Vol.2012-MPS-88 No.9
2012/5/17

ベイズモデルに基づく判別特徴選択

田中 佑典[†1,a)]　上田 修功[†2]　田中 利幸[†1]

概要：本論文では，カテゴリカルデータを対象とし，ベイズモデルに基づいてクラス毎に固有な特徴選択およびクラス識別を行う手法を提案する．提案モデルでは，各特徴が全クラス共通の特徴とクラス固有の特徴のいずれであるかを示す特徴選択のための潜在変数を導入し，ベイズ推定により訓練データからその値を推定する．これにより，従来の特徴選択手法とは異なり，クラス毎に識別に有効な特徴を自動判別することが可能となる．人工データおよび DNA データを用いた実験で，従来の特徴選択手法に対し，提案手法が，汎化性能の観点で優れていることを示す．また，間接的に特徴選択を実現する Lasso やサポートベクトルマシンとの比較により，提案手法が識別器としても顕著に優れていることを示す．

キーワード：ベイズモデル，特徴選択，クラス識別

Bayesian Discriminant Feature Selection

Tanaka Yusuke[†1,a)]　Ueda Naonori[†2]　Tanaka Toshiyuki[†1]

Abstract: Focusing on categorical data, we propose a Bayesian feature selection method in which a set of class-specific features are selected for each class for improving the generalization ability of classification. In the proposed model, we introduce latent variable to each feature and each class to decide whether the feature is specific for the class or common in all classes. The latent variables are estimated from given training data by the framework of the Bayesian inference. Unlike the conventional feature selection methods, this enables us to obtain class dependent subset features which are effective for classification. We demonstrate that the proposed method can be superior to the conventional methods in terms of generalization ability through experiments with synthetic and real DNA data sets. We also show that the proposed method can obtain higher classification accuracy than Lasso and Support Vector Machine which indirectly realize feature selection.

Keywords: Bayesian models, feature selection, classification

1. はじめに

パターン認識では，通常，入力パターンを特徴ベクトルとして表現する．近年，遺伝子情報処理やセンサー情報処理など，観測データの質的な多様性にともない，次元数の大きな特徴ベクトルを取扱う機会が増大している．識別に無関係な冗長な特徴 (特徴ベクトルの要素) が存在すると，一般に，テストデータでの識別性能 (汎化性能) が低下する．それ故，高次元特徴ベクトルから識別に有効な特徴を選択することが汎化性能向上の観点で実用上重要となる．また，クラス固有の特徴を抽出，考察できるという点でも特徴選択は有用である．

Guyon ら [1] は, パターン認識における特徴選択のアプローチを大きく 3 通りに分類している．まず，使用する識別器をブラックボックスとみなし，ある部分特徴を選択したときに性能を評価する手続きを導入することにより特徴選択を行うというアプローチがある．このような性能評価のための手続きをラッパー (wrappers) と呼ぶ．性能評価には, 交差検定法による汎化性能の推定値などがよく用いられる．一方で, 特徴の組み合わせを列挙すると膨大な数

[†1] 現在，京都大学
Presently with Kyoto Uniersity
[†2] 現在，NTT コミュニケーション科学基礎研究所
Presently with NTT Comunication Science Laboratories
[a)] ytanaka@sys.i.kyoto-u.ac.jp

にのぼるため, ラッパーに基づくアプローチは一般的に計算時間の面で実用的なアプローチとは言えない．次に，使用する識別器とは独立に，前処理として特徴選択をするためにフィルタ (filters) を用いるアプローチがある．フィルタに基づく特徴選択手法には，各特徴とラベルの相関に基づく手法 [2] や，各特徴とラベルの相互情報量に基づく手法 [3] などがある．このアプローチは，計算コストの面で実用的であるが，発見的な閾値設定が必要となる．最後に，識別器の学習に特徴選択の機能を組み込むアプローチがあり, [1] では組み込み法 (embedded methods) と総称されている．組み込み法を用いた研究では，サポートベクトルマシンの重みベクトルの変化を基準にして特徴選択を行う SVM-RFE 法 [4] などが提案されている．

特徴選択に関する既存研究の多くは，二クラス識別問題を対象としている．一方で，多クラス識別の問題では，各クラスを特徴づける固有の特徴がそれぞれ存在すると考えられるため，クラス毎に異なるクラス固有の特徴を選択することが性能向上のために有用であると考えられる．これは, クラス数が多い識別問題ではより重要である．実際，クラスタリング問題では，近年，全ての特徴ではなく，部分特徴でクラスタリングを行う bi-clustering 法 [5] や subset clustering 法 [6]，[7] が提案され，サンプルのクラスタリングのみならず，クラスタ固有の特徴のクラスタリングも同時に実現されている．

本論文では，識別問題における汎化性能向上を目的として，この特徴選択型クラスタリングの考え方をクラス識別の問題でモデル化したベイズ判別特徴選択手法を提案する．提案手法は，ベイズモデルに基づき，カテゴリカルデータ (特徴の値が離散値) を対象とする点で上記の subset clustering 法 [6] と共通するが，文献 [6] では 2 値特徴を対象としているのに対し，提案モデルではより一般的な多値特徴とし，かつ，識別問題を対象とするという点で異なり，識別問題でのこのようなアプローチはこれまで提案されていない．提案モデルでは，特徴選択とクラス識別を独立タスクとするのではなく，後述するように特徴選択とクラス識別を同じベイズモデルの枠組みで実現している．従って, 提案モデルは組み込み法の一種であると考えることもできる．人工データ，および実データを用いた実験により，従来手法との比較を行い提案手法の有用性を示す．また，特徴選択手法ではないが間接的に特徴選択を実現する Lasso 法 [8] や，特徴変換に基づく代表的な識別器であるサポートベクトルマシンとの識別性能比較も行い，提案手法の識別器としての性能評価も考察する．

2. ベイズモデルに基づく判別特徴選択

2.1 基本的な考え方

本論文では，図 1 に示すような特徴の値がカテゴリ（離散シンボル）からなるカテゴリカルデータを対象とする．

図 1 カテゴリカルデータの例 (DNA データ)．各クラス 5 つの DNA データを示す．色付きの部分はクラスに関連した特徴を表す．

Fig. 1 Examples of the categorical data (DNA data). Five DNA samples are represented respectively. Color part is the features related to the class.

DNA データ (4 種類の塩基シンボルデータ) などはその典型例である．カテゴリカルデータの場合，図 1 に示すように，同一クラスに属するサンプルの大多数が同じ値をとる特徴は明らかにクラスに関連する特徴と言える．逆に，同一クラスに属するサンプルの値が類似していない特徴はそのクラスに関連しない特徴と言える．当然ながら，あるクラスに関連する特徴が別のクラスでも関連する特徴であるとは限らない．

提案モデルでは，各特徴があるクラスに関連する特徴か否かを示す潜在変数を導入する．$x_j^{(k)}$ をクラス k サンプルの特徴ベクトルの第 j 特徴 (第 j 次元) とし，その値は $1,\ldots,L$ のいずれかのカテゴリカルな値をとるものとする．この時，第 j 特徴がクラス k に関係する (relevant) か否かを，潜在変数 $r_{k,j} \in \{1,0\}$ で表現し，relevant な場合 ($r_{k,j}=1$)，$x_j^{(k)}$ はクラス k 固有の多項分布 (L 項分布) から生成され，そうでない場合 ($r_{k,j}=0$)，$x_j^{(k)}$ は全クラスに共通な多項分布 (L 項分布) から生成されるものと仮定する．

2.2 生成モデル

クラス k に属するサンプルを $\mathrm{X}^{(k)} = \{x_{i,j}^{(k)}\}(i = 1,\ldots,N_k, j = 1,\ldots,M)$ とする．但し, $N_k(k=1,\ldots,K)$ はクラス k に属するサンプル数, M は次元数を表す．また, シンボル $x_{i,j}^{(k)}$ は L 種類の離散シンボル $\{1,\ldots,L\}$ のいずれかの値をとる．提案モデルでは, サンプル $x_{i,j}^{(k)}$ が, 2 種類の多項分布から生成されると仮定する．すなわち, 全データに共通の分布とクラス固有の分布である．全データに共通の分布の多項分布パラメータを $\boldsymbol{\phi} = \{\phi_1,\ldots,\phi_L\}$ とし, クラス毎に固有の多項分布のパラメータを $\boldsymbol{\theta}_{k,j} = \{\theta_{k,j,1},\ldots,\theta_{k,j,L}\}$ とする．ここで, ϕ_l はシンボル l が生起する確率を, $\theta_{k,j,l}$ はクラス k の特徴 j においてシンボル l が生起する確率を表す．明らかに, $\sum_{l=1}^{L}\phi_l = 1$, $\sum_{l=1}^{L}\theta_{k,j,l} = 1$ を満たす．また, 問題によっては特徴毎にとり得るシンボルの種類数が異なる, つまり, L が j に依存する場合も考えられる．その

際, $\boldsymbol{\phi}$ の代わりに, $\boldsymbol{\phi}_j = \{\phi_{j,1}, \ldots, \phi_{j,L_j}\}$ とすればよい. 簡単のため, 以下の定式化では, 全特徴における L は等しいものとするが, 特徴毎に L が異なる場合に以下の定式化は容易に拡張できる.

次いで, サンプルがいずれの分布から生成されるかを示す潜在変数 $\boldsymbol{r}_k \in \{0,1\}^M$ を導入する. 以下に示すように, 潜在変数 $r_{k,j}$ は, パラメータ λ を持つベルヌーイ分布から生成され, $r_{k,j}$ 値に依存して, サンプル $x_{i,j}^{(k)}$ は多項分布により生成されるものと仮定する. λ はクラスや特徴に依らないものとする.

$$
\begin{aligned}
r_{k,j}|\lambda &\sim \mathrm{Bernoulli}(r_{k,j};\lambda),\\
x_{ij}^{(k)}|r_{k,j},\boldsymbol{\phi},\boldsymbol{\theta}_{k,j} &\sim \begin{cases} \mathrm{Multinomial}(x_{i,j}^{(k)};\boldsymbol{\theta}_{k,j}), r_{k,j}=1 \\ \mathrm{Multinomial}(x_{i,j}^{(k)};\boldsymbol{\phi}), \quad r_{k,j}=0 \end{cases}.
\end{aligned}
$$

また, 多項分布のパラメータの事前分布として, ディリクレ分布を用い, ベルヌーイ分布のパラメータの事前分布としてベータ分布を用いる. これらは各分布に対する共役事前分布である. 各パラメータは以下の確率に従って生成される.

$$
\begin{aligned}
\lambda|a,b &\sim \mathrm{Beta}(\lambda;a,b),\\
\boldsymbol{\phi}|\boldsymbol{\alpha} &\sim \mathrm{Dirichlet}(\boldsymbol{\phi};\boldsymbol{\alpha}),\\
\boldsymbol{\theta}_{k,j}|\boldsymbol{\beta}_{k,j} &\sim \mathrm{Dirichlet}(\boldsymbol{\theta}_{k,j};\boldsymbol{\beta}_{k,j}).
\end{aligned}
$$

但し, a, b はベータ分布のハイパーパラメータである. また, $\boldsymbol{\alpha}$ はデータ共通の多項分布のパラメータ $\boldsymbol{\phi}$ に対するディリクレ事前分布のハイパーパラメータであり, $\boldsymbol{\beta}_{k,j}$ はクラス固有の多項分布のパラメータ $\boldsymbol{\theta}_{k,j}$ に対するディリクレ事前分布のハイパーパラメータである.

以上の生成モデルに従って, 全サンプル $\mathrm{X} = \{\mathrm{X}^{(1)}, \ldots, \mathrm{X}^{(K)}\}$ の尤度関数は以下となる.

$$
\begin{aligned}
p(\mathrm{X}|\mathrm{R},\boldsymbol{\phi},\Theta,\lambda) &= \prod_{k=1}^{K}\prod_{i=1}^{N_k}\prod_{j=1}^{M}\prod_{l=1}^{L}\left(\theta_{k,j,l}\phi_l^{1-r_{k,j}}\right)^{I(x_{i,j}^{(k)}=l)}\\
&= \prod_{k=1}^{K}\prod_{j=1}^{M}\prod_{l=1}^{L}\left(\theta_{k,j,l}^{r_{k,j}}\phi_l^{1-r_{k,j}}\right)^{n_{j,l}^{(k)}} \qquad (1)
\end{aligned}
$$

但し, $I(f)$ は f が真であれば値 1, 偽であれば値 0 となる関数を表し, $n_{j,l}^{(k)} = \sum_{i=1}^{N_k} I(x_{i,j}^{(k)} = l)$ と書くことにする. また, $\mathrm{R} = \{r_{k,j}\}(k = 1, \ldots, K, j = 1, \ldots, M)$, $\Theta = \{\boldsymbol{\theta}_{k,j}\}(k = 1, \ldots, K, j = 1, \ldots, M)$ を表す.

2.3 学習

ベイズ学習では, 観測データを与件として全ての未知量 (確率変数) の事後分布 $p(\mathrm{R},\boldsymbol{\phi},\Theta,\lambda|\mathrm{X}) \propto p(\mathrm{X}|\mathrm{R},\boldsymbol{\phi},\Theta)p(\mathrm{R}|\lambda)$ を推定する. さらに, 共役事前分布を用いているため, 次式の様にパラメータを積分消去でき, より効率的な推定が可能となる.

$$
\begin{aligned}
p(\mathrm{X}|\mathrm{R}) &= \int p(\mathrm{X}|\mathrm{R},\boldsymbol{\phi},\Theta)p(\boldsymbol{\phi})p(\Theta)\mathrm{d}\boldsymbol{\phi}\mathrm{d}\Theta\\
&= \left(\prod_{k=1}^{K}\prod_{j=1}^{M}\frac{1}{\mathrm{B}(\beta_{k,j,\bullet})}\frac{\prod_l \Gamma(I(r_{k,j}=1)n_{j,l}^{(k)}+\beta_{k,j,l})}{\Gamma(\sum_l(I(r_{k,j}=1)n_{j,l}^{(k)}+\beta_{k,j,l}))}\right)\\
&\times\left(\frac{1}{\mathrm{B}(\alpha_\bullet)}\frac{\prod_l \Gamma(\sum_k\sum_j I(r_{k,j}=0)n_{j,l}^{(k)}+\alpha_l)}{\Gamma(\sum_l(\sum_k\sum_j I(r_{k,j}=0)n_{j,l}^{(k)}+\alpha_l))}\right). \qquad (2)
\end{aligned}
$$

但し, $\mathrm{B}(\bullet)$ はディリクレ分布の規格化定数であり, $\beta_{k,j,\bullet} = \sum_{l=1}^{L}\beta_{k,j,l}$, $\alpha_\bullet = \sum_{l=1}^{L}\alpha_l$ とした. また, $\Gamma(\bullet)$ はガンマ関数を表す.

同様に,

$$
\begin{aligned}
p(\mathrm{R}) &= \prod_{k=1}^{K}\prod_{j=1}^{M}\int p(r_{k,j}|\lambda)p(\lambda;a,b)\mathrm{d}\lambda\\
&= \frac{1}{\mathrm{B}(a,b)}\frac{\Gamma(\sum_k\sum_j r_{k,j}+a)\Gamma(\sum_k\sum_j(1-r_{k,j})+b)}{\Gamma(KM+a+b)} \qquad (3)
\end{aligned}
$$

を得る. 但し, $\mathrm{B}(a,b)$ はベータ分布の規格化定数である. 式 (2) と式 (3) より, $p(\mathrm{R}|\mathrm{X})$ が得られる. しかし, この分布から直接 R をサンプリングすることは困難であるので, ギブスサンプリングにより推定する. 具体的には, $r_{k,j}$ 以外の R を $r_{\backslash(k,j)}$ と書くこととし, $r_{\backslash(k,j)}$ を既知とした full conditional 分布 $p(r_{k,j}|r_{\backslash(k,j)},\mathrm{X})$ に従ってサンプリングする. この操作を全ての k および j の組み合わせに対して繰り返し, これを一回のサイクルとして収束するまで十分に反復させる.

ギブスサンプリングの更新式について述べる. $r_{k,j} = 1$ となる条件付き確率と $r_{k,j} = 0$ となる条件付き確率の比を γ とし, 式 (2) と式 (3) を用いると以下のようになる.

$$
\begin{aligned}
\gamma &= \frac{p(r_{k,j}=1|r_{\backslash(k,j)},\mathrm{X})}{p(r_{k,j}=0|r_{\backslash(k,j)},\mathrm{X})}\\
&= \left(\prod_{l=1}^{L}\frac{\Gamma(n_{j,l}^{(k)}+\beta_{k,j,l})}{\Gamma(\beta_{k,j,l})}\right)\left(\frac{\Gamma(\sum_{l=1}^{L}\beta_{k,j,l})}{\Gamma(\sum_{l=1}^{L}(n_{j,l}^{(k)}+\beta_{k,j,l}))}\right)\\
&\times\left(\prod_{l=1}^{L}\frac{\Gamma(\sum_{(s,t)\neq(k,j)}I(r_{s,t}=0)n_{t,l}^{(s)}+\alpha_l)}{\Gamma(n_{j,l}^{(k)}+\sum_{(s,t)\neq(k,j)}I(r_{s,t}=0)n_{t,l}^{(s)}+\alpha_l)}\right)\\
&\times\left(\frac{\Gamma(\sum_{l=1}^{L}(n_{j,l}^{(k)}+\sum_{(s,t)\neq(k,j)}I(r_{s,t}=0)n_{t,l}^{(s)}+\alpha_l))}{\Gamma(\sum_{l=1}^{L}(\sum_{(s,t)\neq(k,j)}I(r_{s,t}=0)n_{t,l}^{(s)}+\alpha_l))}\right)\\
&\times\left(\frac{\sum_{(s,t)\neq(k,j)}I(r_{s,t}=1)+a}{\sum_{(s,t)\neq(k,j)}I(r_{s,t}=0)+b}\right). \qquad (4)
\end{aligned}
$$

一方, 明らかに $p(r_{k,j}=1|r_{\backslash(k,j)},\mathrm{X}) + p(r_{k,j}=0|r_{\backslash(k,j)},\mathrm{X}) = 1$ が成り立つので, この関係式と式 (4) より各条件付き確率は以下となる.

$$
\begin{aligned}
p(r_{k,j}=1|r_{\backslash(k,j)},\mathrm{X}) &= \frac{\gamma}{1+\gamma}, \qquad (5)\\
p(r_{k,j}=0|r_{\backslash(k,j),\mathrm{X}}) &= \frac{1}{1+\gamma}. \qquad (6)
\end{aligned}
$$

2.4 クラス識別

本節では, 提案モデルに基づき, クラスラベルが未知の

データ (テストデータ) のクラス識別について述べる．未知データ $\boldsymbol{x} = (x_1, \ldots, x_M)$ は, 観測データ X が与えられた下で, 次式のクラス事後予測分布が最大になるクラスに分類する．

$$k^* = \arg\max_k p(C_k|\boldsymbol{x}, \mathrm{X}) = \arg\max_k p(\boldsymbol{x}|C_k, \mathrm{X})p(C_k). \tag{7}$$

ここで, $p(C_k)$ はクラス事前分布であり，事前知識がなければ一様分布とするのが妥当である．後述する実験でも一様分布としている．また $p(\boldsymbol{x}|C_k, \mathrm{X})$ は,

$$p(\boldsymbol{x}|C_k, \mathrm{X}) = \prod_{j=1}^{M} \sum_{r_{k,j}} p(x_j|r_{k,j})p(r_{k,j}|\mathrm{X}) \tag{8}$$

$$\simeq \prod_{j=1}^{M} \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} p(x_j|r_{k,j}^{(t)}) \tag{9}$$

として，式 (8) での和計算をモンテカルロ近似により求める．これは, 式 (8) において $p(r_{k,j}|\mathrm{X})$ を計算することができないためである．また, 式 (9) の $r_{k,j}^{(t)}$ はギブスサンプリングにおいて t 回目の反復でサンプリングした値を表す．T はギブスサンプリングの反復の上限を表し, $t = 1, \ldots, t_0$ 回目の反復でサンプリングした $r_{k,j}$ は, ギブスサンプリングが収束していないと判断して棄却する．式 (9) における $p(x_j|r_{k,j}^{(t)})$ は, $r_{k,j}^{(t)}$ の値に応じて, 以下のように解析的に計算できる．

$$p(x_j|r_{k,j}^{(t)} = 1) = \int p(x_j|\boldsymbol{\theta}_{k,j})p(\boldsymbol{\theta}_{k,j}|\mathrm{X})\mathrm{d}\boldsymbol{\theta}_{k,j} = \frac{\prod_{l=1}^{L}(n_{j,l}^{(k)} + \beta_{k,j,l})^{I(x_j=l)}}{\sum_{l=1}^{L}(n_{j,l}^{(k)} + \beta_{k,j,l})}, \tag{10}$$

$$p(x_j|r_{k,j}^{(t)} = 0) = \int p(x_j|\boldsymbol{\phi})p(\boldsymbol{\phi}|\mathrm{X})\mathrm{d}\boldsymbol{\phi} = \frac{\prod_{l=1}^{L}(n_{j,l}^{(\bullet)} + \alpha_l)^{I(x_j=l)}}{\sum_{l=1}^{L}(n_{j,l}^{(\bullet)} + \alpha_l)}. \tag{11}$$

但し, $n_{j,l}^{(\bullet)} = \sum_{k=1}^{K} n_{j,l}^{(k)}$ とした．ここで, ハイパーパラメータを無視すると, 式 (10) はクラス k に属するサンプル数に対するクラス k かつ特徴 j において, シンボルが l であるサンプル数の割合を表している．また, 式 (11) は全データ数に対する特徴 j においてシンボルが l であるサンプル数の割合を表しており, 直観的に妥当な結果となっている．

3. 評価実験

3.1 比較手法

提案手法は，$r_{k,j} = 1$，すなわち全ての特徴が relevant と仮定すると，良く知られたナイーブベイズ (NB) モデルと等価となる．そこで，提案手法と NB との比較により，特徴選択の有効性を検証する．また， 全クラス共通に特徴選択を行う手法との比較により，クラス毎に異なった特徴選択を行う提案法の有効性を示す．全クラス共通に特徴選択を行う代表的な手法として, 交差検定法の推定値に基づく特徴の逆方向選択 [1] と比較した．さらに，近年, 最も代表的な識別手法として知られているサポートベクトルマシンと L_1 正則化付ロジスティック回帰との比較により, 提案手法の識別性能そのものを検証する．

3.1.1 ナイーブベイズ (NB)

NB モデルでは，各特徴要素を多項分布とし，それらの積 (各特徴が統計的に独立と仮定) として表現されたモデルで，生成モデルは以下のように書ける．

$$x_{i,j}^{(k)}|\boldsymbol{\theta}_{k,j} \sim \mathrm{Multinomial}(x_{i,j}^{(k)}; \boldsymbol{\theta}_{k,j}),$$
$$\boldsymbol{\theta}_{k,j}|\boldsymbol{\beta}_{k,j} \sim \mathrm{Dirichlet}(\boldsymbol{\theta}_{k,j}; \boldsymbol{\beta}_{k,j}).$$

ここで, 事前分布としてディリクレ分布を用いた．また, パラメータ $\boldsymbol{\theta}_{k,j}$ は最大事後確率 (MAP) 推定で求めた．但し, ディリクレ分布のハイパーパラメータは 5 分割交差検定法により決定した．

3.1.2 交差検定法に基づく特徴選択

交差検定法に基づく特徴の逆方向選択 (Backward Selection：BS) について述べる．BS とは, 全ての特徴から始め, 1 つずつ冗長な特徴を削減する方法である．ある特徴を削除したときの 5 分割交差検定法により求まる識別率が最大となる特徴を削除するものとする．

3.1.3 L_1 正則化付ロジスティック回帰

ロジスティック回帰を用いて, 多クラス識別を行う方法について述べる．このとき, 目的関数に L_1 正則化項を加えることによってスパースな解を得ることができる [8]．これは Lasso 法とも呼ばれる．ロジスティック回帰では, 以下のクラス事後確率が最大となるクラスに識別する．

$$p(C_k|\boldsymbol{x}) = \frac{\exp(a_k)}{\sum_{k'=1}^{K}\exp(a_{k'})}. \tag{12}$$

但し, $a_k = \ln(p(\boldsymbol{x}|C_k)p(C_k))$ であり, モデルが指数型分布族に属する場合, a_k は $a_k = \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x} + w_{k,0}$ のように $\boldsymbol{x}$ の線形結合で表すことができる．また, 目的関数は, T を目標変数とすると, L_1 正則化項を加えて以下のようになる．

$$\mathrm{W}^* = \arg\max_{\mathrm{W}} p(\mathrm{T}|\mathrm{W}) - \eta\sum_{k=1}^{K}\sum_{j=1}^{M}|w_{k,j}|. \tag{13}$$

ここで, $\eta > 0$ は対数尤度最大化と正則化項のトレードオフを決めるパラメータである．η は 5 分割交差検定法により決定した．

3.1.4 サポートベクトルマシン (SVM)

SVM は, 高い汎化性能をもつ 2 クラス識別器として知られている．SVM を多クラス識別問題に拡張する方法として, 一対多手法やペアワイズ手法がある [10]．本実験では一対多手法を用いる．カーネルは線形カーネルおよび RBF カーネルを用い, マージンパラメータとカーネルパラメータは, 5 分割交差検定法によって決定した．

3.2 実験データ

本節では, 実験に用いるデータについて説明する. 提案手法の生成モデルに従って生成した人工データと DNA データを用いて評価を行う. 実験は全てランダムサンプリングして作った 10 セットのデータを用い, 全データの 2/3 を訓練データ, 1/3 をテストデータとした.

3.2.1 人工データ

人工データのクラス数, 次元数, データ数, シンボルの最大値 L およびデータを生成するときに用いたハイパーパラメータを表 1 に示す. $\boldsymbol{\alpha}$ はデータ全体を表すディリクレ分布のハイパーパラメータで, 大きいほど多項分布は一様分布に近づく. $\boldsymbol{\beta}_{k,j}$ はクラス固有のディリクレ分布のハイパーパラメータで, 小さいほど偏った多項分布となる. また, a, b はベータ分布のハイパーパラメータで, b が大きくなるほどベルヌーイ分布の 0 の出る確率が大きくなる.

表 1 人工データ生成時のパラメータの値
Table 1 Value of the parameters to sample synthetic data

	クラス数	次元数	データ数	L	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}_{k,j}$	a, b
data1	5	50	50	5	10	0.2	1, 8
data2	5	50	50	10	10	0.3	1, 6
data3	10	50	100	5	10	0.2	1, 6
data4	10	100	50	10	10	0.2	1, 8
data5	30	50	50	10	10	0.2	1, 6

3.2.2 実データ

実データとして, Promoter データセットおよび Splice データセットの 2 つを用いて実験を行う. DNA は { A,G,C,T } の 4 つの塩基から構成され, これらのシーケンスとして表現される. このシーケンスを学習し, クラスが未知の DNA を識別するという問題を考える. DNA データを用いた研究は広く行われており, 文献 [11] では, ニューラルネットワークに生物学の知識を取り込んで学習を行う KBANN が提案されている.

Promoter データセットは, シーケンスの長さが 57 の 2 クラス問題である. プロモーター (Promoter) とは, DNA の塩基配列の情報が RNA に転写される際に, 転写の開始位置を特定するために使われる短いシーケンスのことである. データには, プロモーターが含まれるものとそうでないものがあり, これらを識別することが目標となる. 各クラスのデータ数は共に 53 である.

Splice データセットは, シーケンスの長さが 60 の 3 クラス問題である. 蛋白質を合成する際, DNA の塩基配列の情報が RNA に転写される. このとき, RNA にはイントロンとエクソンと呼ばれる配列が交互に含まれており, 蛋白質合成には余分なイントロンを除去する必要がある. このような処理をスプライシング (Splicing) という. Splice データセットでは, このイントロンとエクソンの境界を識別する問題を扱う. エクソン-イントロンの境界含むデータをクラス 1, イントロン-エクソンの境界を含むデータをクラス 2, どちらの境界も含まれていないデータをクラス 3 とする. 各クラスのデータ数は, クラス 1 が 767, クラス 2 が 768, クラス 3 が 1655 である.

4. 結果

4.1 人工データでの結果

本節では, 表 1 で示したハイパーパラメータで生成した人工データでの識別結果を述べる. 結果は全て 10 セットでの識別結果の平均識別率と標準偏差で評価する. 尚, 提案手法のハイパーパラメータは 5 分割交差検定法で決定した. ここで, BDFS は提案手法を表し, NB はナイーブベイズ, BS は交差検定法に基づき特徴を逆方向選択 (Backward Selection：BS) する手法を表す.

4.1.1 特徴選択の効果 (人工データ)

表 2 に各特徴選択手法を用いた場合のテストデータの識別率を示す. BDFS と NB を比較すると, 全ての人工データにおいて BDFS の汎化性能が優れており, 提案手法による特徴選択が有効であることが分かる. 特に, data2 や data5 では NB に比べ 15%以上の識別率の改善が見られる. data2 において NB は, 識別に有効でないような特徴に対して過学習を起こしていると考えられ, 実際, 訓練データの識別率は, BDFS は 96.4%, NB は 99.3%であった. data5 ではクラス数が 30 クラスと非常に多いため, クラス毎に有効な特徴選択を行う提案手法が有効に働いていると考えられる. また, 交差検定法に基づく特徴選択手法と比較して, 提案手法が汎化性能向上に非常に有効であることが分かる.

表 2 各特徴選択手法を用いた識別率 (人工データ)(%)
Table 2 Accuracy of using each feature selection method (synthetic data)

	BDFS	NB	BS
data1	96.6 ± 1.8	88.2 ± 4.1	92.1 ± 1.8
data2	96.0 ± 1.1	80.8 ± 4.1	80.7 ± 3.0
data3	92.3 ± 0.3	90.8 ± 1.0	91.0 ± 1.0
data4	93.6 ± 0.6	86.0 ± 1.2	85.2 ± 0.9
data5	87.7 ± 0.1	61.6 ± 1.6	62.2 ± 1.6

4.1.2 識別器としての評価 (人工データ)

表 3 に各識別器により人工データを用いて学習を行ったときの識別率を示す. L1 正則化付ロジスティック回帰を LR+Lasso と表し, 線形カーネルを用いた SVM を SVM(L), RBF カーネルを用いた SVM を SVM(R) と表す. ほとんど全てのデータにおいて, 提案手法の方が従来手法よりも汎化性能が高いことが分かる. LR+Lasso では, 間接的に特徴選択を行っているが, あくまでスパースな解表現を求めるための手法であり, 識別に有効な特徴を選択することは難しかったと考えられる.

表 3 各識別器の識別率 (人工データ)(%)
Table 3 Accuracy of each classifier (synthetic data)

	BDFS	LR+Lasso	SVM(L)	SVM(R)
data1	96.6 ± 1.8	90.5 ± 2.8	79.4 ± 1.9	87.4 ± 2.3
data2	96.0 ± 1.1	78.6 ± 4.6	71.0 ± 3.0	79.9 ± 3.6
data3	92.3 ± 0.3	93.1 ± 3.4	82.3 ± 2.2	87.0 ± 1.9
data4	93.6 ± 0.6	91.0 ± 2.9	84.5 ± 3.1	90.2 ± 2.3
data5	87.7 ± 0.1	74.4 ± 2.5	65.1 ± 1.9	69.1 ± 0.9

4.2 実データ

本節では, 3.2.2 節で述べた二つの DNA データを用いて実験を行った結果を以下に述べる.

4.2.1 特徴選択の効果 (DNA データ)

表 4 に DNA データに対して各特徴選択手法を用いた場合のテストデータの識別率を示す. Promoter データセットにおいて, BDFS が NB および BS に比べ高い汎化性能を示している. BDFS の訓練データの識別率は 98.4%であるのに対し, NB は 99.6%と非常に高く, 提案手法の特徴選択により過学習を緩和できていた. また, BS では多少の識別率の改善が見られたが, BDFS ほどの識別率は実現できていない.

Splice データセットでは, 提案手法は他の手法と比べ, 同等以上の識別性能を示した. 提案手法による大幅な性能改善が見られなかった理由として, 問題が簡単であり, 従来手法において過学習による性能低下が起こらなかったためであると考えられる. 実際, Splice データセットでの訓練データの識別率は, BDFS が 95.8%, NB が 96.1%であった.

表 4 各特徴選択手法を用いたテストデータの識別率 (DNA データ) (%)
Table 4 Test accurary of using each feature selection method (DNA data)

	BDFS	NB	BS
Promoter	91.6 ± 5.2	88.8 ± 3.5	89.4 ± 4.5
Splice	95.7 ± 1.4	94.7 ± 0.7	95.1 ± 0.7

4.2.2 識別器としての評価

DNA データにおいて, 各識別器を学習したときの識別率を表 5 に示す. 提案手法は, LR+Lasso や SVM と比べ同等以上の識別性能を示している. これにより, BDFS が実データにおいても十分に有用であることが分かる.

表 5 各識別器を用いた識別率 (%)
Table 5 Accuracy of using each classifier (DNA data)

	BDFS	LR+Lasso	SVM(L)	SVM(R)
Promoter	91.1 ± 5.2	90.8 ± 5.1	79.6 ± 4.3	86.4 ± 5.2
Splice	95.6 ± 1.4	95.8 ± 0.6	88.3 ± 3.5	94.0 ± 5.2

5. まとめ

本論文では, ベイズモデルに基づきクラス毎に固有の特徴選択およびクラス識別を行う手法を提案した.

評価実験では, 人工データおよび実データを用いて, 提案手法による特徴選択が汎化性能の観点で有効であることを示した. 人工データにおいてクラス数が多い場合には, 従来の全クラス共通に特徴選択を行う手法よりも, 識別性能がより顕著に優れていることを示した. また, L_1 正則化付ロジスティック回帰およびサポートベクトルマシンとの比較により, 識別器としても非常に優れていることを示した.

本論文では, カテゴリカルデータのみを対象にモデル化を行ったが, 連続値データを対象としたモデル化を考えれば, 同様の概念に基づく特徴選択が行えるように拡張することが可能であると考えられる.

参考文献

[1] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
[2] M. A. Hall, “Correlation-based Feature Selection for Machine Learning,” *PhD Thesis*, Department of Computer Science, Waikato University, New Zealand, 1999.
[3] H. Peng, F. Long and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.
[4] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machine,” vol. 46, no. 1-3, pp. 389-422, 2002.
[5] H. Shan and A. Banerjee, “Bayesian Co-clustering,” *Proc. IEEE International Conference on Data Mining (ICDM)*, pp. 530-539, 2008.
[6] P. D. Hoff,“Subset Clustering of Binary Sequences, with an Application to Genomic Abnormality Data,” *Biometrics*, vol. 61, no. 4, pp. 1027-1036, 2005.
[7] K. Ishiguro, N. Ueda and H. Sawada, “Subset Infinite Relational Models,” *Proc. 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, to appear
[8] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of Royal Statistical Society*, vol. 58, no. 1, pp. 267-288, 1996.
[9] C. W. Hsu, C. C. Chung and C. J. Lin, “A Practical Guide to Support Vector Classification,” Technical Report, Department of Computer Science, National Taiwan University, Taipei, 2003.
[10] V. N. Vapnik, *Statistic Learning Theory*, Wiley, New York, 1998.
[11] G. G. Towell and J. W. Shavlik, “Knowledge-Based Artificial Neural Networks,” *Artificial Intelligence*, vol. 70, no. 1-2, pp. 119-165, 1994.