

カーネル多様体の重みつき最小二乗フィッティングと ファジークラスタリング

リラトールライサ[†] 加藤 毅[†] 都丸 拓馬[†] 太田 直哉[†]

[†] 群馬大学大学院工学研究科 〒 376-8515 群馬県桐生市天神町 1-5-1

E-mail: [†]relator-raissa@kato-lab.cs.gunma-u.ac.jp, ^{††}{katotsu, ohta}@cs.gunma-u.ac.jp,

^{†††}tomaru@ail.cs.gunma-u.ac.jp

あらまし 部分空間やアファイン超平面の最小二乗推定は、データの分布をモデル化する手法として次元圧縮やパターン認識などのために広く使われている。一方、近年、非一様な損失の最小化の有用性が示され、注目を浴びている。本稿では、再生核ヒルベルト空間において重みつき最小二乗推定を行うための算法を与える。また、重みつき最小二乗推定を要素技術としてファジークラスタリングに応用する。最後に、異常検知への応用における実験結果を与え、ファジークラスタリングによって学習したモデルがクラスタ数によらず高精度に検出できることが示す。

キーワード カーネル部分空間、カーネルアファイン超平面、重みつき最小二乗推定、ファジークラスタリング、異常検出

1. はじめに

近年、データマイニングの対象となるデータの次元数は飛躍的に増加しており、データの次元圧縮の必要性はますます高まっている。次元圧縮の技術は、パターン認識 [8]、異常検出 [9], [10]、雑音除去 [11]、データ圧縮 [12]、可視化 [13] といった広い応用範囲を持つ。

次元圧縮は、典型的には、多様体を得られたデータ集合にフィッティングすることで行われる。もっともよく使われている次元圧縮法は主成分分析である。主成分分析によって、各データと射影点との距離の平均、すなわち、平均二乗誤差を最小化する部分空間やアファイン超平面^(注1)を得ることができる [14]。特に、超多クラス分類である漢字認識などでは、各クラスの部分空間を構成する方法が広く使われている [15]。部分空間やアファイン超平面のような線形な多様体モデルでは、非線形な多様体に分布するデータを表現できないので、非線形な多様体モデルを導入する研究もある [16]。

カーネル主成分分析 [2], [3], [17] は、非線形な多様体を直接表現するのではなく、非線形カーネルを介してデータを非線形な変換を施すことで、結果的に非線形な多様体のフィッティングを実現している。カーネル主成分分析は線形カーネルを使えば通常の主成分分析になる。

一般に、陽に特徴ベクトルが算法に現れず、ベクトル間の内積のみで表されているような解析法はカーネル法 [18] と呼ばれている。また、線形の解析法を、線形カーネルを用いた時と同じ解析結果が得られ、かつ再生核ヒルベルト空間でも動作するように再定式化すること

がしばしば行われるが、これはカーネル化 (kernelize) と呼ばれている [18]。

カーネル化の利点は、非線形な解析が可能になるだけではない。カーネル法はベクトル間の内積のみで動作するので、 $d \times \ell$ の設計行列 (ℓ 個の d 次元特徴ベクトルを横に並べた行列) は必要とせず、 $\ell \times \ell$ のカーネル行列 (ベクトル間の内積の値を保持する対称行列) さえあれば学習できる。よって、 $d \gg \ell$ のとき、計算時間、記憶容量の両面で有利である。実際に、高次元の特徴ベクトルで陽に特徴抽出するよりも、データから直接ベクトル間の内積の値を効率的に計算したほうがよい場合がある。たとえば、文字列カーネル [19] や グラフカーネル [20], [21] など構造化データと呼ばれるデータのカーネルの多くは、この効果を利用して設計されている [22]。

本論文では、カーネル化可能な形で、部分空間およびアファイン超平面の重みつきフィッティングを行う算法を与えることである。すなわち、 ℓ 個の所与の訓練用例題それぞれに対して非負の重みが与えられているとき、重みつき平均二乗誤差が最小になる部分空間もしくはアファイン超平面を求めるのである (図 1 参照)。多くの学習機械は訓練用例題に対して平均損失が最小になるようにパラメータの値を決定している、重みつき平均損失に変更した場合、多くの解析法では、自明な導出でカーネル化できる。しかし、再生核ヒルベルト空間における部分空間およびアファイン超平面の重みつき平均二乗誤差の最小化の導出は必ずしも自明とはいえない (2. 節および 3. 節参照)。

線形多様体の重みつき最小二乗フィッティングは、様々な解析手法に対して重要な要素技術になりうる。近年、非一様な損失を用いて学習する試みが報告されており [23], [24]、特に、学習用例題と、実際の解析対象の密

(注1): 超平面というと、 n 次元空間に対して、 $n-1$ 次元の線形多様体のみをさすことがあるが、本稿ではさらに低い次元の線形多様体も超平面と呼ぶこととする。

表 1 線形多様体を用いたクラスタリング手法 .

	部分空間	カーネル部分空間	アファイン超平面	カーネルアファイン超平面
クラスタリングなし	Oja [1]	Tsuda [2]	Oja [1]	前田 & 村瀬 [3]
ハードクラスタリング	Ho et al [4]	Li & Fukui [5]	Lu & Vidal [6]	New
ファジークラスタリング	(Li et al [7])	New	New	New

度比が異なる場合、各学習用例題を密度比で重みづけして、重みつき平均損失に基づいて学習することで解析性能が向上することが知られている [25], [26] . そのほかにも、重みつき最小二乗フィッティングはファジークラスタリングへの応用も可能にする . ファジークラスタリングにおいて、各クラスタモデルの学習の際、重みつき平均二乗誤差が必要となる . 5. 節では、重みつきフィッティングの応用例として、異常検出問題 [9], [10] を取り上げ、ファジークラスタリング [8] によって学習した正常クラスモデルが、ほかのモデルと比べて、良好な性能を得ることを示す .

部分空間やアファイン平面を使ったクラスタリングとカーネル法の技術の組み合わせを使ったこれまでの研究は表 1 のようになっている . 部分空間やアファイン超平面へのフィッティングは長年にわたって用いられてきたが、90 年代末期にこれをカーネル主成分分析を使って再生核ヒルベルト空間でも利用できるように拡張された [2], [3], [17] . さらに 2003 年に Ho et al [4] によって部分空間モデルによるクラスタリングに、さらに、2006 年に Lu & Vidal [6] によってアファイン超平面モデルによるクラスタリングに応用された . 2007 年には Li & Fukui [5] によって Ho et al のモデル [4] のカーネル化が施された . 部分空間を使ったファジークラスタリングは Li et al [7] によって行われている . Li et al は要素技術として重みつき特異値分解を開発し、これをファジークラスタリングに応用している . しかし、Li et al は設計行列の要素ごとに重みづけする問題に挑戦しており、例題ごとに重みづけする本研究とは問題設定が異なっている . Li et al の手法は反復的に特異値分解を求めるもので最適解に辿り着く理論的保証は与えられていない . これに対して、本稿で提案する重みつき最小二乗フィッティングは閉じた形で最適解を与えているので、必ず重みつき平均二乗誤差が最小化される . 重みつき主成分分析 [27], [28] も提案されている . しかし、導出には次元が 1 のアファイン部分空間のみに対して、行われている . 本稿では、異なる証明方法を使って、任意の次元数の部分空間に対する導出を明示的に与えるものである .

なお、現状では、部分空間クラスタリング (subspace clustering) という言葉は二つの全く異なる意味で用いられていることに注意されたい . 一つは、各クラスタモデルが部分空間になっているもの [4], [5] で、もう一つは、各クラスタが高次元中の少数の次元のみから構成されているものである [29] . 筆者らが調査する限り、後者の意味で使われることのほうが多いようである .

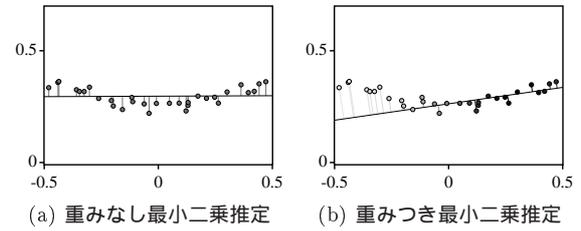


図 1 アファイン超平面のフィッティング . (a) は重みなし最小二乗推定の結果、(b) は重みつき最小二乗推定の結果である . 重みは $\exp(-(x - 0.3)^2/8)$ で与えた . 各点の濃度は重みの大きさを示している . 濃い色の点ほど多様体との距離がより小さくなるように学習されているのが分かる .

本稿では次の記法を用いる . 太字の小文字はベクトルを表す . 太字の大文字は行列を表す . ベクトルや行列の要素には細字を用いる . 行列 A の転置は A^T と書く . 行列 A の逆行列は A^{-1} と書く . $n \times n$ の単位行列は I_n と書く . n 次元ベクトルですべての要素が 1 のものを $\mathbf{1}_n$ と書く . n 次元ベクトルですべての要素が 0 のものを $\mathbf{0}_n$ と書く . \mathbb{R} は実空間、 \mathbb{N} は自然数の集合である . \mathbb{R}^n は n 次元実ベクトル空間、 $\mathbb{R}^{m \times n}$ は $m \times n$ 実数行列の集合、 \mathbb{R}_+ , \mathbb{R}_{++} はそれぞれ非負および正の実数の集合を表す . 任意の自然数 $n \in \mathbb{N}$ に対し、 \mathbb{N}_n は n 以下の自然数の集合を表すのに用いる . \mathbb{S}^n は $n \times n$ 対称行列を表す . $d \times m$ 正規直交行列を $\mathbb{O}^{d \times m} \equiv \{U \in \mathbb{R}^{d \times m} \mid U^T U = I_m\}$ と定義する . Δ_n は確率単体を表す : $\Delta_n \equiv \{x \in \mathbb{R}_+^n \mid \mathbf{1}_n^T x = 1\}$. \mathcal{P}_K は \mathbb{N}_K の全単射写像の集合である . D_x は対角要素に x を持つ対角行列を表す . つまり、 $D_x = \text{diag}(x)$. シンボル \leq および \geq はスカラー間の不等号のみならず、ベクトルの要素間の不等号も表す .

2. 部分空間の重みつきフィッティング

本節では、部分空間の重みつき最小二乗推定法をカーネル法に直接拡張できる形で与える .

部分空間は、 $d \times m$ 個の正規直交行列 $U \in \mathbb{O}^{d \times m}$ を使って、

$$S_{ss}(U) = \{y \in \mathbb{R}^d \mid \exists \alpha \in \mathbb{R}^m \text{ s.t. } y = U\alpha\}$$

のようにあらわされる . 任意のベクトル $x \in \mathbb{R}^d$ と部分空間との二乗ユークリッド距離は

$$d(x, S_{ss}(U)) = \min_{y \in S_{ss}(U)} \|x - y\|^2 = \|(I_d - UU^T)x\|^2$$

で与えられる [2], [5] . 例題集合 $\{x_i\}_{i=1}^{\ell}$ に対して、重みなし平均二乗誤差

$$\frac{1}{\ell} \sum_{i=1}^{\ell} d(x, \mathcal{S}_{\text{ss}}(U))$$

を最小にするパラメータは次のように求められる：行列 $\sum_{i=1}^{\ell} x_i x_i^{\top}$ を計算し、この行列の主要な m 個の固有ベクトルを列とする行列で正規直交行列 U を構成する。

本節の目的は、各例題に重み $v_i \in \mathbb{R}_+$ が所与として、重みつき平均二乗誤差

$$J_{\text{ss}}(U; v) \equiv \frac{\sum_{i=1}^{\ell} v_i d(x_i, \mathcal{S}_{\text{ss}}(U))}{\sum_{i=1}^{\ell} v_i}$$

を最小にする U をカーネル法に直接拡張できる形で与えることである。

訓練用例題集合から $d \times \ell$ の設計行列 X を $X \equiv [x_1, \dots, x_{\ell}]$ と構成する。任意のベクトル $x, y \in \mathbb{R}^d$ に対して、関数 $K(x, y)$ を2つのベクトルの内積、すなわち、 $K(x, y) \equiv \langle x, y \rangle$ と定義する。学習機械の中で、陽に特徴ベクトルが現れず、関数 $K(x, y)$ のみで表されるように定式化したものはカーネル法と呼ばれる[18]。この内積の定義を任意の正定値カーネルに置き換えることで、非線形化など様々な効果を得ることができる。訓練用例題集合 $\{x_i\}_{i=1}^{\ell}$ に対し、第 (i, j) 要素が $K_{i,j} = \langle x_i, x_j \rangle$ なる $\ell \times \ell$ の対称行列を $K \in \mathbb{S}^{\ell}$ と定義する。この行列 K は必ず半正定値になることが知られている。また、 ℓ 次元のベクトルを返すベクトル値関数 $k: \mathbb{R}^d \mapsto \mathbb{R}^{\ell}$ を $k(x) \equiv [\langle x_1, x \rangle, \dots, \langle x_{\ell}, x \rangle]^{\top}$ と定義する。

部分空間の重みつき最小二乗推定は次のように与えられる。

定理 2.1. 対称行列 \tilde{K} を $\tilde{K} \equiv D_v^{1/2} K D_v^{1/2}$ と定義する。 $r \equiv \text{rank}(\tilde{K})$ とし、 $r \geq m$ と仮定する。 \tilde{K} の最も大きい m 個の固有値を $\lambda_1, \dots, \lambda_m \in \mathbb{R}_{++}$ とし、それぞれの固有値に対応する固有ベクトルを $b_1, \dots, b_m \in \mathbb{R}^{\ell}$ とする。 $\lambda \equiv [\lambda_1, \dots, \lambda_m]^{\top}$ および $B \equiv [b_1, \dots, b_m]$ とおき、

$$\hat{U}_{\text{ss}} \equiv X D_v^{1/2} B \text{diag}(\lambda)^{-1/2}$$

とすると、

$$\hat{U}_{\text{ss}} \in \underset{U \in \mathbb{O}^{d \times m}}{\text{argmin}} J_{\text{ss}}(U; v)$$

が成り立つ。□

(Proof sketch) 式変形により $J_{\text{ss}}(U; v) = \frac{1}{2} \text{tr}(X D_v X^{\top}) - \frac{1}{2} \text{tr}(U^{\top} X D_v X^{\top} U)$ となるので、 U の列ベクトルは対称行列 $X D_v X^{\top}$ の m 個の主要なベクトルになる。固有ベクトルは X の列ベクトルのうち $v_i > 0$ なるベクトルの張る部分空間上にあるので、 $U = X D_v^{1/2} A$ なる $A \in \mathbb{R}^{\ell \times m}$ が存在するはず。これを $J_{\text{ss}}(U; v)$ に代入すると、次の最適化問題に帰着される：

$$\begin{aligned} & \max \text{tr} \left(A^{\top} \tilde{K}^2 A \right), \quad \text{wrt } A \in \mathbb{R}^{\ell \times m}, \\ & \text{subj to } A^{\top} \tilde{K} A = I_m. \end{aligned}$$

この解となる行列の列は、それぞれ、 \tilde{K} の m 個の主要なベクトルに比例することから、

$$\forall h \in \mathbb{N}_m, \exists t_h \in \mathbb{R}_+ : a_h = t_h b_h.$$

を得る。よって、 $\forall h \in \mathbb{N}_m$ に対して、

$$1 = \|u_h\|^2 = \|X D_v^{1/2} t_h b_h\|^2 = t_h^2 \text{tr} \left(b_h^{\top} \tilde{K} b_h \right) = t_h^2 \lambda_h$$

となるため、 $t_h = \lambda_h^{-1/2}$ を得る。したがって、 $\forall h \in \mathbb{N}_m$ に対して、 $u_h = \lambda_h^{-1/2} X D_v^{1/2} b_h$ となり、これを行列で表現すると題意は示される。□

カーネル法に拡張できるようにするためには陽に特徴ベクトルを含まないように表さなくてはならないが、定理 2.1 の結果は、基底ベクトルの行列 \hat{U}_{ss} は、陽に特徴ベクトルを含んだままである。しかし、実際に部分空間を用いる場合は、基底ベクトルそのものは必要なく、任意のベクトルから部分空間への距離さえ計算できればよい。

系 2.1.

$$d(x, \mathcal{S}_{\text{ss}}(\hat{U}_{\text{ss}})) = K(x, x) - \|\text{diag}(\lambda)^{-1/2} A D_v^{1/2} k(x)\|^2.$$

□

よって、 \hat{U}_{ss} の列が張る部分空間 $\mathcal{S}_{\text{ss}}(\hat{U}_{\text{ss}})$ との距離 $d(x, \mathcal{S}_{\text{ss}}(\hat{U}_{\text{ss}}))$ は陽に特徴ベクトルを含まず、内積のみで表されているのでカーネル化可能である。

入力空間に対して線形な部分空間でよく、 $d \ll \ell$ の場合、カーネルを用いないほうが効率的に計算できる。定理 2.1 では、 $\ell \times \ell$ 行列の固有値分解を要したが、次の系は $d \times d$ 行列の固有値分解で済むことを示している。

系 2.2. 定理 2.1 で与えた \hat{U}_{ss} の m 個の列ベクトルは、 $\sum_{i=1}^{\ell} v_i x_i x_i^{\top}$ の主要な m 個の固有ベクトルである。また、 $\sum_{i=1}^{\ell} v_i x_i x_i^{\top}$ の主要な m 個の固有ベクトルは、 $J_{\text{ss}}(U; v)$ を最小化する。□

3. アファイン超平面の重みつきフィッティング

本節では、アファイン超平面に対する重みつき最小二乗推定法について議論する。アファイン超平面はオフセットベクトル $\mu \in \mathbb{R}^d$ と $d \times m$ 個の正規直交行列 $U \in \mathbb{O}^{d \times m}$ を使って、

$$\mathcal{S}_{\text{sa}}(U, \mu) = \{y \in \mathbb{R}^d \mid \exists \alpha \in \mathbb{R}^m \text{ s.t. } y - \mu = U\alpha\}$$

のようにあらわされる。任意のベクトル $x \in \mathbb{R}^d$ とアファイン超平面との二乗ユークリッド距離は

$$\begin{aligned} d(x, \mathcal{S}_{\text{sa}}(U, \mu)) &= \min_{y \in \mathcal{S}_{\text{sa}}(U, \mu)} \|x - y\|^2 \\ &= \|(\mathbf{I}_d - U U^{\top})(x - \mu)\|^2 \end{aligned}$$

で与えられる。

各例題に重み $v_i \in \mathbb{R}_+$ が所与として、重みつき平均二乗誤差

$$J_{\text{sa}}(\mathbf{U}, \boldsymbol{\mu}; \mathbf{v}) \equiv \frac{\sum_{i=1}^{\ell} v_i d(\mathbf{x}_i, \mathcal{S}_{\text{sa}}(\mathbf{U}, \boldsymbol{\mu}))}{\sum_{i=1}^{\ell} v_i}$$

を最小にする $(\mathbf{U}, \boldsymbol{\mu})$ を求めたい。 \mathbf{v} に定数をかけても最適解集合は不変なので、 $\sum_{i=1}^{\ell} v_i = 1$ を仮定する。各訓練用例題から d -次元ベクトル $X\mathbf{v}$ だけ差し引いたものを

$$\forall i \in \mathbb{N}_{\ell} : \quad \bar{\mathbf{x}}_i \equiv \mathbf{x}_i - X\mathbf{v}$$

と表すことにし、さらに、 $d \times \ell$ の行列 \bar{X} を

$$\bar{X} \equiv [\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_{\ell}]$$

と定義する。前節では、関数 $K, k(\cdot)$ 、行列 K は、ベクトル間の内積を表すために用意したが、各特徴ベクトルから $X\mathbf{v}$ だけ差し引いたものどうしの内積を表すために、関数 $\bar{K} : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ 、 $\bar{k}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^{\ell}$ を

$$\begin{aligned} \bar{K}(\mathbf{x}, \mathbf{y}) &\equiv \langle \mathbf{x} - X\mathbf{v}, \mathbf{y} - X\mathbf{v} \rangle, \\ \bar{k}(\mathbf{x}) &\equiv [\langle \bar{\mathbf{x}}_1, \mathbf{x} - X\mathbf{v} \rangle, \dots, \langle \bar{\mathbf{x}}_{\ell}, \mathbf{x} - X\mathbf{v} \rangle]^{\top} \end{aligned}$$

とおく。また、 $\ell \times \ell$ の対称行列 \bar{K} を第 (i, j) 要素が $\langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j \rangle$ となるように定義する。

アファイン超平面の重みつき最小二乗推定は次のように与えられる。

定理 3.1. 対称行列 \bar{K} を $\bar{K} \equiv D_v^{1/2} \bar{K} D_v^{1/2}$ と定義する。 $r \equiv \text{rank}(\bar{K})$ とし、 $\bar{r} \geq m$ と仮定する。 \bar{K} の最も大きい m 個の固有値を $\bar{\lambda}_1, \dots, \bar{\lambda}_m \in \mathbb{R}_{++}$ とし、それぞれの固有値に対応する固有ベクトルを $\bar{\mathbf{b}}_1, \dots, \bar{\mathbf{b}}_m \in \mathbb{R}^{\ell}$ とする。 $\bar{\boldsymbol{\lambda}} \equiv [\bar{\lambda}_1, \dots, \bar{\lambda}_m]^{\top}$ および $\bar{B} \equiv [\bar{\mathbf{b}}_1, \dots, \bar{\mathbf{b}}_m]$ とおき、

$$\hat{U}_{\text{sa}} \equiv \bar{X} D_v^{1/2} \bar{B} \text{diag}(\bar{\boldsymbol{\lambda}})^{-1/2}, \quad \hat{\boldsymbol{\mu}}_{\text{sa}} \equiv X\mathbf{v},$$

とすると、

$$(\hat{U}_{\text{sa}}, \hat{\boldsymbol{\mu}}_{\text{sa}}) \in \underset{\mathbf{U} \in \mathbb{O}^{d \times m}, \boldsymbol{\mu} \in \mathbb{R}^d}{\text{argmin}} J_{\text{sa}}(\mathbf{U}, \boldsymbol{\mu}; \mathbf{v})$$

が成り立つ。 □

(Proof sketch) まず、 $J_{\text{sa}}(\mathbf{U}, \boldsymbol{\mu}; \mathbf{v})$ を $\boldsymbol{\mu}$ で偏微分して $\mathbf{0}_d$ とおくと、任意の $\boldsymbol{\beta} \in \mathbb{R}^m$ に対して $\boldsymbol{\mu} = X\mathbf{v} + \mathbf{U}\boldsymbol{\beta}$ が最適値になることが分かる。 $\boldsymbol{\beta} = \mathbf{0}_m$ において、 $J_{\text{sa}}(\mathbf{U}, \boldsymbol{\mu}; \mathbf{v})$ に代入すると定理 2.1 と同様な手順で導くことができる。 □

系 3.1.

$$\begin{aligned} d(\mathbf{x}, \mathcal{S}_{\text{sa}}(\hat{U}_{\text{sa}}, \hat{\boldsymbol{\mu}}_{\text{sa}})) \\ = \bar{K}(\mathbf{x}, \mathbf{x}) - \|\text{diag}(\bar{\boldsymbol{\lambda}})^{-1/2} \bar{B} D_v^{1/2} \bar{k}(\mathbf{x})\|^2. \end{aligned}$$

□

よって、重みつき最小二乗推定された部分空間およびアファイン超平面とのユークリッド距離も閉形式で求まり、カーネル法に直接拡張できることが示された。

カーネルを用いない算法も次の系に与えておく。

系 3.2. 定理 3.1 で与えた \hat{U}_{sa} の m 個の列ベクトルは、 $d \times d$ 対称行列 $\sum_{i=1}^{\ell} v_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{sa}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{sa}})^{\top}$ の主要な m 個の固有ベクトルである。また、 $\sum_{i=1}^{\ell} v_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{sa}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{sa}})^{\top}$ の主要な m 個の固有ベクトルは、 $J_{\text{sa}}(\mathbf{U}, \boldsymbol{\mu}; \mathbf{v})$ を最小化する。 □

4. 異常検出への応用

4.1 問題設定

異常検出は、正常クラスの訓練用例題のみから学習を行う問題である。正常クラスの訓練用例題のみから学習を行うのは、異常検出における応用においてこれはどのような異常データが現れるか予測できないので、あらかじめ手元にある異常データを使って学習すると予期せぬ異常データを異常と予測できなくなるからである [30]。 ℓ 個の正常クラスの訓練用例題が所与として $\mathbf{x}_i \in \mathbb{R}^d$ ($i = 1, \dots, \ell$) と書くことにする。これらの訓練用例題から正常クラスモデルを構築し、正常クラスモデルへの距離によって異常か正常か判別することができる。正常クラスモデルは特徴ベクトルの空間の部分集合 $S(\Theta) (\subset \mathbb{R}^d)$ で与えられるとする。ただし、 Θ はモデルのパラメータである。未知の例題 $\mathbf{x} \in \mathbb{R}^d$ が与えられたとき、 \mathbf{x} と $S(\Theta)$ との二乗ユークリッド距離

$$d(\mathbf{x}, S(\Theta)) \equiv \min_{\mathbf{y} \in S(\Theta)} \|\mathbf{x} - \mathbf{y}\|^2$$

の大きさを正常か異常か予測することとする。ユークリッド距離の代わりに、マハラノビス距離を使う方法 [31] や距離関数を学習する方法 [32] もあるが、これらは回転不変ではないためカーネル法に拡張できない。

4.2 正常クラスモデルへの距離

正常クラス集合のモデル $S(\Theta)$ の与え方をいくつか紹介する。以下にあげるモデルのパラメータは、訓練用例題との平均二乗誤差

$$J(\Theta) \equiv \frac{1}{\ell} \sum_{i=1}^{\ell} d(\mathbf{x}_i, S(\Theta))$$

を最小化するように求める。

部分空間モデル

部分空間を正常クラスモデルとして用いる場合、 $S(\Theta) = \mathcal{S}_{\text{ss}}(\mathbf{U})$ となり、モデルパラメータは $\Theta = \mathbf{U}$ となる。このモデルの学習には 2. 節において $\mathbf{v} = \mathbf{1}_{\ell}/\ell$ とおけばよい。

アファイン超平面モデル

アファイン超平面を正常クラスモデルとして用いる場合、 $\mathcal{S}(\Theta) = \mathcal{S}_{\text{sa}}(U, \mu)$ となり、モデルパラメータは $\Theta = (U, \mu)$ となる。このモデルの学習には 3. 節において $v = \mathbf{1}_\ell / \ell$ とおけばよい。

部分空間集合モデル

部分空間集合モデルは、 K_{clst} 個の部分空間の和集合 $\mathcal{S}_{\text{ms}}(U) = \bigcup_{k=1}^{K_{\text{clst}}} \mathcal{S}_{\text{ss}}(U_k)$ によって、正常クラスの集合モデルを形成する。ただし、 $U = \{U_k\}_{k \in \mathbb{N}_{K_{\text{clst}}}}$ とした。モデルパラメータは $\Theta = U$ となる。Ho et al のクラスタリング手法 [4] を使うと、 U の局所最適解を求めることができる。この手法は、 K -平均法と類似した算法を用いて局所最適解を探索するものである。

アファイン超平面集合モデル

K_{clst} 個のアファイン超平面の和集合 $\mathcal{S}_{\text{ma}}(U, M) = \bigcup_{k=1}^{K_{\text{clst}}} \mathcal{S}_{\text{sa}}(U_k, \mu_k)$ によって、正常クラスの集合モデルを形成することもできる。ただし、 $U = \{U_k\}_{k \in \mathbb{N}_{K_{\text{clst}}}}$ 、 $M = [\mu_1, \dots, \mu_{K_{\text{clst}}}] \in \mathbb{R}^{d \times K_{\text{clst}}}$ とした。モデルパラメータは $\Theta = \{U, M\}$ となる。このモデリングによる方法も K -平均法と似たような算法を使って局所最適解を探索することができる [6]。

4.3 クラスタリングによる学習

前節に紹介したモデルのうち、部分空間集合モデル、アファイン超平面集合モデルは、いずれもモデルとの距離を

$$d(x, \mathcal{S}(\Theta)) = \min_{k \in \mathbb{N}_{K_{\text{clst}}}} d(x, \mathcal{S}_k(\Theta_k)) \quad (1)$$

のように書くことができる。ただし、 $\mathcal{S}_k(\Theta_k)$ は第 k クラスターのモデルを示しており、部分空間集合モデルの場合 $\mathcal{S}_k(\Theta_k) = \mathcal{S}_{\text{ss}}(U_k)$ 、アファイン超平面集合モデルの場合 $\mathcal{S}_k(\Theta_k) = \mathcal{S}_{\text{sa}}(U_k, \mu_k)$ となる。訓練用例題との距離の平均は

$$J(\Theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} \min_{k_i \in \mathbb{N}_{K_{\text{clst}}}} d(x_i, \mathcal{S}_{k_i}(\Theta_{k_i}))$$

と表すことができる。 $J(\Theta)$ を最小化することは、訓練用例題集合の最適な排他的な分割 $\{\mathcal{I}_k\}_{k=1}^{K_{\text{clst}}}$ (例題の添え字集合 \mathbb{N}_ℓ を $\bigcup_{k=1}^{K_{\text{clst}}} \mathcal{I}_k = \mathbb{N}_\ell$ 、かつ $\forall k \in \mathbb{N}_{K_{\text{clst}}}, \forall k' \in \mathbb{N}_{K_{\text{clst}}} \setminus \{k\}: \mathcal{I}_k \cap \mathcal{I}_{k'} = \emptyset$ を満たすように K_{clst} グループに分けるような分割) を求めることに他ならない。ひとたび、最適な分割 $\{\mathcal{I}_k^*\}_{k=1}^{K_{\text{clst}}}$ が見つければ

$$\min_{\Theta} J(\Theta) = \frac{1}{\ell} \sum_{k=1}^{K_{\text{clst}}} \min_{\Theta_k} \sum_{i \in \mathcal{I}_k^*} d(x_i, \mathcal{S}_k(\Theta_k))$$

となるので、モデルパラメータ Θ の学習の際、各訓練用例題は、 K_{clst} 個のクラスターのモデルのうち 1 個だけ寄与することになり、各クラスターのモデルはほかのクラスターのモデルと訓練用例題を共有しないことになる。

クラスター間で訓練用例題を共有するモデルとしては、自己組織化マップ (SOM) [33] があげられる。SOM では、各訓練用例題が複数のクラスターに寄与するように学習算法が設計されているので、各クラスターが“繋がった”ように学習される。これによって、正常クラスのデータが分布する連続した多様体を学習することができると期待される。しかし、現実の問題は必ずしも各クラスターが“繋がっている”とは限らないので、本研究では別のアプローチで正常クラスの多様体を学習することにする。

κ -距離によるクラスタリング

次のようなあらかじめ重み係数 $\kappa \in \mathbb{R}_+^{K_{\text{clst}}}$ を定めておく：

$$\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_{K_{\text{clst}}} \geq 0. \quad (2)$$

学習時には、モデルとの距離として、式 (1) の代わりに、次のように拡張したものをを用いることとする：

$$d(x, \mathcal{S}(\Theta)) = \min_{\pi \in \mathcal{P}_{K_{\text{clst}}}} \sum_{k=1}^{K_{\text{clst}}} \kappa_{\pi(k)} d(x, \mathcal{S}_{\pi(k)}(\Theta_{\pi(k)})). \quad (3)$$

すなわち各クラスターとの距離の線形和に置き換えたのである。式 (3) で与えた距離を κ -距離と呼ぶことにする。重み係数を、 $\kappa_1 = 1$ 、 $\kappa_k = 0 (\forall k \geq 2)$ とおけば式 (1) と等しくなるので、 κ -距離は、ユークリッド距離 (1) の一般化といえる。式 (2) の仮定から、最適な順列 π^* は

$$d(x, \mathcal{S}_{\pi^*(1)}(\Theta_{\pi^*(1)})) \leq \dots \leq d(x, \mathcal{S}_{\pi^*(K_{\text{clst}})}(\Theta_{\pi^*(K_{\text{clst}})}))$$

を満たす。ひとたび各訓練用例題に対する最適な順列 π_i^* ($\forall i \in \mathbb{N}_\ell$) が見つかったならば、訓練用例題との距離の総和の最小値は

$$\min_{\Theta} J(\Theta) = \frac{1}{\ell} \sum_{k=1}^{K_{\text{clst}}} \min_{\Theta_k} \sum_{i=1}^{\ell} v_{k,i} d(x_i, \mathcal{S}_k(\Theta_k)) \quad (4)$$

と表される。ただし、 $v_{k,i}$ は

$$\forall i \in \mathbb{N}_\ell, \forall k \in \mathbb{N}_{K_{\text{clst}}}: v_{\pi^*(k), i} = \kappa_k \quad (5)$$

を満たす係数である。式 (4) に表されているように、各クラスターモデルは、各例題を重み $v_{k,i}$ 分だけ取り込むことになるので、クラスターモデル間で例題を共有することになり、なめらかにつながったモデルが形成される。

κ -距離による $J(\Theta)$ を最小化するモデルパラメータを求める Block coordinate ascent 法 [34] は次のようになる。

算法 4.1. 次の 2 つのステップを収束するまで繰り返す。

- ステップ 1: 各例題 $i \in \mathbb{N}_\ell$ に対して順列 $\pi_i^{(t+1)}$ を

$$\forall k \in \mathbb{N}_{K_{\text{clst}}-1}: d(x_i, \mathcal{S}_{\pi_i^{(t+1)}(k)}(\Theta_{\pi_i^{(t+1)}(k)})) \leq d(x_i, \mathcal{S}_{\pi_i^{(t+1)}(k+1)}(\Theta_{\pi_i^{(t+1)}(k+1)}))$$

を満たすように更新し、式 (5) を使って $v_{k,i}^{(t+1)}$ を計算する。

- ステップ 2: 各クラスタのモデルパラメータ Θ_k を更新する:

$$\forall k \in \mathbb{N}_{K_{\text{clst}}} : \Theta_k^{(t+1)} := \operatorname{argmin}_{\Theta_k} \sum_{i=1}^{\ell} v_{k,i}^{(t+1)} d(x_i, \mathcal{S}_k(\Theta_k))$$

□

この算法から得られる数列 $\{J(\Theta^{(t)})\}_{t \in \mathbb{N}}$ は単調減少することが保証されている。また、 $|\mathcal{P}_{K_{\text{clst}}}| < \infty$ なので、 $J(\Theta^{(t)})$ は有限回の反復で一定値となる。

ステップ 2 におけるモデルパラメータの更新の際は、部分空間集合モデルの場合は、各クラスタモデルに定理 2.1 および系 2.1 を適用すればよく、アフライン超平面集合モデルの場合は、各クラスタモデルに定理 3.1 および系 3.1 を適用すればよい。

ファジー距離によるクラスタリング

クラスタモデル間で例題を共有するもう一つの方法としては、1 以上の定数 $b_{\text{bez}} \in \mathbb{R}$ を使って、式 (1) を次のように拡張することが考えられる:

$$d(x, \mathcal{S}(\Theta)) = \min_{w \in \Delta_{K_{\text{clst}}}} \sum_{k=1}^{K_{\text{clst}}} w_k^{b_{\text{bez}}} d(x, \mathcal{S}_k(\Theta_k)). \quad (6)$$

このような距離はファジークラスタリング [8] において用いられている。 $b_{\text{bez}} = 1$ の場合は、式 (1) に等しくなるが、 $b_{\text{bez}} > 1$ の場合、ラグランジェ乗数法を使うと、最適な重み w^* は

$$w_k^* \propto \frac{1}{(d(x, \mathcal{S}_k(\Theta_k)))^{1/(b_{\text{bez}}-1)}} \quad (7)$$

で与えられることを容易に示すことができる [8]。式 (6) で与えた距離をファジー距離と呼ぶことにする。ファジー距離によるクラスタリングの算法を次に与える。

算法 4.2. 次の 2 つのステップを収束するまで繰り返す。

- ステップ 1: 各例題に対して重み $w_i^{(t+1)}$ を式 (7) を使って更新する。 $v_{k,i}^{(t+1)} = ([w_i^{(t+1)}]_k)^{b_{\text{bez}}}$ とする。
- ステップ 2: 算法 4.1 のステップ 2 と同様にして、各クラスタのモデルパラメータ Θ_k を更新する。

□

ステップ 2 は算法 4.1 のステップ 2 と同じである。算法 4.2 から生成される数列 $\{J(\Theta^{(t)})\}_{t \in \mathbb{N}}$ は単調減少することが保証されている。また、 $J(\cdot)$ の定義より、この数列は非負であるので、算法 4.1 は必ず収束する。

5. 実験

前節で述べた各モデルが、異常検知という応用における検出能力を定量評価するため、顔画像、生活音、文字列の 3 種類のデータを用いた。

顔画像データは Extended Yale Face Database B に登録されているものを用いた。このデータベースには 39 人の顔画像が含まれている。一人当たり最大 60 枚の画

像が登録されている。このうち最初の 3 人を正常クラスとし、そのほか 36 人を異常クラスとした (図 2 参照)。このデータセットが含んでいる画像 $X, X' \in \mathbb{R}^{192 \times 168}$ に対して、カーネル関数として $K(X, X') = \operatorname{tr}(X^T X')$ を用いた。

生活音データは、風呂場から音を採取し、パワースペクトルを抽出して 256 次元の特徴ベクトルを得た。特徴ベクトルから線形カーネルを使ってカーネル値を計算した。このデータセットは正常クラスは 1,406 個からなり、異常クラスは 94 個からなる。

文字列データとしてアミノ酸配列を用いた。蛋白質立体構造データベース SCOP [35] から 12 種類のフォールドにおけるアミノ酸配列 3,427 本を用い、そのうち最初の 4 フォールドを正常クラス、それ以外を異常クラスとした。内訳は正常クラスは 909 本、異常クラスは 2,518 本になる。それぞれのアミノ酸配列から Lodhi et al の文字列カーネル [19] を用いてカーネル値を得た。Lodhi et al の文字列カーネルは特徴ベクトルを経ずに直接文字列から効率的に値を計算できるカーネルである。

正常クラスのデータのうち 80% を無作為に選んで訓練用例題とし、残りの正常クラスのデータと異常クラスのデータを評価用に用いた。部分空間、およびアフライン超平面を単独で用いるモデルは、累積寄与率 0.95 を超えない最大の多様体の次元 m を選んだ。部分空間集合モデル、およびアフライン超平面集合モデルを用いる場合は、クラスタ数は 10, 20, 30 を用いた。多様体の次元数は次のように決めた: まず、それぞれのクラスタ数で、single linkage によるクラスタリングを行い、次に、それぞれのクラスタで累積寄与率 0.95 を超えない最大の多様体の次元を選んだ。 κ -距離のパラメータは $\kappa_1 = 0.9$, $\kappa_2 = 0.1$ それ以外を $\kappa_k = 0$ とした。ファジー距離のパラメータは $b_{\text{bez}} = 2$ とした。

SS は部分空間モデル, SA はアフライン超平面モデル, MS- κ C および MA- κ C は、それぞれ、部分空間集合モデル、およびアフライン超平面集合モデルを κ -距離によるクラスタリングで学習したもの、MS-FC および MA-FC は、それぞれ、部分空間集合モデル、およびアフライン超平面集合モデルをファジー距離によるクラスタリングで学習したもの、MS-HC および MA-HC は、それぞれ、部分空間集合モデル、およびアフライン超平面集合モデルをハードクラスタリングで学習したものである。したがって、MS- κ C, MA- κ C, MS-FC, MA-FC が提案法になる。筆者らの知る限り、カーネル版においては MA-HC も新しい。それら以外が従来法となる。予測時にはすべてユークリッド距離を用いた。

性能を表 2 に結果をまとめた。表は ROC カーブ [36] の AUC を報告している。AUC の値は大きいほど予測性能がよいことをしめし、ランダムに予測した場合、0.5 になる。クラスタ数 10,20,30 の 3 種類のうち最高性能 (Highest) をとったものと、3 つの平均をとったもの

表 2 異常検出における性能評価．4 回リポートにおける AUC の平均を示している．カッコ内は標準偏差．太字は最高性能，下線は最高性能と統計的有意差がないもの．

顔画像	SS	MS- κ C	MS-FC	MS-HC	SA	MA- κ C	MA-FC	MA-HC
Highest	0.757 (0.034)	<u>0.954</u> (0.010)	<u>0.952</u> (0.004)	0.921 (0.018)	0.835 (0.028)	0.955 (0.007)	0.950 (0.007)	<u>0.911</u> (0.040)
Average	0.757 (0.034)	<u>0.924</u> (0.018)	<u>0.918</u> (0.020)	0.842 (0.026)	0.835 (0.028)	0.936 (0.012)	<u>0.930</u> (0.008)	0.860 (0.025)
生活音	SS	MS- κ C	MS-FC	MS-HC	SA	MA- κ C	MA-FC	MA-HC
Highest	0.846 (0.019)	0.955 (0.007)	0.944 (0.004)	0.964 (0.004)	0.846 (0.019)	0.993 (0.004)	0.993 (0.001)	<u>0.990</u> (0.006)
Average	0.846 (0.019)	0.911 (0.014)	0.943 (0.004)	0.908 (0.015)	0.846 (0.019)	<u>0.988</u> (0.005)	0.990 (0.001)	0.976 (0.004)
文字列	SS	MS- κ C	MS-FC	MS-HC	SA	MA- κ C	MA-FC	MA-HC
Highest	0.811 (0.014)	0.838 (0.017)	0.841 (0.017)	0.813 (0.015)	0.813 (0.014)	0.836 (0.017)	0.845 (0.017)	0.814 (0.015)
Average	0.811 (0.014)	0.836 (0.016)	0.834 (0.015)	0.811 (0.015)	0.813 (0.014)	0.831 (0.017)	0.838 (0.016)	0.813 (0.015)

(Average) の 2 種類で性能評価を行った．訓練用例題と評価用例題の分割を 4 通り作って，それぞれ上記の手続きで学習し，それぞれで AUC を計算した．それらの平均と標準偏差を表に示している．性能の差が統計的有意か調査するため，1 標本 t 検定を行った．有意水準を 1% とした．表では太字が最高 AUC，下線が統計的有意差が検出されなかったものを示している．顔画像では Highest で MA- κ C が最高性能を得た．MA-HC は平均 AUC は 0.044 下がったが，標準偏差が大きかったために，MA- κ C の AUC との統計的有意差は検出されなかった．Highest はクラスタ数を変えて最も高い性能をとっている．クラスタ数などのハイパーパラメータは，教師つき学習の場合，予測性能などを指標にした交差確認法で推定できる．しかし，異常検出における応用では，教師なし学習になるので，交差確認法によるクラスタ数の推定が困難になる．Average は，Highest と比べて，MA-HC は平均 AUC が 0.051 も下がったが，MA- κ C は MA- κ C は Highest と Average の差は 0.019 にとどまった．生活音でも同様な傾向があった．文字列では MA-FC が最高性能を得，ほかのどの手法とも統計的に有意な差が検出された．

MA- κ C，MS- κ C，MA-FC，MS-FC は学習時には， κ 距離やファジー距離を用いているが，表 2 に掲載されている予測性能は，予測時にはユークリッド距離を使ったものであった．これに対して，予測時にも κ 距離やファジー距離を用いる方法も考えられる．これも試したが，予測時にはユークリッド距離に戻したほうが僅差ながら性能がよいことが多かった．

カーネル法は非線形の効果を期待して使われることがしばしばあるが [18]，本実験で用いたカーネルはいずれも非線形の効果が無いものである．一方，得られているカーネル値から非線形なカーネルを得るのは容易である：任意の入力ベクトル x, x' とカーネル関数 $K(\cdot, \cdot)$ に対して，多項式カーネルならば $(K(x, x') + c)^p$ ，RBF カーネルならば $\exp(-\gamma(K(x, x) + K(x', x') - 2K(x, x')))$ のように計算できる [18]．ただし， c, p, γ は定数である．本研究でもこれらの算法を使って非線形カーネルを試したが，本実験で用いたデータでは非線形カーネルが検出性能のさらなる向上を産むことはなかった．

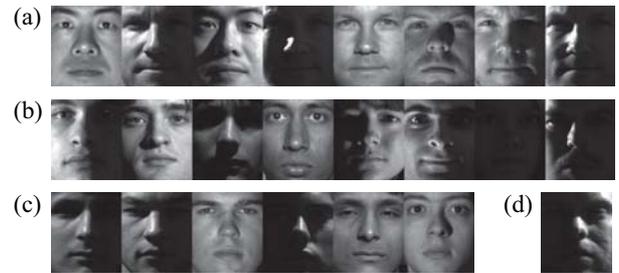


図 2 顔画像の例．(a) 正常データ．(b) 異常データ．(c) 従来法で検出できなかったが MA- κ C で検出できた異常データ．(d) いずれの従来法でも検出してしまったが MA- κ C で検出しなかった正常データ．

6. 結 論

本論文の貢献をまとめると次のようになる：

- 重みつき最小二乗推定された部分空間およびアフライン超平面とのユークリッド距離が閉形式で求まること，またカーネル化できることを示した．
- 再生核ヒルベルト空間における部分空間およびアフライン超平面を用いたファジークラスタリングを行う算法を与えた．
- 実データを使った異常検出の応用に適用して実験を行ったところ，部分空間およびアフライン超平面を用いたファジークラスタリングはクラスタ数によらず良好な検出性能を得ることが分かった．

本論文では，再生核ヒルベルト空間における線形多様体の重みつき最小二乗推定をファジークラスタリングに応用したが，共変量シフト (covariate shift) [25], [26] 下の学習においても重みつき損失平均の最小化は重要となる．このように，線形多様体の重みつき最小二乗推定を要素技術として用いることにより，機械学習およびデータ解析の可能性はさらに拡大すると考えられる．それらの技術の開発，および性能評価が今後の課題となる．

今回の実験では，風呂場の生活音から異常を検出する実験を行ったが，近い将来，高齢化社会になる我が国では，このような異常検出は不可欠な技術となるだろう．今後は，本論文で開発した技術を，ほかの生活音にも適用して，異常検出を行っていきたい．

文 献

- [1] E. Oja: "New aspects on the subspace methods of pattern recognition", *Electron. Electr. Eng. Res. Stud. Pattern Recognition and Image Processing Ser. 5*, Letchworth, UK, pp. 55-64 (1984).
- [2] K. Tsuda: "Subspace classifier in the hilbert space", *Pattern Recogn. Lett.*, **20**, pp. 513-519 (1999).
- [3] 前田英作, 村瀬洋: "カーネル非線形部分空間法によるパターン認識", *信学論*, **82-D-II**, pp. 600-612 (1999).
- [4] J. Ho, M. hsuan Yang, J. Lim, K. chih Lee and D. Kriegman: "Clustering appearances of objects under varying illumination conditions", In *CVPR*, pp. 11-18 (2003).
- [5] X. Li and K. Fukui: "Nonlinear k-subspaces based appearances clustering of objects under varying illumination conditions", *ACCV Workshop Subspace 2007*, pp. 46-52 (2007).
- [6] L. Lu and R. Vidal: "Combined central and subspace clustering for computer vision applications", *Proceedings of the 23rd international conference on machine learning, ICML '06*, New York, NY, USA, ACM, pp. 593-600 (2006).
- [7] X. Li, Z. Ning and L. Xiang: "Robust multi-body motion segmentation based on fuzzy k-subspace clustering", *IEICE - Trans. Inf. Syst.*, **E88-D**, pp. 2609-2614 (2005).
- [8] R. O. Duda, P. E. Hart and D. G. Stork: "Pattern Classification, Second Edition", John Wiley & Sons, Inc. (2001).
- [9] V. Chandola, A. Banerjee and V. Kumar: "Anomaly detection: A survey", *ACM Comput. Surv.*, **41**, pp. 15:1-15:58 (2009).
- [10] A. Patcha and J.-M. Park: "An overview of anomaly detection techniques: Existing solutions and latest technological trends", *Comput. Netw.*, **51**, pp. 3448-3470 (2007).
- [11] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz and G. Rätsch: "Kernel PCA and denoising in feature spaces", *Proceedings of the 1998 conference on Advances in neural information processing systems II*, Cambridge, MA, USA, MIT Press, pp. 536-542 (1999).
- [12] A. Abadpour and S. Kasaei: "Color PCA eigenimages and their application to compression and watermarking", *Image Vision Comput.*, **26**, pp. 878-890 (2008).
- [13] C. M. Bishop and M. E. Tipping: "A hierarchical latent variable model for data visualization", *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**, pp. 281-293 (1998).
- [14] T. Hastie, R. Tibshirani and J. H. Friedman: "The Elements of Statistical Learning", Springer (2003).
- [15] 張放, 菅谷至寛, 大町真一郎, 阿曾弘具: "部分空間法と高速フーリエ変換を用いた傾斜文字認識", *画像電子学会誌*, **39**, pp. 598-605 (2010).
- [16] S. T. Roweis and L. K. Saul: "Nonlinear dimensionality reduction by locally linear embedding", *Science*, **290**, pp. 2323-2326 (2000).
- [17] B. Schölkopf, A. Smola and K.-R. Müller: "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, **10**, 5, pp. 1299-1319 (1998).
- [18] B. Schölkopf and A. J. Smola: "Learning with kernels", MIT Press, Cambridge, MA (2002).
- [19] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins: "Text classification using string kernels", *J. Mach. Learn. Res.*, **2**, pp. 419-444 (2002).
- [20] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor and K. M. Borgwardt: "Graph kernels", *J. Mach. Learn. Res.*, **11**, pp. 1201-1242 (2010).
- [21] A. Moschitti, D. Pighin and R. Basili: "Tree kernels for semantic role labeling", *Comput. Linguist.*, **34**, pp. 193-224 (2008).
- [22] T. Gärtner: "A survey of kernels for structured data", *SIGKDD Explor. Newsl.*, **5**, pp. 49-58 (2003).
- [23] R. L. Givan, E. K. P. Chong and H. S. Chang: "Scheduling multiclass packet streams to minimize weighted loss", *Queueing Syst. Theory Appl.*, **41**, pp. 241-270 (2002).
- [24] J. Gutierrez, J.-L. Rouas and R. Andre-Obrecht: "Weighted loss functions to make risk-based language identification fused decisions", *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02, ICPR '04*, Washington, DC, USA, IEEE Computer Society, pp. 863-866 (2004).
- [25] S. Bickel, M. Brückner and T. Scheffer: "Discriminative learning under covariate shift", *J. Mach. Learn. Res.*, **10**, pp. 2137-2155 (2009).
- [26] 杉山将: "共変量シフト下での教師付き学習", *日本神経回路学会誌*, **13**, 3, pp. 111-118 (2006).
- [27] C. Alzate and J. A. K. Suykens: "Multiway spectral clustering with out-of-sample extensions through weighted kernel pca", *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, pp. 335-347 (2010).
- [28] C. Alzate and J. A. K. Suykens: "A weighted kernel pca formulation with out-of-sample extensions for spectral clustering methods", *IJCNN*, pp. 138-144 (2006).
- [29] L. Parsons, E. Haque and H. Liu: "Subspace clustering for high dimensional data: a review", *SIGKDD Explor. Newsl.*, **6**, pp. 90-105 (2004).
- [30] S. D. Villalba and P. Cunningham: "An evaluation of dimension reduction techniques for one-class classification", *Technical Report UCD-CSI-2007-9*, University College Dublin (2007).
- [31] G. R. Lanckriet, L. E. Ghaoui and M. I. Jordan: "Robust novelty detection with single-class mpm", In *advances in neural information processing systems*, MIT Press, pp. 905-912 (2003).
- [32] K. Q. Weinberger and L. K. Saul: "Fast solvers and efficient implementations for distance metric learning", *Proceedings of the 25th international conference on machine learning, ICML '08*, New York, NY, USA, ACM, pp. 1160-1167 (2008).
- [33] T. Kohonen, M. R. Schroeder and T. S. Huang Eds.: "Self-Organizing Maps", Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition (2001).
- [34] J. M. Ortega and W. C. Rheinboldt: "Iterative Solution of Nonlinear Equations in Several Variables", Academic Press, New York, NY (1970).
- [35] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia and A. G. Murzin: "SCOP database in 2004: refinements integrate structure and sequence family data", *Nuclear Acid Research*, **32**, pp. D226-D229 (2004).
- [36] M. Gribskov and N. L. Robinson: "Use of receiver operating characteristic (roc) analysis to evaluate sequence matching.", *Comput Chem*, **20**, 1, pp. 25-33 (1996).