情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

条件付きエントロピー最小化に基づく教師付き次元削減手法
Supervised Dimensionality Reduction by Conditional Entropy Minimization

日野英逸* 村田昇†
Hideitsu Hino Noboru Murata

Abstract: データが有する本質的な情報を失うことなくデータの次元を削減することは情報処理における重要な課題である. 学習データにクラスラベルが付随している教師付き次元削減手法としては, Fisher Discriminant Analysis(FDA) が広く用いられている. しかし, FDA による次元削減で得られる判別曲面は非常に限定された理想的状況においてのみ最適であり, 実際上も適切な判別曲面が得られないことが多い. 本報告では, 情報論的な観点から教師付き次元削減問題を捉え, 条件付きエントロピー最小化に基づく次元削減の枠組みを提案する. 推定した条件付きエントロピーの最小化を勾配法により実行し, 可視化及び判別問題への適用結果を記す. さらに, カーネル法を用いた非線型教師付き次元削減への拡張も試みる.

Keywords: dimensionality reduction, conditional entropy, classification, visualization

1 まえがき

情報処理の重要な目的の一つとしてデータの縮約, あるいはコンパクトな表現がある. 特に, 情報処理の対象として画像, 音声, テキスト, 遺伝子発現データなど, 極めて高次元なデータを扱う場合, 本質的な情報を保持したまま次元を低減することは必須である. 次元削減により, 後に続く判別やクラスタリングといった処理が高速に実現でき, また, データを低次元で表現することで可視化も可能となる.

教師無しの線型次元削減手法としては, 主成分分析 [1](Principal Component Analysis:PCA) が代表的である. 一方, 代表的な教師付き線型次元削減手法として, Fisher の判別分析 [2](Fisher Discriminant Analysis:FDA) がある. FDA は, 特徴データとそのクラスラベルが観測された状況で, クラス内のデータの分散を小さく保ちつつ, クラス間の分散が大きくなるような方向への特徴データの射影を求める手法である.

各クラスのデータが共分散構造の等しい正規分布に従っている時は, FDA は最適なクラス分離を与える方向を発見することができる. しかし, 多くの問題では正規性の仮定は成り立たず, 判別性の高い射影を得ることができないことがある. FDA の自然な拡張として, 同一のクラスに属するデータ同士の類似度を考慮した, Local Fisher Discriminant Analysis(LFDA) が提案されている [3]. これはデータの局所性をデータ間の類似度行列という形で導入するものであり, 判別の前処理として用いた場合に FDA を大きく上回る判別精度が得られると報告されている.

本報告では, 情報論的観点から次元削減問題にアプローチする. 上述の代表的な次元削減手法をエントロピーの観点から理解し, 続いて次元削減のための目的関数として条件付きエントロピーが自然であることを論じる. 提案する手法は, データの分布を仮定せず, ノンパラメトリックにデータの密度推定を行った上でクラス条件付きエントロピーを最小化するような変換を探索するものであり, FDA における最適性の仮定から大きく外れたデータに対してもよい結果が期待できる.

以下, 本報告の構成を記す. 第 2 節において, 線型次元削減問題を定式化し, その情報論的な解釈を与える. 第 3 節では, 条件付きエントロピーが情報論的に自然な次元削減の基準となることを示す. エントロピーの推定と最適化に用いるカーネル密度推定にも言及する. 条件付きエントロピー最小化による次元削減を提案し, 人工データを用いた実験により, FDA がうまく働かない状況においても提案手法が最適な射影方向を発見できる例を示す. 第 4 節では, 条件付きエントロピーを変形すること

*早稲田大学, 169-8555, 東京都新宿区大久保 3-4-1 tel. 03-5286-3383, e-mail hideitsu.hino@toki.waseda.jp,
Waseda University, 3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan
†e-mail noboru.murata@eb.waseda.ac.jp

により, 提案手法が教師付き次元削減 (より一般には特徴抽出) 手法として機能する原理を考察する. 第 5 節ではベンチマークデータセットを用いて, 多クラスデータに対する可視化の例と, 提案手法によって得られた低次元特徴空間における判別実験の結果を示す. 第 6 節では, 提案手法の非線型次元削減への拡張を検討する.

2 次元削減手法の情報論的理解

データ $\{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^n$ が与えられたときに, これを m 次元ベクトル $f(\boldsymbol{x}_i) = \boldsymbol{z}_i \in \mathbb{R}^m$, $(m < n)$ に変換する次元削減問題を考える. 線型の次元削減の場合には, 変換 f は行列 A によって

$$\boldsymbol{z}_i = A^T \boldsymbol{x}_i, \quad A \in \mathbb{R}^{n\times m}. \tag{1}$$

で記述できる. ここでは, 次元削減手法をエントロピーの観点から考察する. 本稿で用いるエントロピーとは, (Shannon) 微分エントロピーを意味し, 確率変数 X に対して,

$$H(X) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \tag{2}$$

で定義される. ただし p は X の密度関数である.

教師無し次元削減手法として広く用いられている PCA は, データを低次元に射影したときに分散ができるだけ大きくなるような射影軸を求める手法である. これは, データが正規分布に従うという条件下で, もとのデータの情報量 $-\log p(\boldsymbol{x})$ を平均としてできるだけ保つような低次元構造を抽出することに相当する. 実際, 1 次元正規分布 $\mathcal{N}(x; \mu, \sigma^2)$ の平均情報量, すなわちエントロピーは

$$-\int_{-\infty}^{\infty} p(x) \log p(x) dx = \frac{1}{2} \log(2\pi e \sigma^2) \tag{3}$$

となり, 分散 σ^2 が大きいことはエントロピーが大きいことと等価である. PCA ではラベル情報が利用できないので, 判別のための射影軸の決定ではなく, 情報量を保存することでデータの特徴を最もよく捉えると考えられる射影軸を定めている.

一方, 教師データが与えられている状況では, データ $\boldsymbol{x}$ が属するクラスラベルを用いた判別的な次元縮約として, FDA がよく利用される. FDA は, データ $\{\boldsymbol{x}_i\}_{i=1}^N$ とそのクラスラベル $\{y_i\}_{i=1}^N$, $y_i \in \{1, 2, \ldots, C\}$ が与えられたときに, 最もクラス判別が容易になる小数個の射影軸を求める手法である. データ $\boldsymbol{x} \in \mathbb{R}^n$ を式 (1) のように変換するとして, 変換後のデータのクラス間分散 $A^T \Sigma_b A$ とクラス内分散 $A^T \Sigma_w A$ の比が最大になるような変換を求める. ただし,

$$\begin{aligned}
\Sigma_w &= \frac{1}{N} \sum_{y=1}^{C} \sum_{\boldsymbol{x}\in D_y} (\boldsymbol{x} - \boldsymbol{\mu}_y)(\boldsymbol{x} - \boldsymbol{\mu}_y)^T = \sum_{y=1}^{C} \frac{N_y}{N} \Sigma_y, \\
\Sigma_b &= \frac{1}{N} \sum_{y=1}^{C} N_y (\boldsymbol{\mu}_y - \boldsymbol{\mu})(\boldsymbol{\mu}_y - \boldsymbol{\mu})^T,
\end{aligned}$$

であり, D_y はクラス y に属するデータの集合, $\boldsymbol{\mu}_y$ と Σ_y は D_y に属するデータの平均と共分散行列, $\boldsymbol{\mu}$ は全てのデータの平均を表す. 以上の準備の下, FDA は, $\log |A^T \Sigma_w A| / |A^T \Sigma_b A|$ を最小化する変換 A を求めることと理解される. ただし, $|\cdot|$ で行列の行列式を表すものとする. ここで, log の中身の分母, 分子を定数倍してもこの目的関数の値は変わらないことから, 通常は $|A^T \Sigma_b A|$ を一定に固定した上での最小化問題

$$\min_A \log |A^T \Sigma_w A| \quad \text{subject to} \quad |A^T \Sigma_b A| = const. \tag{4}$$

を解く.

データが各クラスで同一の共分散行列 $\Sigma_y = \Sigma_c$, $y = 1, \ldots, C$ を持つ正規分布に従う場合には, FDA によって Bayes 最適な判別曲面が得られることが知られている. この理想的な状況では $\Sigma_w = \Sigma_c$ であり, 従って

$$\begin{aligned}
H(A^T X|Y) &= \log(2\pi)^{m/2} e + \sum_{y=1}^{C} \frac{N_y}{2N} \log |A^T \Sigma_y A| \\
&= \log(2\pi)^{m/2} e + \frac{1}{2} \log |A^T \Sigma_c A| \\
&= \log(2\pi)^{m/2} e + \frac{1}{2} \log |A^T \Sigma_w A|
\end{aligned}$$

から FDA における $\log |A^T \Sigma_w A|$ の最小化はクラス条件付きエントロピーの最小化と等価である.

以上より, 教師無しの代表的な次元削減手法である PCA は, データ分布の正規性の仮定の下でエントロピーを最大にするように変換をすることで特徴を抽出しようとするものであり, 教師付きの代表的な次元削減手法である FDA は, データが各クラスで同一の共分散行列を持つ正規分布に従うという仮定の下で, 条件付きエントロピーを最小にすることで特徴を抽出するものであることがわかる. 本節における考察を踏まえて, 次節では条件付きエントロピー最小化を基準とした教師付き次元削減方法を提案する.

3 提案する枠組み

データのクラスラベルが与えられている教師付き学習の枠組みにおいて, データの低次元空間での表現としては, 類似したデータ同士がコンパクトにまとまっている

ことが望ましい．データを $\boldsymbol{x}$ から $\boldsymbol{z}$ に変換してコンパクトな表現を得ることを考えると, 相互情報量

$$I(X;Z) = H(Z) - H(Z|X) \tag{5}$$

が小さいほど圧縮性の高い良い変換であるといえる [4]. 学習の際, クラスラベル y などの情報が利用できる教師付き学習の枠組みでは, 教師データ y で条件付けた場合にデータがコンパクトに表現できればよく, このときは, 変換の良さの尺度として条件付き相互情報量

$$I(X;Z|Y) = H(Z|Y) - H(Z|X,Y) \tag{6}$$

を考えるのが自然である．さらに, データの変換 $f : \boldsymbol{x} \mapsto \boldsymbol{z}$ が確定的な場合を考えているので $H(Z|X,Y)$ は 0 になり, $H(Z|Y)$ が小さいことが, 教師付きの枠組みにおけるコンパクトなデータ表現の基準となる．条件付きエントロピー $H(Z|Y)$ を変換 $f : \boldsymbol{x} \mapsto \boldsymbol{z}$ に関して最小化することで次元削減のための変換を学習するというのが提案する枠組みである．ただし, $H(Z|Y)$ は全てのデータ $\boldsymbol{x}$ を一点に写像する f によって最小化される．また, 変換関数 f の自由度が高い場合には与えられた学習データに過度に適合する可能性もある．こうした問題を避けるために, 実際上は何らかの制約条件が必要となる．ここでは一般に, 制約の強さを表すパラメタ ε と, 変換関数 f 及び学習データ D に依存して定まる制約関数 $\Psi(f, D)$ を用いて, 制約項を $\varepsilon\Psi(f, D)$ で表現する．従って,

$$\min_{f:\boldsymbol{x}\mapsto\boldsymbol{z}} H(Z|Y) + \varepsilon\Psi(f, D) \tag{7}$$

なる最適化問題を考えることになる．制約関数 $\Psi(f, D)$ の具体的な形状は対象とする問題に応じて適切に定める必要がある．

まずは, 線型変換 $A \in \mathbb{R}^{n\times m}$ について式 (7) の最小化問題を考える．エントロピーの最小化のために, Gaussian カーネル密度推定に基づくエントロピーの推定を行う．エントロピー推定の方法は数多くあるが, ここでは, Leave-One-Out(LOO) に基づく方法を採用した [5]．まず, データ $D = \{\boldsymbol{x}_i\}_{i=1}^N$ が与えられたとき, $\boldsymbol{x}$ の密度を

$$\hat{p}(\boldsymbol{x}; D, h) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi h^2}} \exp\left(-||\boldsymbol{x} - \boldsymbol{x}_i||^2/2h^2\right) \tag{8}$$

で推定する[1]．その上で, 一次元のエントロピーの近似を以下のように行う．推定した密度を用いてエントロピーを

$$H(X) \approx \tilde{H}(X) = -E[\log \hat{p}(X; D, h)] \tag{9}$$

のように近似し, さらに $\hat{p}(\boldsymbol{x}; D, h)$ を $\hat{p}(\boldsymbol{x}_j; D\backslash\{\boldsymbol{x}_j\}, h)$ で置き換えて, LOO 法によって期待値の計算を

$$\tilde{H}(X) \approx \hat{H}(X) = -\frac{1}{N}\sum_{j=1}^{N} \log \hat{p}(\boldsymbol{x}_j; D\backslash\{\boldsymbol{x}_j\}, h) \tag{10}$$

のように近似する．以下ではエントロピーを上式で推定することとして, 推定値であることを表す $\hat{\cdot}$ は省略する．

多次元確率変数の密度推定及び結合エントロピーの推定も, 上述のようにカーネル密度推定を介して行うことができる．しかし, 十分な精度で多次元の密度推定を行うことは困難であるため, ここでは, 多次元のデータ $\boldsymbol{z} \in \mathbb{R}^m$ の結合エントロピーの推定をするのではなく, 変換された特徴量ベクトルの各次元での周辺エントロピーの和で与えられる上界を求め, それを最小化する [4]．つまり, 変換行列 A の第 l 列のベクトル $\boldsymbol{a}_l$ によって射影される第 l 次元特徴量に関する周辺エントロピー

$$H_l(\boldsymbol{a}_l^T X) = H_l(Z_l) = -\int p(z_l)\log p(z_l) dz_l$$

の和が, 結合エントロピーの上界

$$\overline{H(Z)} = \sum_{l=1}^{m} H_l(Z_l) \geq H(Z) \tag{11}$$

を与えるので, 各次元の周辺エントロピー $H_l(\boldsymbol{a}_l^T X)$ を $\boldsymbol{a}_l$ に関して最小化することで, 目的関数である結合エントロピーを近似的に最小化する．また, このエントロピーをクラス別に計算したものを $\overline{H(Z|Y=y)} = \sum_{l=1}^{m} H_l(Z_l|Y=y)$ として, これをクラス事前確率 (各クラスのデータ数と全データ数の比で推定) で重みを付けて加えたものが $H(Z|Y)$ の上界

$$\begin{aligned}\overline{H(Z|Y)} &= \sum_{y=1}^{C} p(y)\overline{H(Z|Y=y)} \\ &= \sum_{y=1}^{C} p(y)\sum_{l=1}^{m} H_l(Z_l|Y=y)\end{aligned} \tag{12}$$

である．以下では, 多次元のエントロピーを実際に最小化する際には各次元に関する周辺エントロピーの和を最小化するものとする．

3.1 勾配法による最適化

条件付きエントロピーの近似を, A に関する勾配法で最小化する．周辺エントロピーの和を最適化するので, 変換行列 A の第 l 列を $\boldsymbol{a}_l$ として, 変換特徴量 $\boldsymbol{a}_l^T\boldsymbol{x} =$

[1] 写像 f に関する最適化を繰り返すことで, 変換後のデータの密度推定に適したバンド幅は変わっていくため, バンド幅の推定方法は高速に行えることが必要である．カーネルのバンド幅 h は, “Silverman's Rule of Thumb” と呼ばれるヒューリスティックな方法を用いて推定した [6].

$z_l, l = 1, 2 \ldots, m$ についての周辺エントロピーの勾配を求めてから, A の各列毎に更新を行った.

解くべき最適化問題は

$$\min_A \overline{H(A^T X|Y)} + \varepsilon \Psi(A, D) \qquad (13)$$

である. ただし, 全データ集合を $D = \{\boldsymbol{x}_i\}_{i=1}^N$, 各クラス y に属するデータ集合を $D_y = \{\boldsymbol{x}_j\}_{j=1}^{N_y}$, $y = 1, \ldots, C$ として, 条件付きエントロピーは

$$\begin{aligned} H(A^T X|Y) &= \sum_{y=1}^{C} \frac{N_y}{N} H(A^T X|Y = y) \\ &= -\sum_{y=1}^{C} \frac{1}{N} \sum_{\boldsymbol{x}_j \in D_y} \log \hat{p}(A^T \boldsymbol{x}_j; D_y \backslash \{\boldsymbol{x}_j\}, h) \end{aligned}$$

で計算される. $H(A^T X)$ の A の第 l 列 $\boldsymbol{a}_l$ に関する条件付き周辺エントロピーの導関数は

$$\frac{\partial H(\boldsymbol{a}_l^T X|Y)}{\partial \boldsymbol{a}_l} =$$

$$\frac{1}{h^2 N} \sum_{y=1}^{C} \sum_{\boldsymbol{x}_j \in D_y} \frac{\sum_{\boldsymbol{x}_i \in D_y \backslash \{\boldsymbol{x}_j\}} e^{\left(-\frac{||\boldsymbol{a}_l^T \boldsymbol{x}_j - \boldsymbol{a}_l^T \boldsymbol{x}_i||^2}{2h^2}\right)} \boldsymbol{v}_{ji}}{\sum_{\boldsymbol{x}_i \in D_y \backslash \{\boldsymbol{x}_j\}} e^{\left(-\frac{||\boldsymbol{a}_l^T \boldsymbol{x}_j - \boldsymbol{a}_l^T \boldsymbol{x}_i||^2}{2h^2}\right)}},$$

$$\boldsymbol{v}_{ji} = (\boldsymbol{x}_j - \boldsymbol{x}_i)^T \boldsymbol{a}_l \cdot (\boldsymbol{x}_j - \boldsymbol{x}_i) \in \mathbb{R}^n$$

であり, 制約条件なしの勾配法と, 近似的な制約充足のために後述の準直交化を繰り返すことで式 (13) の最適化が行える.

3.1.1 準直交化

周辺エントロピーで結合エントロピーの上界を最小化する場合, 各周辺エントロピーを個別に最小化しても, 単一の方向ベクトルしか得られない. そこで準直交化 [7] を各ステップで行い, 勾配法の各ステップにおいて, 射影した各次元でのデータ $\boldsymbol{a}_1^T \boldsymbol{x}, \ldots, \boldsymbol{a}_m^T \boldsymbol{x}$ を無相関化する. つまり, データ $\{\boldsymbol{x}_i\}_{i=1}^N$ は白色化してあるとして, 周辺エントロピーの最小化の途中で得られている行列 $A \in \mathbb{R}^{n \times m}$ を, $||A^T A - I_m||_F$ を近似的に満たすように修正する. ここで I_m は $m \times m$ の単位行列であり, $|| \cdot ||_F$ は行列のフロベニウスノルムである. すなわち, 式 (13) における制約関数を $\Psi(A, D) = \Psi(A) = ||A^T A - I_m||_F$ と定めたことになる. この準直交化は, 次の 2 つの操作を適当な回数繰り返すことで行われる:

1. $A \leftarrow \frac{3}{2} A - \frac{1}{2} A A^T A$.
2. A の各列のノルムを 1 に正規化.

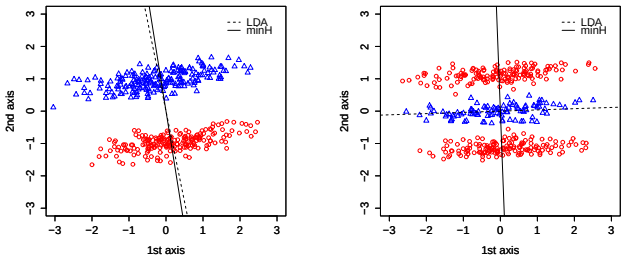
実際, 対称行列 $A^T A$ の対角化を, 直交行列 $E \in \mathbb{R}^{m \times m}$ と $A^T A$ の固有値 $\{d_i\}_{i=1}^m$ を並べた対角行列 D を用いて $A^T A = EDE^T$ で表すと, 上記の step.1 によって

$$\begin{aligned} A^T A &\mapsto \frac{1}{4}(3A - AA^T A)^T (3A - AA^T A) \\ &= E\left(\frac{9}{4} D - \frac{6}{4} D^2 + \frac{1}{4} D^3\right) E^T \end{aligned}$$

となる. A の各列は正規化してあるので $d_i \in [0, 1]$ であり, $A^T A$ の変換後の固有値を $\{h(d_i)\}_{i=1}^m$ とすると $h(d_i) = \frac{1}{4}(9d_i - 6d_i^2 + d_i^3)$ となる. 従って, $h(d_i) - d_i = \frac{d_i}{4}\{(d_i - 3)^2 - 4\} \geq 0$ より $h(d_i) \geq d_i$ が成立するので, $A^T A$ の固有値は 1 に収束する.

3.2 人工データを用いた実験結果

図 3.2 に, 2 種類の 2 クラスデータを用いて, FDA 及び提案手法を用いて求めた判別軸を示す. 2 クラス判別問題なので, FDA により射影軸が 1 つ得られる. 提案手法では, 変換 $A = \boldsymbol{a}$ を 2 次元ベクトルとして学習することで, FDA と同様に射影軸が 1 つ得られる. 点線が FDA による判別軸で, 実線が $H(\boldsymbol{a}^T X|Y)$ の最小化で得られた $\boldsymbol{a}$ で射影される軸である. 各クラスが単峰性の分



(a) Unimodal data. (b) Multimodal data.

図 1: Discriminative axes found by FDA and proposed method.

布に従うような単純な場合 (a) には, FDA と提案手法はほぼ同一の射影軸を与える. 一方, 同クラスで modality が二つある場合 (b) では, 提案手法が優れた結果を与えていることがわかる.

4 提案手法の正当性の検討

$H(Z|Y)$ を書き換える事で, なぜ条件付きエントロピーの最小化アプローチがうまくいくのかを再考する. まず, 確率変数 X のネゲントロピー [8] に注目する. ネゲントロピー及び条件付きネゲントロピーは, それぞれ $J(Z) = H_G(Z) - H(Z), J(Z|Y) = H_G(Z|Y) - H(Z|Y)$ で定義される. ただし, $H_G(Z)$ は, $p(\boldsymbol{z})$ と同じ共分散構造を持つ正規分布のエントロピーである. このネゲントロピーを用いると, $\boldsymbol{z} = A^T \boldsymbol{x}$ とクラスラベル y の相互

情報量が次のように変形できる:

$$\begin{aligned} I(Z;Y) &= H(Z) - H(Z|Y) \\ &= \{H_G(A^T X) - H_G(A^T X|Y)\} \\ &\quad -\{J(A^T X) - J(A^T X|Y)\} \\ &= \frac{1}{2}\log \frac{|A^T \Sigma A|}{\prod_{y=1}^{C} |A^T \Sigma_y A|^{p(y)}} \\ &\quad -\{J(A^T X) - J(A^T X|Y)\}. \end{aligned}$$

ただし, Σ, Σ_y はそれぞれ全データの共分散行列, クラス y に属するデータの共分散行列であり, $p(y)$ はクラス事前確率である. 従って, 最小化の目的である $H(Z|Y)$ は,

$$\begin{aligned} H(Z|Y) &= H(A^T X) + \{J(A^T X) - J(A^T X|Y)\} \\ &\quad -\frac{1}{2}\log \frac{|A^T \Sigma A|}{\prod_{y=1}^{C} |A^T \Sigma_y A|^{p(y)}} \\ &= H_G(A^T X) - J(A^T X|Y) \\ &\quad -\frac{1}{2}\log \frac{|A^T \Sigma A|}{\prod_{y=1}^{C} |A^T \Sigma_y A|^{p(y)}} \end{aligned} \tag{14}$$

の 3 項に分ける事ができる.

4.1 全データが正規分布に従うとした時の全データのエントロピー

変形した条件付きエントロピーである式 (14) の第一項 $H_G(A^T X)$ は, データ全体が正規分布に従うとした場合のエントロピーである. これは全データの共分散行列の行列式で完全に決まる項であり, この項を小さくすることは, データ全体をコンパクトに表現することにつながる. ただし, データのスケール変換によりいくらでも小さくなる項であり, 判別性には関係しない. 従って, 変換の正規化 (直交化など) を行うという状況では, 本質的にはこの項は影響しない.

4.2 条件付きネゲントロピー

第二項に現れる (条件付き) ネゲントロピーは, 主に独立性分分析の分野において正規分布からの離れ具合の指標として用いられている [8]. 式 (14) の最適化により $J(A^T X|Y)$ が大きくなることで, クラスラベルで条件付けたときにネゲントロピーが大きくなるようにデータが分布することになる. つまり条件付けすることでデータの構造がより顕著に現れる (特徴的な構造が現れる) ようにデータを変換していることに相当する.

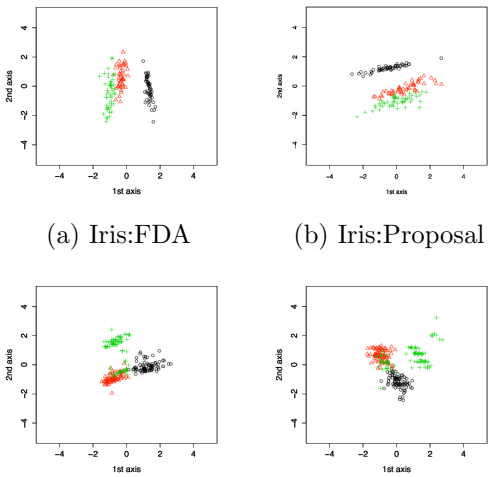
4.3 クラス内分散が一定でない判別分析規準

式 (14) の第三項は, 各クラスの分散構造及び事前分布が異なる状況での線型判別分析 (Heteroscedastic discriminant analysis;[9]) の目的関数そのものであり, この項の最適化はクラス判別に大きく寄与することが理解できる.

5 実験

5.1 多クラスデータの可視化

まず, 提案手法の可視化への適用例を示す. ここで, 条件付きエントロピーを直接最小化するのではなく, 式 (14) の最小化によっても理論的には同じ効果があるはずである. その実証のために, ここでは式 (14) の各項の勾配を計算し, 勾配法による最適化によって実データの次元削減を行う. 当然, 直接 $H(Z|Y)$ を最小化するのと比べて計算量は大きくなるので, 式 (14) の最適化は本節の実験のみにとどめる. また, FDA による可視化も行うが, C をクラス数としたとき, FDA では $C-1$ 次元までしか判別軸を得ることができないため, 3 クラス以上のデータセットを用いる. データは UCI レポジトリから, Iris と Soybeans データを利用した. それぞれ, 4 次元 3 クラス, 22 次元 3 クラスのデータである.



(a) Iris:FDA (b) Iris:Proposal

(c) Soybean:FDA (d) Soybean:Proposal

図 2: Visualization result of multidimensional data.

どちらのデータに対しても, クラスがよく分離されるような 2 次元への射影が得られていることがわかる.

5.2 One-nearest-neighbor 法による判別

データの分離性を測る指標はいくつもあるが, ここでは単純に, one-nearest-neighbor 法 [10] で判別した場合の誤判別率を用いる. 実験には, IDA データセット[2]を用いた. 表 1 に, 利用したデータセットの名称, 特徴量

[2]以前は http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm から入手可能であったが, 現在は公開されていない

表 2: Misclassification rate of linear methods.

Data name	min $H(Z\|Y)$	LFDA	FDA	PCA	Euclidean
banana	**13.64**(0.765)[2]	13.7(0.8)	38.34(3.966)	13.99(0.849)[2]	13.64(0.761)
breast-cancer	**33.90**(4.704)[5]	34.7(4.3)	34.91(5.076)	40.71(7.085)[3]	32.73(4.824)
diabetes	31.98(4.703)[7]	32.0(2.5)	**31.32**(2.813)	38.44(5.019)[4]	30.12(2.051)
flare-solar	36.50(1.936)[4]	39.2(5.0)	**36.42**(1.875)	48.64(6.920)[5]	36.47(1.880)
german	34.91(3.024)[7]	**29.9**(2.8)	32.03(2.577)	41.83(4.452)[2]	29.46(2.469)
heart	27.68(3.909)[7]	**21.9**(3.7)	22.93(4.105)	46.27(23.894)[4]	23.16(3.735)
image	5.72(1.712)[7]	**3.2**(0.8)	22.12(0.860)	37.33(9.546)[2]	3.381(0.540)
ringnorm	**20.25**(1.303)[7]	21.1(1.3)	31.72(1.016)	28.04(5.075)[10]	35.03(1.362)
splice	31.36(6.958)[2]	**16.9**(0.9)	20.35(0.783)	43.90(4.894)[2]	28.77(1.524)
thyroid	4.674(2.535)[5]	**4.6**(2.6)	17.92(4.888)	9.05(4.366)[2]	4.36(2.210)
titanic	**22.510**(1.107)[1]	33.1(11.9)	22.53(1.066)	26.41(8.392)[1]	22.50(1.057)
twonorm	**3.359**(0.4241)[13]	3.5(0.4)	3.54(0.496)	7.55(18.770)[3]	6.68(0.718)
waveform	24.76(3.167)[7]	**12.5**(1.0)	18.61(1.162)	31.69(18.714)[9]	15.83(0.654)

次元, 学習データとテストデータの数と, 学習/テストデータの組のセット数を示す. PCA 及び提案手法によ

表 1: IDA data specifications.

Data name	dim	train(test) data size	set 数
banana	2	400(4900)	100
breast-cancer	9	200(77)	100
diabetes	8	468(300)	100
flare-solar	9	666(400)	100
german	20	700(300)	100
heart	13	170(100)	100
image	18	1300(1010)	20
ringnorm	20	400(7000)	100
splice	60	1000(2175)	20
thyroid	5	140(75)	100
titanic	3	150(2051)	100
twonorm	20	400(7000)	100
waveform	21	1000(1000)	100

る変換後のデータの次元は, 各データセットから 5 セット分の学習データを取り出し, それらを用いた 5 fold cross validation によって定めた. 表 2 に, one-nearest-neighbor 法による誤判別の平均及び標準偏差 (全て%表記) を示す. 提案手法及び PCA において選択された最適な次元 D は表中で $[D]$ のように記した. また, LFDA による同一のデータに対する判別結果は [3] より引用した. 提案手法による判別精度は, 多くのデータに対して PCA, FDA を上回っている. LFDA と比較すると, titanic データのように有意に優れた結果が得られる場合がある一方で, waveform, splice データのように精度が大幅に落ちる場合もあるが, その他の多くのデータに対して同程度の判別力を有していることがわかる. また, 表の最右行 (Euclidean) に示した次元削減をしていない特徴量を用いた one-nearest-neighbor 法による判別結果と比較しても, 多くの場合ほとんど判別精度が落ちず, 精度が向上しているケースも見られる.

6 非線型変換への拡張

非線型変換によって特徴抽出を行い, 次元を削減する方法を検討する. FDA のカーネル法を用いた非線型化 (Kernel Fisher Discriminant Analysis:KFDA) が提案されており [11], 線型分離不可能な複雑なデータに対して有効であることが知られている. 本節では, 提案手法の非線型化を, カーネル法の枠組みで試みる.

6.1 Kernel Fisher Discriminant Analysis

1 次元線型射影関数を, $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ で定義する. 多次元への射影は, 射影ベクトル $\boldsymbol{a}$ を列に並べた行列を用いることで実現するものとして, 以下では簡単のために 1 次元への射影のみを扱う. データ $\boldsymbol{x}$ を何らかの写像 Φ によって特徴量空間に写像した空間においてデータの判別を行うことを考えると, これは $f(\boldsymbol{x}) = \boldsymbol{a}^T\Phi(\boldsymbol{x})$ となる. ここで, 一般にカーネル法では適当な正則条件の下で $\boldsymbol{a} = \sum_{i=1}^N \alpha_i\Phi(\boldsymbol{x}_i)$ と書けることと, 特徴量空間における内積を表すカーネル関数 $\langle\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j)\rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ を用いると,

$$f(\boldsymbol{x}) = \sum_{i=1}^N \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) \tag{15}$$

と書き直せる. FDA のカーネル法による非線型化である KFDA は, クラス間分散行列 Σ_b 及びクラス内分散行列 Σ_w をカーネル関数を用いて書き直すことで容易に導出できる. 与えられたカーネル関数に関するデータのグラム行列を $K = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{ij} \in \mathbb{R}^{N\times N}$ とする. また, グラム行列から計算されるクラス毎の平均と全データの平

均に対応するベクトルを

$$\bar{\boldsymbol{k}}^y = \frac{1}{N_y}\sum_{\boldsymbol{x}_i\in D_y}\Big(k(\boldsymbol{x}_i,\boldsymbol{x}_1),\quad\cdots,\quad k(\boldsymbol{x}_i,\boldsymbol{x}_N)\Big)^T\in\mathbb{R}^N,$$

$$\bar{\boldsymbol{k}} = \frac{1}{N}\sum_{\boldsymbol{x}_i\in D}\Big(k(\boldsymbol{x}_i,\boldsymbol{x}_1),\quad\cdots,\quad k(\boldsymbol{x}_i,\boldsymbol{x}_N)\Big)^T\in\mathbb{R}^N$$

とすると, クラス間分散行列 V_b 及びクラス内分散行列を V_w はそれぞれ

$$V_b = \frac{1}{N}\sum_{y=1}^{C}N_y(\bar{\boldsymbol{k}}^y-\bar{\boldsymbol{k}})(\bar{\boldsymbol{k}}^y-\bar{\boldsymbol{k}})^T,$$

$$V_w = \frac{1}{N}\sum_{y=1}^{C}\sum_{i\in D_y}(\boldsymbol{k}_i-\bar{\boldsymbol{k}}^y)(\boldsymbol{k}_i-\bar{\boldsymbol{k}}^y)^T$$

となる. ただし $\boldsymbol{k}_i$ はグラム行列 K の第 i 列からなるベクトルを表す. 以上より, KFDA は $\log|\boldsymbol{\alpha}^TV_w\boldsymbol{\alpha}|/|\boldsymbol{\alpha}^TV_b\boldsymbol{\alpha}|$ の最小化問題になり, FDA と同様に $\boldsymbol{\alpha}^TV_b\boldsymbol{\alpha}$ を固定した上での最小化問題として定式化される. ただし, カーネル関数を用いた場合, その表現能力の高さから, 過学習を起こす可能性が非常に高い. そこで, 何らかの方法で正則化することが重要である. ここではクラス内分散行列を, 正則化項を加えた $V_w+\zeta K$ で置き換える. ここで ζ は適当な正則化パラメタである. 以上より, KFDA は次の最小化問題として定式化される:

$$\min_{\boldsymbol{\alpha}}\ \log|\boldsymbol{\alpha}^T(V_w+\zeta K)\boldsymbol{\alpha}|\quad\text{subject to}\quad|\boldsymbol{\alpha}^TV_b\boldsymbol{\alpha}|=const.$$

6.2 条件付きエントロピー最小基準による非線型次元削減

条件付きエントロピー最小化に基づく次元削減手法の非線型化を, KFDA と組み合わせることで実現する. KFDA によって定まるカーネル関数値の結合係数 $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_N)^T$ を初期値として, 特徴空間に分布するデータを判別軸に射影した値 $f(\boldsymbol{x})=\boldsymbol{a}^T\Phi(\boldsymbol{x})=\sum_{i=1}^N\alpha_ik(\boldsymbol{x},\boldsymbol{x}_i)$ のクラス条件付きエントロピーを $\boldsymbol{\alpha}$ に関して最適化する.

ここで, 最適化の目的関数であるクラス条件付きエントロピーに, 以下のような考えから導かれる正則化項を加える. KFDA によって得られた $\boldsymbol{\alpha}=\boldsymbol{\alpha}_0$ を条件付きエントロピー最小化基準で修正するということは, 直観的には KFDA の最適基準を多少外れてでも, 各クラスでデータの分散が小さくなるように $\boldsymbol{\alpha}$ を調整するということと理解できる. このとき, KFDA で定まる射影軸上での各クラスの平均値を保ちつつ, クラス毎の分散が小さくなることが, 判別の観点からは望ましい. そこで, KFDA によって定まる射影軸上の各クラスのデータの平均を $\boldsymbol{\alpha}_0^T\bar{\boldsymbol{k}}^y, y=1,\ldots,C$ として, グラム行列を用いて表現した射影軸上のクラス中心 $\frac{1}{N_y}\sum_{\boldsymbol{x}_i\in D_y}\sum_{j=1}^N\alpha_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{\alpha}^T\bar{\boldsymbol{k}}^y$ と $\boldsymbol{\alpha}_0^T\bar{\boldsymbol{k}}^y$ との二乗誤差を正則化項として用いる. これは, 式 (7) において $\Psi(f,D)=\sum_{y=1}^C(\boldsymbol{\alpha}^T\bar{\boldsymbol{k}}^y-\boldsymbol{\alpha}_0^T\bar{\boldsymbol{k}}^y)^2$ としたことに相当する. つまり最小化の目的関数は

$$H(f(X)|Y)+\varepsilon\sum_{y=1}^{C}(\boldsymbol{\alpha}^T\bar{k}^y-\boldsymbol{\alpha}_0\bar{\boldsymbol{k}}^y)^2 \qquad (16)$$

であり, KFDA によって得られた解 $\boldsymbol{\alpha}$ を初期値として, 上式を最小化する $\boldsymbol{\alpha}$ を求める. 最小化は, 線型の場合と同様に勾配法を採用した. 目的関数 (16) の $\boldsymbol{\alpha}$ に関する導関数は, 条件付きエントロピーの導関数

$$\frac{\partial H(f(X)|Y)}{\partial\boldsymbol{\alpha}} = \frac{1}{h^2N}\sum_{y=1}^{C}\sum_{\boldsymbol{x}_j\in D_y}\frac{\sum_{\boldsymbol{x}_i\in D_y\setminus\{\boldsymbol{x}_j\}}e^{-\frac{|\boldsymbol{\alpha}^T(\boldsymbol{k}_j-\boldsymbol{k}_i)|^2}{2h^2}}\boldsymbol{v}_{ji}}{\sum_{\boldsymbol{x}_i\in D_y\setminus\{\boldsymbol{x}_j\}}e^{-\frac{|\boldsymbol{\alpha}^T(\boldsymbol{k}_j-\boldsymbol{k}_i)|^2}{2h^2}}},$$

$$\boldsymbol{v}_{ji}=(\boldsymbol{k}_j-\boldsymbol{k}_i)^T\boldsymbol{\alpha}\cdot(\boldsymbol{k}_j-\boldsymbol{k}_i)\in\mathbb{R}^N$$

と, 正則化項の導関数

$$\frac{\partial\Psi(f,D)}{\partial\boldsymbol{\alpha}}=2\left((\bar{\boldsymbol{k}}^y)^T\bar{\boldsymbol{k}}^y\cdot\boldsymbol{\alpha}-\boldsymbol{\alpha}_0^T\bar{\boldsymbol{k}}^y\cdot\bar{\boldsymbol{k}}^y\right)$$

から計算できる.

6.3 One-nearest-neighbor による判別

KFDA による判別結果と, KFDA によって得られた結合係数 $\boldsymbol{\alpha}_0$ を $H(f(X)|Y)$ のエントロピー最小化に基づき修正したものによる判別結果を表 3 に記す. KFDA に用いるカーネル関数は Gaussian カーネル

$$k(\boldsymbol{x}_j,\boldsymbol{x}_i)=\exp\left(-\lambda||\boldsymbol{x}_j-\boldsymbol{x}_i||^2\right) \qquad (17)$$

であり, カーネルパラメタ λ は線型の場合の変換後の次元数と同様に, 各データセットから取り出した 5 つの学習データセットを用いた 5 fold cross validation によって定めた. KFDA と提案手法によるその修正による判別結果を表 3 に示す. ただし, 非線型化により表 2 における線型次元削減手法による最高精度を上回ったデータセットのみを示している.

KFDA によって求めた $\boldsymbol{\alpha}_0$ から開始して条件付きエントロピー基準 (16) を最小化することで得た解は, 条件付きエントロピーの観点からはより良いものになっているが, 表 3 の ringnorm データに見られるように, 判別の観点からは KFDA を改悪する可能性がある. 実用上は, こうした改悪は以下のようにして避けられる. 上述のように, KFDA のカーネルパラメタは cross validation によって選択した. この cross validation を行う際に, KFDA 単体と, 条件付きエントロピー最小化による修正を施した結果を用いた判別精度を計算し, バリデーションセットに対して判別精度の低下が見られる際には提案手法の適用を控えれば良い.

表 3: Misclassification rate by KFDA and modified KFDA.

Data name	KFDA	modified KFDA
breast-cancer	**33.753**(4.9489)	33.917(5.1108)
diabetes	29.474(2.2658)	29.474(2.2658)
flare-solar	**35.608**(1.9819)	36.334(1.8159)
german	28.277(2.2762)	**28.046**(2.4585)
heart	21.116(3.7185)	**21.084**(3.7831)
ringnorm	**2.056**(0.4546)	9.739(5.3771)
waveform	11.673(0.7438)	**11.61**(0.7425)

7 おわりに

情報論的観点から, クラス条件付きエントロピー最小化に基づく教師付き次元削減手法を提案した. 提案手法は通常の FDA では正しく判別曲面を得られないような状況でも正しく機能しうることを実験的に示した. また, 提案する条件付きエントロピー最小化基準による次元削減によってクラス判別に有効な特徴的な低次元空間が得られる仕組みを理論的に考察した. 大規模なベンチマークデータを用いた実験により, 提案する手法によって, FDA や PCA といった既存の次元削減手法より優れた判別を可能とする特徴量が得られることを示した. 近年, 局所性を考慮した次元削減手法が盛んに研究されており, 特にデータ間の類似度関数を陽に用いるものとしては, LFDA の他に Locality Preserving Projection[12](LPP) やその非線型化である Laplacian Eigenmap[13](LE) がよく知られている. LFDA は FDA におけるクラス間, クラス内分散比の最大化基準を, 局所性を考慮して定式化したものである. LPP 及び LE は, 低次元空間に射影したデータ間の距離を, データの類似度で重みつけて最小化する手法であり, データから構成されるグラフのラプラシアンを用いて, 一般化固有値問題として定式化される. 一方, 提案手法は目的関数が変換後のデータのクラス条件付きエントロピーであり, その定義には陽に局所性が含まれていない. しかし, カーネル密度関数に基づくデータの密度推定を介してデータ分布の局所構造が自然に反映され, それにより類似度関数を陽に用いた LFDA と同程度の判別力を有する方法になっていると考えられる.

さらに, 非線型次元削減 (特徴抽出) への拡張として, カーネル法に基づく一つの方法を提案した. ここで提案した方法では, 表 2 に示したようにほとんどのデータに関して提案手法による有意な精度向上は見られず, また, 精度が劣化するケースもあった. 一方, 紙面の制約上, 本報告には載せられなかったが, カーネル関数は正定値性を保つような多くの変換について再びカーネル関数となる [14] という性質を用いて, 複数のカーネル関数の組み合わせ最適化を行う multiple kernel learning (MKL) を, 条件付きエントロピー最小化基準によって行うことができる. これは, 今までは主に SVM のマージン最適化に基づいて行われてきた MKL (e.g., [15]) に対する新しいアプローチであり, 実データを用いた実験により, 既存の MKL 手法と同程度の判別結果が得られている.

参考文献

[1] K.Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[2] R.A.Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936.

[3] M.Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.*, 8:1027–1061, 2007.

[4] T.M.Cover and J.A.Thomas. *Elements of information theory*. John Wiley and Sons, Inc., 1991.

[5] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Meulen. Nonparametric entropy estimation: An overview. *International Journal of the Mathematical Statistics Sciences*, 6:17–39, 1997.

[6] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall/CRC, December 1994.

[7] A. Hyvärinen. Fast and robust fixed point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[8] A.Hyvärinen, J.Karhunen, and E.Oja. *Independent Component Analysis*. J. Wiley, New York, 2001.

[9] N. Kumar and A.G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Commun.*, 26(4):283–297, 1998.

[10] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[11] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müllers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48, 1999.

[12] X. He and P. Niyogi. Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[13] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[14] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[15] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004.