) 上田修功，斉藤和巳：多重トピックテキストの確率モデル—テキストモデル研究の最前線 (1)，(2)，情報処理，Vol.45, pp.184–190, 282–289 (2004).

(平成 18 年 1 月 20 日受付)
(平成 18 年 7 月 4 日採録)

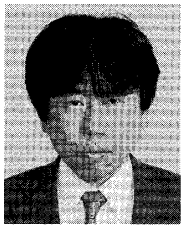


藤野 昭典（正会員）

1972 年生．1995 年京都大学工学部精密工学科卒業．1997 年同大学大学院工学研究科精密工学専攻修士課程修了．同年 NTT 入社．機械学習等の研究に従事．現在，NTT コミュニケーション科学基礎研究所に所属．電子情報通信学会 PRMU 研究奨励賞（2004 年度），FIT 論文賞（2005 年）各受賞．電子情報通信学会会員．

上田 修功（正会員）

1958 年生．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協創情報研究部長，奈良先端大客員教授．電気通信普及財団賞（1997 年，2006 年），電子情報通信学会論文賞（2002 年，2004 年）等受賞，電子情報通信学会，日本神経回路学会，IEEE 各会員．



斉藤 和巳（正会員）

1963年生．1985年慶應義塾大学理工学部数理科学科卒業．工学博士．同年NTT入社．1991年より1年間オタワ大学客員研究員．神経回路網，機械学習，複雑ネットワーク等の研究に従事．現在，NTTコミュニケーション科学基礎研究所主任研究員（特別研究員），奈良先端大客員助教授．情報処理学会論文賞（1997年），人工知能学会論文賞（1999年）等受賞．電子情報通信学会，人工知能学会，日本神経回路学会，IEEE各会員．