非定常環境下での学習：共変量シフト適応，クラスバランス変化適応，変化検知

杉山　将（東京工業大学）
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/˜sugi
山田　誠（Yahoo! Labs）
makotoy@yahoo-inc.com
ドゥ・プレシ　マーティヌス・クリストフェル（東京工業大学）
christo@sg.cs.titech.ac.jp
リウ　ソン（東京工業大学）
song@sg.cs.titech.ac.jp

概 要

一般的な教師付き学習法では，訓練データとテストデータが同じ確率分布に従うという仮定のもとで学習を行う．しかし実際には，標本の選択バイアスや環境の非定常性などにより，この大前提が満たされないことがある．このような状況では，標準的な教師付き学習法は大きな推定バイアスを持ち，汎化性能が低下してしまう．本論文では，入力変数の確率分布が変化する共変量シフトと呼ばれる状況と，分類問題においてクラス事前確率が変化するクラスバランス変化と呼ばれる状況を考え，重要度重み付けによる半教師付き適応学習法を紹介する．また，確率分布が変化しているかどうかを検知する手法も紹介する．

キーワード

共変量シフト，クラスバランス変化，重要度重み付け，密度比推定，ダイバージェンス近似，変化検知

Learning under Non-Stationarity: Covariate Shift Adaptation, Class-Balance Change Adaptation, and Change Detection

Masashi Sugiyama (Tokyo Institute of Technology)
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/˜sugi

Makoto Yamada (Yahoo! Labs)
makotoy@yahoo-inc.com

Marthinus Christoffel du Plessis (Tokyo Institute of Technology)
christo@sg.cs.titech.ac.jp

Song Liu (Tokyo Institute of Technology)
song@sg.cs.titech.ac.jp

Abstract

In standard supervised learning algorithms training and test data are assumed to follow the same probability distribution. However, because of a sample selection bias or non-stationarity of the environment, this important assumption is often violated in practice, which causes a significant estimation bias. In this article, we review semi-supervised adaptation techniques for coping with such distribution changes. We focus on two scenarios of such distribution change: the covariate shift (input distributions change but the input-output dependency does not change) and the class-balance change in classification (class-prior probabilities change but class-wise input distributions remain unchanged). We also show methods of change detection in probability distributions.

1 まえがき

回帰や分類などの**教師付き学習**の目的は，入力 $\boldsymbol{x}$ と出力 y が組になった訓練データの背後に潜む入出力関係を推定することである．これにより，テスト入力点 $\boldsymbol{x}'$ に対する出力値 y' を予測できるようになる．これまでに多数の教師付き学習法が開発され，様々な応用問題を通してその有用性が実証されてきた．

標準的な教師付き学習法では，訓練データとテストデータが同じ確率分布に従うという仮定のもと，その妥当性を保証している (Vapnik, 1998; Hastie et al., 2001; Bishop, 2006)．しかし，近年の応用問題ではこの仮定が成り立たないことがあり，標準的な教師付き学習法は強い推定バイアスを持ってしまう．

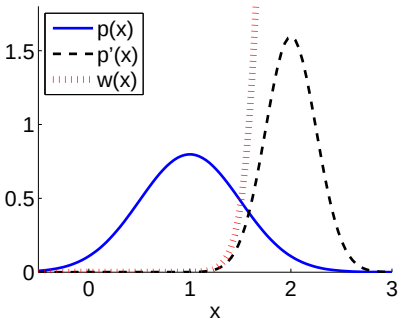
本論文では，訓練データとテストデータが異なる確率分布に従う二つの状況を考える．一つ目は，**共変量シフト** (Shimodaira, 2000; Sugiyama & Kawanabe, 2012) とよばれる状況であり，訓練データとテストデータの入力点が異なる確率分布に従うが，入出力関係そのものは変化しないという設定である．もう一つは，分類問題における**クラスバランス変化** (Saerens et al., 2002; du Plessis & Sugiyama, 2012) とよばれる状況であり，訓練データとテストデータのクラス事前確率が異なるが，各クラスの入力分布は変化しないという設定である．以下では，これらの問題設定に対して，**重要度重み付け**を用いた**半教師付き適応学習法**を紹介する．なお，本論文の第 1 節から第 3.3 節および第 5 節は，英語で出版された文献 (Sugiyama et al., 2013b) の日本語訳に相当し，図も再利用している．第 3.4 節と第 4 節は，今回新たに加筆した．

具体的には，入出力が組になった訓練データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ に加えて，入力のみのテストデータ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ もが与えられる**半教師付き学習**の問題を考える．標準な半教師付き学習の研究 (Chapelle et al., 2006) では，訓練データとテストデータは同じ確率分布に従うと仮定するが，本論文では，訓練データとテストデータの従う確率分布が異なる状況を考える．すなわち，訓練データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ は同時確率密度 $p(\boldsymbol{x}, y)$ を持つ確率分布に独立に従うと仮定し，テストデータ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ は確率密度 $\int p'(\boldsymbol{x}, y)\mathrm{d}y$ を持つ確率分布に独立に従うと仮定する．ここで，$p(\boldsymbol{x}, y)$ と $p'(\boldsymbol{x}, y)$ は一般に異なる状況を考える．

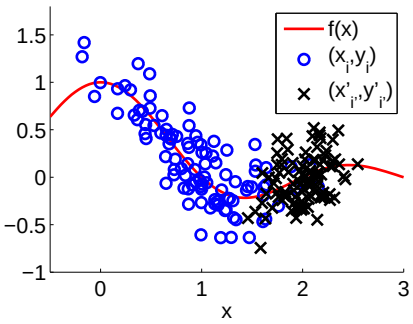
$$p(\boldsymbol{x}, y) \neq p'(\boldsymbol{x}, y)$$

本論文での学習の目的は，入出力の訓練データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ と入力のみのテストデータ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ を用いて，新たなテスト入力点 $\boldsymbol{x}'$ に対する出力値 y' を予測することである．訓練データとテストデータが異なる確率分布に従う状況下での学習は，非定常環境適応，データセットシフト適応，転移学習，ドメイン適応などとよばれる．また，半教師付きの転移学習やドメイン適応は，テスト出力データが得られないことから，教師なし転移や教師なし適応とよばれることもある．

以下，第 2 節では共変量シフトに対する適応学習法を紹介し，第 3 節ではクラスバランス変化に対する適応学習法を紹介する．そして，第 4 節で確率分布が変化したかどうかを検知する手法を紹介する．



(a) 訓練入力密度 $p(\boldsymbol{x})$，テスト入力密度 $p'(\boldsymbol{x})$，重要度 $w(\boldsymbol{x})$



(b) 学習したい真の関数 $f(\boldsymbol{x})$，訓練データ $(\boldsymbol{x}_i, y_i)$，テストデータ $(\boldsymbol{x}'_{i'}, y'_{i'})$

図 1: 共変量シフト．入力分布は変化するが学習したい真の関数は変化しない．

2 共変量シフトに対する適応学習法

共変量シフト (Shimodaira, 2000; Sugiyama & Kawanabe, 2012) とは，入力分布が変化するが，出力の入力に対する条件付き分布は変化しない状況を指す．

$$p(\boldsymbol{x}) \neq p'(\boldsymbol{x}), \quad p(y|\boldsymbol{x}) = p'(y|\boldsymbol{x})$$

図 1 に共変量シフト下での回帰問題の例を示す．この例では，訓練入力データ $\{\boldsymbol{x}_i\}_{i=1}^n$ は入力空間の左側にあり，テスト入力データ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ は入力空間の右側にある．テスト出力の予測は訓練入力データの確率密度が低い領域で行われるため，共変量シフト下での学習は**外挿**を伴う．

2.1 重要度重み付き学習

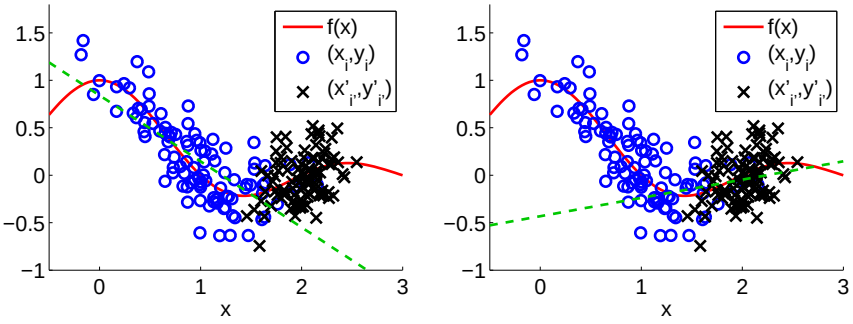
図 1(b) に示した訓練データに，単純な直線モデル

$$f_{\boldsymbol{\theta}}(x) = \theta_1 + \theta_2 x$$

を最小二乗法

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2$$

によって適合させた結果を図 2(a) に示す．破線で示した学習結果の関数は，$\circ$ 印で示した訓練データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ にはうまく適合できているが，$\times$ 印で示したテスト入力データ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ に対する出力の予測はうまくできていない．



(a) 通常の最小二乗法　　(b) 重要度重み付き最小二乗法

図 2: 共変量シフト下での回帰．破線は学習した関数を表す．

直感的には，訓練データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ のうち，テスト入力データ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ に近い入力を持つもののみを用いて学習を行うと，共変量シフトに対応できると思われる．この考え方は，$p'(\boldsymbol{x})$ と $p(\boldsymbol{x})$ の比によって定義される**重要度**

$$w(\boldsymbol{x}) := \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

に基づいて損失関数を重み付けすることにより実現できる．**重要度重み付き最小二乗法** (Shimodaira, 2000)，すなわち，

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} w(\boldsymbol{x}_i)\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\Big)^2$$

によって得られた学習結果を図 2(b) に示す．これより，重要度重み付けによってテスト出力の予測精度が向上する事がわかる．

この重要度重み付き最小二乗法では，重要度重み付けによって**汎化誤差**（期待テスト誤差）を近似している．

$$G := \iint \mathrm{loss}(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))p'(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

ここで，$\mathrm{loss}(y, \widehat{y})$ は y を $\widehat{y}$ で予測したときの損失を表す．具体的には，訓練損失の重要度重み付き平均によって，汎化誤差 G を以下のように近似できる．

$$\begin{aligned}
G &= \iint \mathrm{loss}(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))p'(y|\boldsymbol{x})p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&= \iint \mathrm{loss}(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))p'(y|\boldsymbol{x})\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&= \iint \mathrm{loss}(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))w(\boldsymbol{x})p(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&\approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{loss}(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))w(\boldsymbol{x}_i)
\end{aligned}$$

この重要度重み付けは，例えば，フィッシャー判別分析 (Sugiyama et al., 2007)，ロジスティック回帰 (Yamada et al., 2010a)，条件付き確率場 (Tsuboi et al., 2009) など，尤度や損失に基づくあらゆる学習アルゴリズムに適用できる．更に，重要度重み付けは，能動学習や実験計画におけるバイアスの低減にも役立つ (Wiens, 2000; Kanamori & Shimodaira, 2003; Sugiyama, 2006; Kanamori, 2007; Sugiyama & Rubens, 2008; Sugiyama & Nakajima, 2009)．重要度重み付けに関する更なる議論は，Sugiyama and Kawanabe (2012) を参照せよ．

重要度重み付き学習を実行するためには，重要度 $\{w(\boldsymbol{x}_i)\}_{i=1}^n$ の値が必要である．しかし，訓練入力密度 $p(\boldsymbol{x})$ とテスト入力密度 $p'(\boldsymbol{x})$ は一般に未知であるため，重要度の値はデータから推定する必要がある．訓練入力密度 $p(\boldsymbol{x})$ を訓練入力データ $\{\boldsymbol{x}_i\}_{i=1}^n$ から，テスト入力密度 $p'(\boldsymbol{x})$ をテスト入力データ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ からそれぞれ推定し，それらの比を取れば密度比を推定できる．しかし，比を取ることにより確率密度の推定誤差が増幅される恐れがあるため，このような二段階の重要度推定法は一般に精度が良くない．

そこで，$p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ を推定することなく，それらの比 $w(\boldsymbol{x})$ を直接推定する様々な方法が開発された．

- $p(\boldsymbol{x})$ が一様分布になるように入力空間を変換した後，$p'(\boldsymbol{x})$ の推定を行う方法 (Ćwik & Mielniczuk, 1989; Chen et al., 2009)．
- $p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ から生成されたデータを，ロジスティック回帰によって分離する方法 (Qin, 1998; Cheng & Chu, 2004; Bickel et al., 2007)．
- 重要度のモデル $w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ を考え，$p(\boldsymbol{x})w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ の積率を適合する方法 (Qin, 1998; Gretton et al., 2009; Kanamori et al., 2012)．
- $p(\boldsymbol{x})w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ に関する積分方程式を解く方法 (Vapnik et al., 2013; Que & Belkin, 2013)．
- $p(\boldsymbol{x})w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ をカルバック・ライブラー距離 (Kullback & Leibler, 1951) のもとで適合する方法 (Sugiyama et al., 2008; Nguyen et al., 2010; Tsuboi et al., 2009; Yamada & Sugiyama, 2009; Yamada et al., 2010b)．
- $w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ と $w(\boldsymbol{x})$ を最小二乗適合する方法 (Kanamori et al., 2009; Kanamori et al., 2012)．
- $w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ と $w(\boldsymbol{x})$ をブレグマン距離 (Bregman, 1967) のもとで適合する方法 (Sugiyama et al., 2012b)．

これらの方法の中で特に，重要度モデルを最小二乗適合する方法 (Kanamori et al., 2009; Kanamori et al., 2012) が次のような優れた性質を持っており，実用上非常に有用である．

- 最適解が解析的かつ効率的に計算できる．

- ハイパーパラメータの調整に交差確認が利用できる．
- パラメトリックとノンパラメトリック両方の設定で，最適な学習率が達成できる (Kanamori et al., 2009; Kanamori et al., 2012)．
- あるクラスの重要度推定量の中で，条件数の意味で最適な安定性を持っている (Kanamori et al., 2013)．

更に，重要度推定を次元削減と組み合わせることにより，高次元空間における重要度推定の精度を向上させられる (Sugiyama et al., 2010; Sugiyama et al., 2011; Yamada & Sugiyama, 2011)．重要度推定に関する更なる議論は，Sugiyama et al. (2012a) を参照せよ．

2.2 相対重要度重み付き学習

図 1 と図 2 に示した共変量シフト回帰の例では，$x = 2$ あたりにある少数の訓練データのみが大きな重要度を持ち，それ以外の訓練データの重要度重みはほとんどゼロである．従って，重要度重み付き学習は，事実上 $x = 2$ あたりにあるわずかな訓練データのみを用いて行われていることになり，不安定である．

このような不安定性は，重要度関数 $w(\boldsymbol{x})$ が極端に大きな値を取ることに原因があり，**相対重要度**を用いることにより解決できる (Yamada et al., 2013)．

$$w^{(\beta)}(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{\beta p'(\boldsymbol{x}) + (1-\beta)p(\boldsymbol{x})}$$

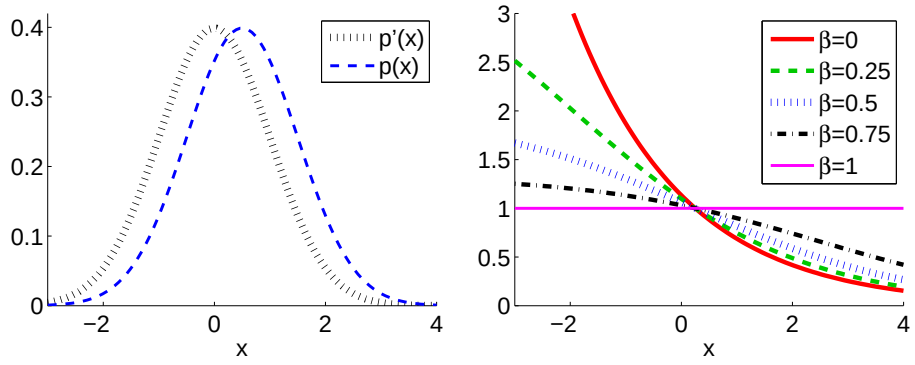
ここで，$\beta \in [0, 1]$ は相対化係数とよばれ，相対重要度の滑らかさを調整する．相対重要度 $w^{(\beta)}(\boldsymbol{x})$ は，$\beta = 0$ のときにもとの重要度 $w(\boldsymbol{x})$ と一致し，β を増加させると段々と平坦になっていく．そして，$\beta = 1$ のときに一様重み $w(\boldsymbol{x}) = 1$ になる（図 3 参照）．このことは，重要度関数の非負性 $p'(\boldsymbol{x})/p(\boldsymbol{x}) \geq 0$ より，相対重要度関数が常に $1/\beta$ 以下になることからも確認できる．

$$w^{(\beta)}(\boldsymbol{x}) = \frac{1}{\beta + (1-\beta)\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}} \leq \frac{1}{\beta}$$

相対重要度で重み付けした最小二乗法を，**相対重要度重み付き最小二乗法**とよぶ．

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} w^{(\beta)}(\boldsymbol{x}_i) \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2$$

相対重要度重み付き最小二乗法では，相対化係数 β はバイアスとバリアンスのトレードオフを調整する役割を担う．



(a) 確率密度　　(b) 相対重要度 $w^{(\beta)}(x)$

図 3: 相対重要度．$p'(x)$ は標準正規分布であり，$p(x)$ はそれを 0.5 だけシフトさせたものである．

以下，二つのデータ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ から，相対重要度 $w^{(\beta)}(\boldsymbol{x})$ を直接推定する方法を紹介する．相対重要度 $w^{(\beta)}(\boldsymbol{x})$ の近似に，パラメータに関する線形モデル

$$w_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{b} \alpha_j \psi_j(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\psi}(\boldsymbol{x})$$

を用いる．ここで，

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top$$

はパラメータのベクトルを表し，

$$\boldsymbol{\psi}(\boldsymbol{x}) = (\psi_1(\boldsymbol{x}), \ldots, \psi_b(\boldsymbol{x}))^\top$$

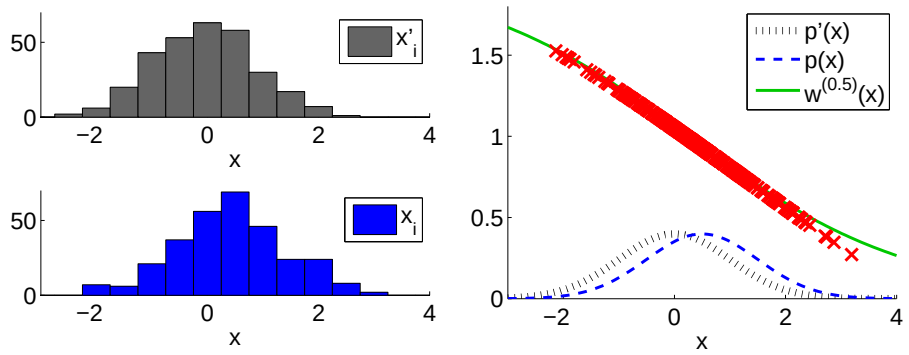
は基底関数のベクトルを表す．実際には，ガウスカーネルモデル

$$w_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{n'} \alpha_j \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'_j\|^2}{2\sigma^2}\right)$$

がよく用いられる．σ^2 はガウス関数のバンド幅を表す．

相対重要度モデル $w_{\boldsymbol{\alpha}}(\boldsymbol{x})$ のパラメータ $\boldsymbol{\alpha}$ は，以下の二乗誤差規準 $J(\boldsymbol{\alpha})$ を最小にするように学習する．

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \int \Big(w_{\boldsymbol{\alpha}}(\boldsymbol{x}) - w^{(\beta)}(\boldsymbol{x})\Big)^2 \Big(\beta p'(\boldsymbol{x}) + (1-\beta)p(\boldsymbol{x})\Big) \mathrm{d}\boldsymbol{x} \\ &= \int \boldsymbol{\alpha}^\top \boldsymbol{\psi}(\boldsymbol{x}) \boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\alpha} \Big(\beta p'(\boldsymbol{x}) + (1-\beta)p(\boldsymbol{x})\Big) \mathrm{d}\boldsymbol{x} \\ &\quad - 2 \int \boldsymbol{\alpha}^\top \boldsymbol{\psi}(\boldsymbol{x}) p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + C \end{aligned}$$



(a) 訓練データとテストデータ　　(b) 相対重要度 ($\beta = 0.5$)

図 4: 最小二乗相対重要度適合法の実行例．図 4(b) の × 印は，推定した相対重要度の $\boldsymbol{x}_i$ での値を表す．

ここで，第三項目の

$$C = \int w^{(\beta)}(\boldsymbol{x}) p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

は定数であるため，無視することにする．第一項目と第二項目の期待値を標本平均で近似し，ℓ_2 正則化項を追加すれば，以下の学習規準が得られる．

$$\min_{\boldsymbol{\alpha}} \left[\boldsymbol{\alpha}^\top \widehat{\boldsymbol{G}}_\beta \boldsymbol{\alpha} - 2\widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \|\boldsymbol{\alpha}\|^2 \right]$$

ここで，$\lambda \geq 0$ は正則化の強さを調整する正則化係数であり，$\widehat{\boldsymbol{G}}_\beta$ と $\widehat{\boldsymbol{h}}$ はそれぞれ次式で定義される $b \times b$ 行列と b 次元ベクトルである．

$$\widehat{\boldsymbol{G}}_\beta = \frac{\beta}{n'} \sum_{i'=1}^{n'} \boldsymbol{\psi}(\boldsymbol{x}'_{i'}) \boldsymbol{\psi}(\boldsymbol{x}'_{i'})^\top + \frac{1-\beta}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{x}_i) \boldsymbol{\psi}(\boldsymbol{x}_i)^\top, \quad \widehat{\boldsymbol{h}} = \frac{1}{n'} \sum_{i'=1}^{n'} \boldsymbol{\psi}(\boldsymbol{x}'_{i'})$$

この学習規準はパラメータ $\boldsymbol{\alpha}$ に関して凸二次関数であるため，最小解 $\widehat{\boldsymbol{\alpha}}$ は解析的に

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{G}}_\beta + \lambda \boldsymbol{I} \right)^{-1} \widehat{\boldsymbol{h}}$$

によって求められる．この方法を，**最小二乗相対重要度適合法**とよぶ (Yamada et al., 2013)．ハイパーパラメータである正則化係数 λ とガウスカーネルのバンド幅 σ^2 は，学習規準 J に関する交差確認によって最適化することができる．最小二乗相対重要度適合法の実行例を図 4 に示す．

2.3　重要度重み付きモデル選択

教師付き学習において高い汎化性能を獲得するためには，相対化係数 β や学習モデルの基底関数などのモデル選択が非常に重要である．一般に教師付き学習のモデル選択には，

赤池情報量規準 (Akaike, 1974)，**部分空間情報量規準** (Sugiyama & Ogawa, 2001)，**交差確認** (Stone, 1974) などが用いられるが，共変量シフト下ではこれらのモデル選択の妥当性が保証されない．

共変量シフト下では，これらのモデル選択法を重要度で重み付けすることにより，理論的な妥当性が保証される (Shimodaira, 2000; Sugiyama & Müller, 2005; Sugiyama et al., 2007)．その中で特に，**重要度重み付き交差確認法**が実用上有用である．以下に，重要度重み付き交差確認法の手続きを示す．

1. 訓練データ $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ を，m 個の重なりのない（ほぼ）同じ大きさの部分集合 $\{\mathcal{T}_i\}_{i=1}^m$ にランダムに分割する．

2. $i = 1, \ldots, m$ に対して以下を繰り返す．

   (a) $\mathcal{T} \backslash \mathcal{T}_i$（すなわち，$\mathcal{T}_i$ 以外の全てのデータ）を用いて学習結果の関数 f_i を求める．

   (b) 取っておいたデータ $\mathcal{T}_i$ に対する汎化誤差を評価する．

$$\widehat{G}_i = \begin{cases} \dfrac{1}{|\mathcal{T}_i|} \displaystyle\sum_{(\boldsymbol{x},y)\in\mathcal{T}_i} w(\boldsymbol{x}) \Big(f_i(\boldsymbol{x}) - y \Big)^2 & \text{（回帰）} \\ \dfrac{1}{|\mathcal{T}_i|} \displaystyle\sum_{(\boldsymbol{x},y)\in\mathcal{T}_i} \dfrac{w(\boldsymbol{x})}{2} \Big(1 - \mathrm{sign}\big(f_i(\boldsymbol{x}) y \big) \Big) & \text{（分類）} \end{cases}$$

   ここで，$|\mathcal{T}_i|$ は集合 $\mathcal{T}_i$ の要素数を表す．

3. $\{\widehat{G}_i\}_{i=1}^m$ の平均値 $\widehat{G}$ を最終的な汎化誤差の予測値として出力する．

$$\widehat{G} = \frac{1}{m} \sum_{i=1}^m \widehat{G}_i$$

2.4 応用例

共変量シフト適応学習は，ブレインコンピュータインターフェース (Sugiyama et al., 2007; Li et al., 2010)，ロボット制御 (Hachiya et al., 2009; Akiyama et al., 2010; Hachiya et al., 2011; Zhao et al., 2013)，音声からの話者認識 (Yamada et al., 2010a)，顔画像からの年齢予測 (Ueki et al., 2011)，加速度センサーからの行動識別 (Hachiya et al., 2012)，日本語の単語分割 (Tsuboi et al., 2009)，迷惑メールフィルタ (Bickel & Scheffer, 2007)，ターゲット広告配信 (Bickel et al., 2009)，HIV 検査 (Bickel et al., 2008)，半導体露光装置におけるウエハ位置合わせ (Sugiyama & Nakajima, 2009) など，様々な実応用問題においてその有用性が実証されている．以下，動画からの三次元姿勢推定に対する応用例を紹介する (Yamada et al., 2012)．

三人の被験者が，歩く，ジョギングする，ボクシング動作をする，投げる，受ける，ジェスチャをするという動作をしたときの，多視点動画とモーションキャプチャデータからな

る HumanEva-I データセット (Sigal & Black, 2006) を用いる．入力 $\boldsymbol{x}$ として，三つの視点の動画から得た各 9630 フレームの画像に対して，270 次元の**勾配方向ヒストグラム特徴** (Bo & Sminchisescu, 2010) を抽出する．訓練データとして，$3 \times 4815 = 14445$ フレームからランダムに n 個のデータを選び，残りの 14445 フレームをテストデータとして用いる．

以下の二つの設定で実験を行う．

被験者選択バイアス： 訓練データとして三人の被験者全てのデータを用い，テストデータとして一人の被験者のみデータを用いる．

被験者間転移： 訓練データとして二人の被験者のデータを用い，テストデータとして残りの一人の被験者のデータを用いる．

学習法として，**カーネル回帰法** (KR) (Agarwal & Triggs, 2005)，**対ガウス過程回帰法** (TGP) (Bo & Sminchisescu, 2010)，**重み付き k 最近傍法** (WkNN) (Shakhnarovich et al., 2003) を用いる．これらの学習法の詳細は，Yamada et al. (2012) を参照せよ．以下では，KR，TGP の重要度重み付き版を IWKR，IWTGP とよぶ．

各姿勢は，20 個の三次元関節マーカ

$$\boldsymbol{y} = [\boldsymbol{y}^{(1)\top}, \ldots, \boldsymbol{y}^{(20)\top}]^\top \in \mathbb{R}^{60}$$

によって表される．真の姿勢 $\boldsymbol{y}^*$ とその推定 $\widehat{\boldsymbol{y}}$ の誤差を

$$\mathrm{Error}(\boldsymbol{y}^*, \widehat{\boldsymbol{y}}) = \frac{1}{20} \sum_{m=1}^{20} \|\widehat{\boldsymbol{y}}^{(m)} - \boldsymbol{y}^{*(m)}\|$$

で評価する．

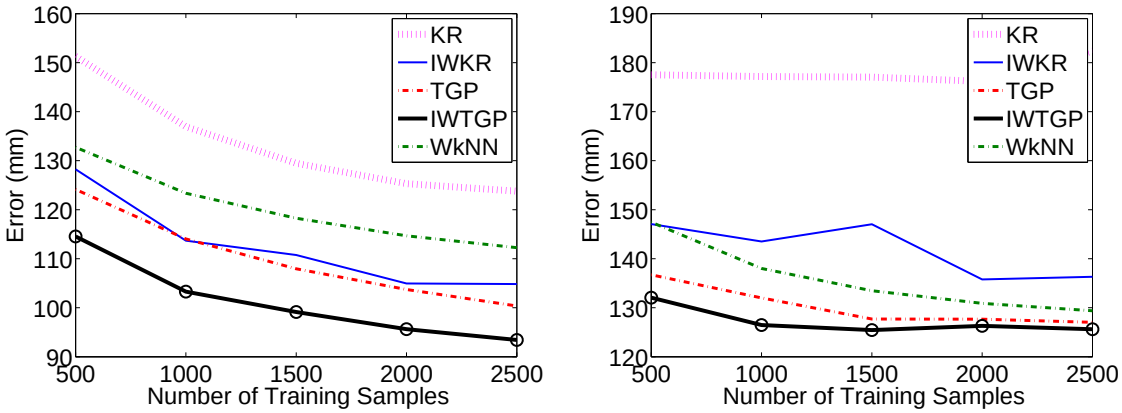
各動作 10 回の試行に対する平均姿勢推定誤差を，訓練データ数 n の関数として図 5 に示す．これらのグラフより，重要度重み付き学習法である IWTGP と IWKR は，重要度で重み付けをしていない TGP と KR，および，WkNN より性能が良いことがわかる．

3 クラスバランス変化に対する適応学習法

分類問題における**クラスバランス変化** (Saerens et al., 2002; du Plessis & Sugiyama, 2012) とは，クラス事前確率が変化するが各クラスの入力分布は変化しない状況を指す．

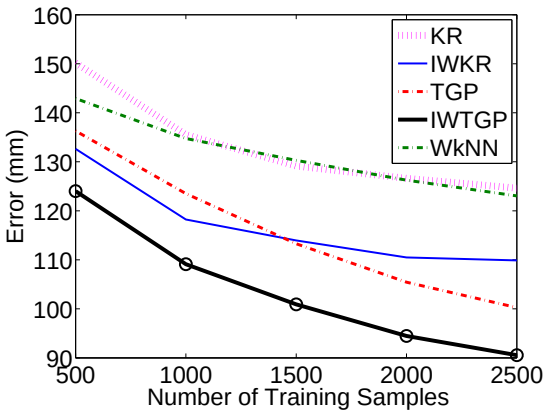
$$p(y) \neq p'(y), \quad p(\boldsymbol{x}|y) = p'(\boldsymbol{x}|y) \tag{1}$$

クラスバランス変化の例を図 6 に示す．訓練データとテストデータのクラス事前確率が異なるとき，訓練データを用いて単純に分類器を学習すると推定バイアスを持ってしまう．

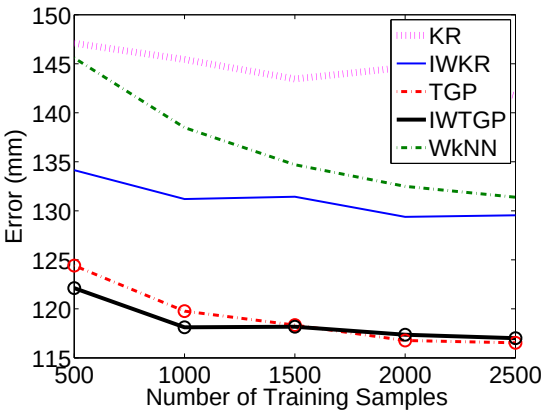


(a) 被験者選択バイアス S1

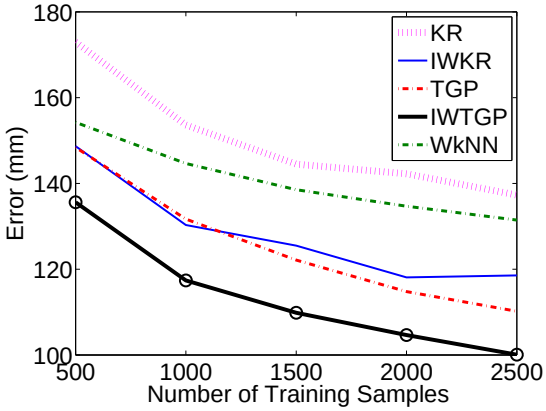
(b) 被験者間転移 S1



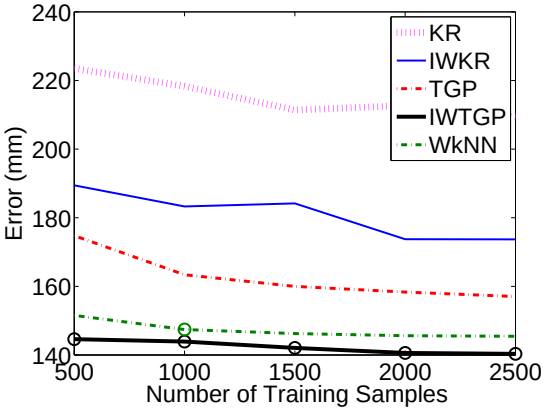
(c) 被験者選択バイアス S2



(d) 被験者間転移 S2

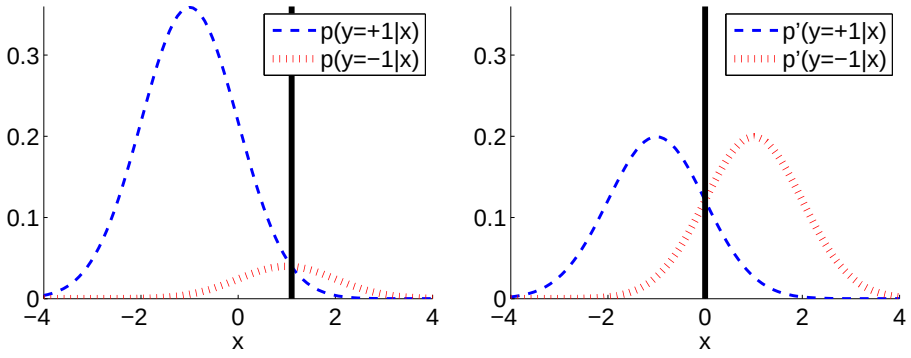


(e) 被験者選択バイアス S3



(f) 被験者間転移 S3

図 5: 三次元の姿勢推定実験．各動作 10 回の試行に対する平均姿勢推定誤差をプロット．平均誤差が最小の手法と，有意水準が 5% の t 検定によって同等だと判定された手法の結果に ○ 印を付した．



(a) 訓練データ:$p(y=+1)=0.9$, $p(y=-1)=0.1$

(b) テストデータ : $p'(y=+1)=p'(y=-1)=0.5$

図 6: クラスバランス変化．分類問題において，クラス事前確率が $p(y)$ から $p'(y)$ に変化するが，各クラスの入力分布は変化しない，すなわち，$p(\boldsymbol{x}|y)=p'(\boldsymbol{x}|y)$．クラスバランス変化によって，最適な分類境界が変化する．

クラスバランス変化による推定バイアスは，訓練データに対する損失を**クラス事前確率比**によって重み付けすることにより解消することができる．

$$w(y)=\frac{p'(y)}{p(y)}$$

以下では，クラスラベル y が $+1$ か -1 の値のみを取る二値分類問題を考える．

3.1 クラス事前確率推定

訓練データのクラス事前確率 $p(y)$ は，訓練データのクラス比 n_y/n によって簡単に推定できる．ここで，n_y は n 個の訓練データ $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$ のうちクラス y に属するものの数を表す．テストデータのクラス事前確率 $p'(y)$ も，ラベル付きテストデータ $\{(\boldsymbol{x}'_{i'},y'_{i'})\}_{i'=1}^{n'}$ が与えられれば同様に推定できる．しかし，ここでは半教師付き学習の設定を考えているため，テストデータは入力データ $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ のみしか与えらない．従って，$p'(y)$ は単純には推定できない．

式 (1) のもとでの半教師付き学習では，訓練データのクラス毎の入力密度 $p(\boldsymbol{x}|y)$ の混合密度 $q_\pi(\boldsymbol{x})$ をテスト入力密度 $p'(\boldsymbol{x})$ に適合させることにより，テストデータのクラス事前確率 $p'(y)$ を推定できる（図 7 参照）．

$$q_\pi(\boldsymbol{x})=\pi p(\boldsymbol{x}|y=+1)+(1-\pi)p(\boldsymbol{x}|y=-1)$$

ここで，パラメータ π の値は $p'(y=+1)$ に対応し，$1-\pi$ は $p'(y=-1)$ に対応する．

q_π と p' との適合には，例えば，**カルバック・ライブラー距離** (Kullback & Leibler, 1951)

$$\mathrm{KL}(p'\|q_\pi)=\int p'(\boldsymbol{x})\log\frac{p'(\boldsymbol{x})}{q_\pi(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$$

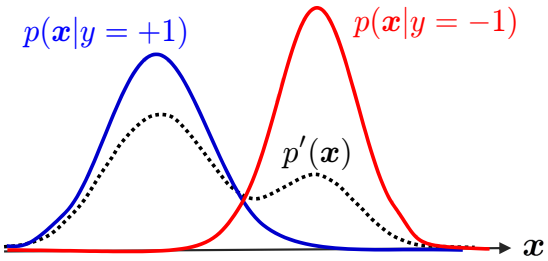


図 7: クラス毎の訓練入力密度 $p(\boldsymbol{x}|y)$ の混合密度 $q_\pi(\boldsymbol{x})$ をテスト入力密度 $p'(\boldsymbol{x})$ に適合させることによって，テストデータのクラス事前確率 $p'(y)$ を推定できる．

やピアソン距離 (Pearson, 1900)

$$\mathrm{PE}(p'\|q_\pi) = \int q_\pi(\boldsymbol{x}) \left(\frac{p'(\boldsymbol{x})}{q_\pi(\boldsymbol{x})} - 1\right)^2 \mathrm{d}\boldsymbol{x}$$

を用いる．これらの距離は，$p'(\boldsymbol{x})$ と $q_\pi(\boldsymbol{x})$ を推定せずに密度比 $p'(\boldsymbol{x})/q_\pi(\boldsymbol{x})$ を直接推定することにより，精度良く近似できる (Sugiyama et al., 2012a)．しかし，密度比関数 $p'(\boldsymbol{x})/q_\pi(\boldsymbol{x})$ は小さな変動に敏感なため，外れ値の影響を受けやすいという問題がある．

そこで以下では，外れ値の影響を受けにくい L^2 距離を用いることにする．

$$L^2(p', q_\pi) = \int \Big(p'(\boldsymbol{x}) - q_\pi(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}$$

L^2 距離は，密度差 $p'(\boldsymbol{x}) - q_\pi(\boldsymbol{x})$ を直接推定することにより，精度良く近似できる (Sugiyama et al., 2013a)．

L^2 距離のもとでの混合比 π のノンパラメトリック推定は，古くは Hall (1981) によって議論されている．その後，カーネル密度推定に基づく手法 (Titterington, 1983) が提案され，更にカーネルのバンド幅の同時選択法が開発された (Hall & Wand, 1988)．L^2 距離のもとでのカーネルのバンド幅の選択に関する更なる議論は，Anderson et al. (1994) を参照せよ．

3.2 L^2 距離の近似

ここでは，密度差の直接推定による L^2 距離の近似法 (Kim & Scott, 2010; Sugiyama et al., 2013a) を紹介する．説明を簡単にするために，確率密度 p と p' との間の L^2 距離

$$L^2(p, p') = \int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}, \qquad f(\boldsymbol{x}) = p(\boldsymbol{x}) - p'(\boldsymbol{x}) \tag{2}$$

を，二つのデータ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ から近似する問題を考える．

密度差 $f(\boldsymbol{x})$ の近似に，ガウスカーネルモデル

$$f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{n+n'} \alpha_j \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_j\|^2}{2\sigma^2}\right)$$

を用いる．ここで，

$$(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n, \boldsymbol{c}_{n+1}, \ldots, \boldsymbol{c}_{n+n'}) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'})$$

はガウスカーネルの中心を表す．密度差モデルのパラメータ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n+n'})^\top$ は，以下の二乗誤差規準 $J(\boldsymbol{\alpha})$ を最小にするように学習する．

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \int \Big(f_{\boldsymbol{\alpha}}(\boldsymbol{x}) - f(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x} \\ &= \int f_{\boldsymbol{\alpha}}(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - 2\int f_{\boldsymbol{\alpha}}(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + C \end{aligned}$$

ここで，第三項目の

$$C = \int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}$$

は定数であるため無視することにする．第一項目は，解析的に計算できる．

$$\int f_{\boldsymbol{\alpha}}(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} = \boldsymbol{\alpha}^\top \boldsymbol{U} \boldsymbol{\alpha}$$

ただし，$\boldsymbol{U}$ は第 (j, j') 要素が

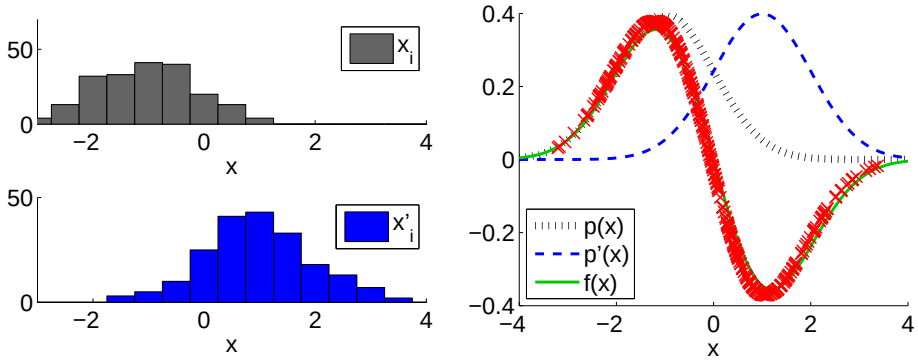
$$\begin{aligned} U_{j,j'} &= \int \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_j\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_{j'}\|^2}{2\sigma^2}\right) \mathrm{d}\boldsymbol{x} \\ &= (\pi\sigma^2)^{d/2} \exp\left(-\frac{\|\boldsymbol{c}_j - \boldsymbol{c}_{j'}\|^2}{4\sigma^2}\right) \end{aligned}$$

で与えられる $(n+n') \times (n+n')$ 行列である．第二項目 $\int f_{\boldsymbol{\alpha}}(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ に含まれる期待値を標本平均で近似し，ℓ_2 正則化項を追加すれば，以下の学習規準が得られる．

$$\min_{\boldsymbol{\alpha}} \left[\boldsymbol{\alpha}^\top \boldsymbol{U} \boldsymbol{\alpha} - 2\widehat{\boldsymbol{v}}^\top \boldsymbol{\alpha} + \lambda \|\boldsymbol{\alpha}\|^2\right]$$

ここで，$\lambda \geq 0$ は正則化の強さを調整する正則化係数であり，$\widehat{\boldsymbol{v}}$ は第 j 要素が

$$\widehat{v}_j = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_j\|^2}{2\sigma^2}\right) - \frac{1}{n'} \sum_{i'=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x}'_{i'} - \boldsymbol{c}_j\|^2}{2\sigma^2}\right)$$



(a) データ　　(b) 密度差

図 8: 最小二乗密度差推定法の実行例．図 8(b) の × 印は，推定した密度差の $\boldsymbol{x}_i$ と $\boldsymbol{x}'_{i'}$ での値を表す．

で与えられる $(n+n')$ 次元ベクトルである．この学習規準はパラメータ $\boldsymbol{\alpha}$ に関して凸二次関数であるため，最小解 $\widehat{\boldsymbol{\alpha}}$ は解析的に

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{U} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{v}}$$

によって求められる．この方法を，**最小二乗密度差推定法**とよぶ (Sugiyama et al., 2013a)．ハイパーパラメータである正則化係数 λ とガウスカーネルのバンド幅 σ^2 は，学習規準 J に関する交差確認によって最適化することができる．最小二乗密度差推定法の実行例を図 8 に示す．

L^2 距離 (2) に含まれる密度差関数 f を最小二乗密度差推定量 $f_{\widehat{\boldsymbol{\alpha}}}$ で置き換えれば，L^2 距離推定量 $\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \widehat{\boldsymbol{\alpha}}$ が得られる．同様に，L^2 距離の別表現

$$L^2(p, p') = \int f(\boldsymbol{x}) \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x}) \Big) \mathrm{d}\boldsymbol{x}$$

より，別の L^2 距離推定量 $\widehat{\boldsymbol{v}}^\top \widehat{\boldsymbol{\alpha}}$ が得られる．これらの線形結合

$$2\widehat{\boldsymbol{v}}^\top \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \widehat{\boldsymbol{\alpha}}$$

は，推定バイアスが小さいことが知られており，実用上有用である (Sugiyama et al., 2013a)．

3.3 数値例

UCI **データセット** [1] を用いた数値例を示す．各データセットに対して，正負それぞれのクラスからラベル付きの訓練データを 10 個ずつ選び，真のクラス事前確率

$$\pi^* = 0.1, 0.2, \ldots, 0.9$$

[1] `http://archive.ics.uci.edu/ml/`

に従って，ラベルなしのテストデータを 50 個選ぶ．最小二乗密度差推定法の性能を，以下の方法と比較する．

KDEi：ガウスカーネル関数に対するカーネル密度推定法を用いて $p'(\boldsymbol{x})$ と $q_\pi(\boldsymbol{x})$ をデータからそれぞれ推定し，推定した確率密度関数間の L^2 距離を計算する (Titterington, 1983)．それぞれのガウスカーネル関数のバンド幅は，確率密度の二乗誤差に関する 5 分割交差確認によって**独立に**最適化する (Härdle et al., 2004)．

KDEj：同様にカーネル密度推定法を用いるが，それぞれのガウスカーネル関数のバンド幅を，確率密度差の二乗誤差に関する 5 分割交差確認によって**同時に**最適化する (Hall & Wand, 1988).

EM：EM アルゴリズムを用いたクラス事前確率推定法 (Saerens et al., 2002)．この手法はカルバック・ライブラー距離のもとでの分布適合に対応する．

真のクラス事前確率 π とその推定値との二乗誤差の 1000 回の試行に対する平均値と標準誤差を，図 9 の左側の列に示す．これより，最小二乗密度差推定法によってクラス事前確率を精度良く推定できている事がわかる．

次に，推定したクラス事前確率を用いて分類器を学習する．ここでは，クラス事前確率比で重み付けした ℓ_2 正則化最小二乗分類器 (Rifkin et al., 2003) を用いる．すなわち，テスト入力 $\boldsymbol{x}'$ に対するクラスラベルの予測 $\widehat{y}'$ を

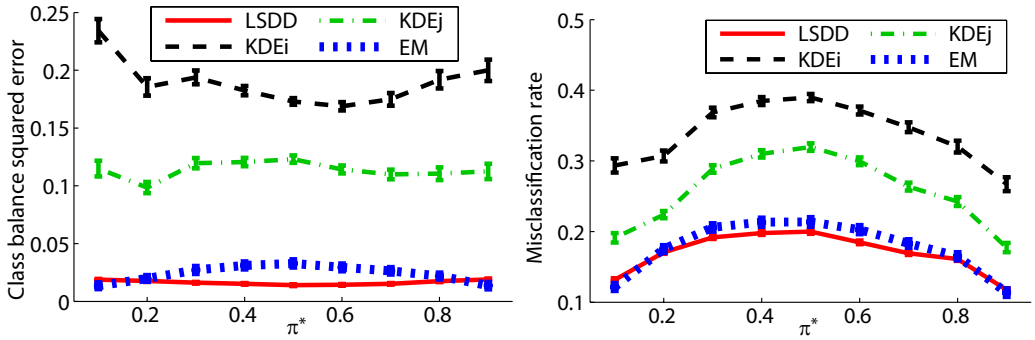
$$\widehat{y}' = \mathrm{sign}\left(\sum_{j=1}^{n} \widehat{\theta}_j K(\boldsymbol{x}', \boldsymbol{x}_j)\right)$$

によって求める．ここで $K(\boldsymbol{x}, \boldsymbol{x}')$ はバンド幅 κ のガウスカーネルを表し，パラメータ $\{\widehat{\theta}_j\}_{j=1}^n$ は

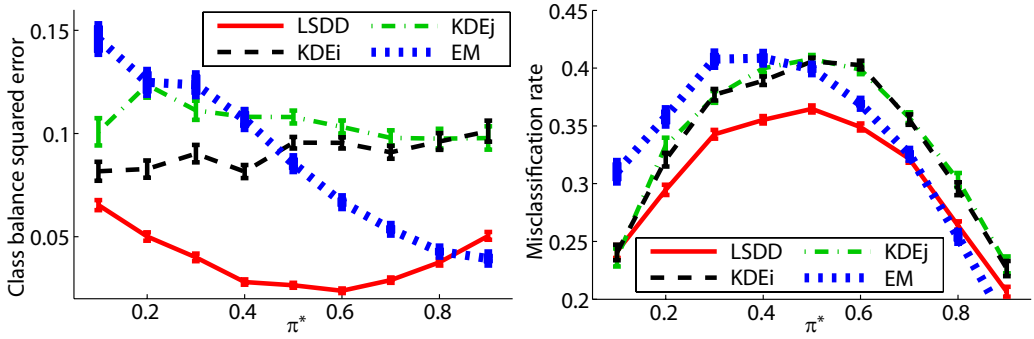
$$(\widehat{\theta}_1, \ldots, \widehat{\theta}_n) := \mathop{\mathrm{argmin}}_{\theta_1,\ldots,\theta_n} \left[\sum_{i=1}^{n} \frac{\pi_{y_i}}{n_{y_i}/n}\left(\sum_{j=1}^{n} \theta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - y_i\right)^2 + \delta \sum_{j=1}^{n} \theta_j^2\right]$$

によって求められる．また，$\pi_{+1} = \widehat{\pi}$，$\pi_{-1} = 1 - \widehat{\pi}$，$\widehat{\pi}$ はクラス事前確率の推定値を表し，$\delta \geq 0$ は正則化係数を表す．ガウスカーネルのバンド幅 κ と正則化係数 δ は，誤分類率に関する 5 分割クラス事前確率比重み付き交差確認法 (Sugiyama et al., 2007) により決定する．

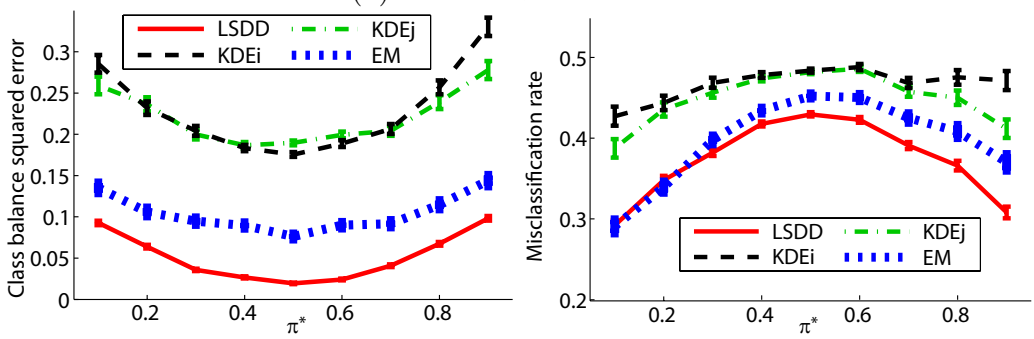
1000 回の試行に対する誤分類率の平均と標準誤差を，図 9 の右側の列に示す．これより，最小二乗密度差推定法によって誤分類率を低く抑えられていることがわかる．これは，テストデータのクラス事前確率がうまく推定できていることによると考えられる．



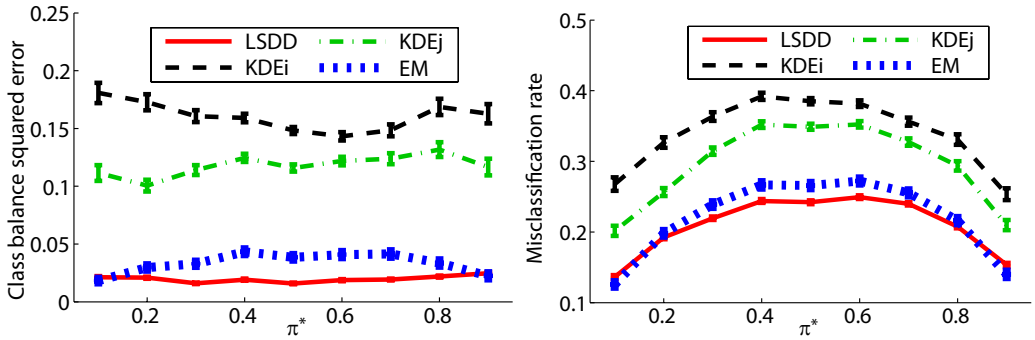
(a) Australian データセット



(b) Diabetes データセット

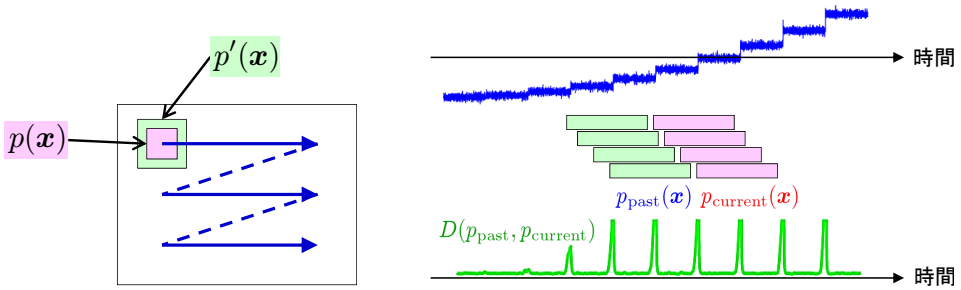


(c) German データセット



(d) Statlogheart データセット

図 9: クラスバランス適応の数値例．左：クラス事前確率の二乗推定誤差．右：クラス事前確率比重み付き ℓ_2 正則化最小二乗分類器による誤分類率．ハイパーパラメータはクラス事前確率比重み付き交差確認により決定．



(a) 画像中の注目領域検出　　(b) 時系列データの変化点検知

図 10: 確率分布間の距離に基づく変化検知の応用例.

3.4　正例のみからのクラス事前確率推定

ここまで，正負両方のクラスのラベル付き訓練データを用いたクラス事前確率推定を紹介してきた．しかし，1 クラス分類問題 (Li et al., 2011) や異常検出問題 (Hido et al., 2011) などの状況では，正のクラスのラベル付き訓練データしか与えられないため，上記の手法をそのまま用いることができない．

このような場合でも，テスト入力密度 $p'(\boldsymbol{x})$ に正のラベル付き訓練データの確率密度の定数倍 $\gamma p(\boldsymbol{x}|y=+1)$ を適合させることにより，テストデータのクラス事前確率を推定できる (Elkan & Noto, 2008; du Plessis & Sugiyama, 2014)．この推定法によって求めた γ は，正負のクラスの訓練データの確率密度 $p(\boldsymbol{x}|y=\pm 1)$ が重なりを持たないとき，すなわち，$p(\boldsymbol{x}|y=+1)p(\boldsymbol{x}|y=-1)=0$ のとき，正のテストデータのクラス事前確率 $p'(y=+1)$ の不偏推定量になっている．

4　教師なし変化検知

本節では，確率分布の変化を検知する手法を紹介する．教師なし変化検知の目的は，確率密度 $p(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}_i\}_{i=1}^{n}$ と確率密度 $p'(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ が与えられたとき，$p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ が等しいかどうかを判定することである．

4.1　確率分布間の距離に基づく変化検知

第 3.1 で，密度比推定を用いたカルバック・ライブラー距離，ピアソン距離，L^2 距離の推定法を紹介した．これらの手法を用いて p と p' の距離を推定すれば，p と p' が同じかどうかを判定することができる (Liu et al., 2013)．

図 10 に示すような考え方で確率分布間の距離を推定することにより，**画像中の注目領域の検出** (Yamanaka et al., 2013b) や**時系列データ** (Kawahara & Sugiyama, 2012; Yamanaka et al., 2013a) の変化点の検知を行うことができる．

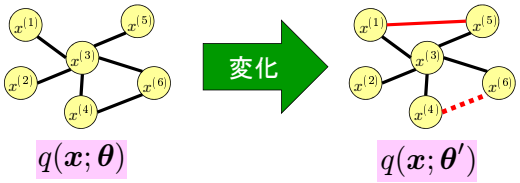


図 11: 二次のマルコフネットワークは，$k = 1, \ldots, d$ が頂点で $\|\boldsymbol{\theta}_{k,k'}\| > 0$ のときに頂点 k, k' 間に枝があるグラフとして表現できる．マルコフネットワークの変化は，グラフの枝の有無の変化に対応する．

4.2 確率分布の構造の変化検知

上記の手法では，確率分布間の距離をノンパラメトリックに推定することにより，柔軟かつロバストな変化検知を行うことができた．一方，$p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ に二次のマルコフネットワーク

$$q(\boldsymbol{x};\boldsymbol{\theta}) \propto \exp\left(\sum_{k\geq k'} \boldsymbol{\theta}_{k,k'}^{\top} \boldsymbol{f}(x^{(k)}, x^{(k')})\right)$$

を仮定することにより，このモデルのパラメータ $\boldsymbol{\theta}$ の値の変化から，多次元ベクトル $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^{\top}$ の要素 $x^{(k)}, x^{(k')}$ 間の相互作用の変化を捉える事ができる（図 11）．

$\boldsymbol{f}(x, x')$ は，マルコフネットワークの特徴ベクトルである．$\boldsymbol{f}(x, x') = xx'$ とおけば，二次のマルコフネットワークは正規モデルと一致し，$x^{(k)}, x^{(k')}$ 間の共分散をモデル化することに対応する．$\boldsymbol{f}(x, x') = [x^t, x^{t-1}x', \ldots, x, x', 1]^{\top}$ のような高次特徴を考えると，$x^{(k)}, x^{(k')}$ 間の高次の共分散を考慮したモデル化が行える．

モデル $q(\boldsymbol{x};\boldsymbol{\theta})$ を標本 $\{\boldsymbol{x}_i\}_{i=1}^{n}$ と $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ にそれぞれ適合させ，推定したパラメータ $\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}'$ の差 $\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}'$ を求めることにより，変数間の相互作用の変化を捉えられる．例えば，**グループスパース正則化** (Yuan & Lin, 2007; Friedman et al., 2008) を施した最尤推定法

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\log q(\boldsymbol{x}_i;\boldsymbol{\theta}) - \lambda \sum_{k\geq k'} \|\boldsymbol{\theta}_{k,k'}\|\right], \quad \lambda \geq 0$$

$$\max_{\boldsymbol{\theta}'} \sum_{i'=1}^{n'} \left[\log q(\boldsymbol{x}'_{i'};\boldsymbol{\theta}') - \lambda' \sum_{k\geq k'} \|\boldsymbol{\theta}'_{k,k'}\|\right], \quad \lambda' \geq 0$$

を用いれば，スパースなパラメータ変化を捉えることができる．しかし，我々が知りたいのはパラメータが変化したかどうかだけであり，個々のパラメータの値 $\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}'$ を求めることが目的ではない．

そこで，二つのパラメータの差 $\boldsymbol{\theta} - \boldsymbol{\theta}'$ を直接スパース化する手法が提案された (Tibshirani

et al., 2005; Zhang & Wang, 2010)．

$$\max_{\boldsymbol{\theta},\boldsymbol{\theta}'}\left[\sum_{i=1}^{n}\log q(\boldsymbol{x}_i;\boldsymbol{\theta})+\sum_{i'=1}^{n}\log q(\boldsymbol{x}'_{i'};\boldsymbol{\theta}')-\gamma\sum_{k\geq k'}\|\boldsymbol{\theta}_{k,k'}-\boldsymbol{\theta}'_{k,k'}\|\right],\quad \gamma\geq 0$$

しかし，この方法でもまだ二つのパラメータ $\boldsymbol{\theta},\boldsymbol{\theta}'$ を明示的に推定している．また，マルコフネットワークの正規化項

$$\int\exp\left(\sum_{k\geq k'}\boldsymbol{\theta}_{k,k'}^{\top}\boldsymbol{f}(x^{(k)},x^{(k')})\right)\mathrm{d}\boldsymbol{\theta}$$

は，正規モデルやノンパラ正規モデル (Liu et al., 2009) など，限られたモデルに対してしか効率良く計算することができないという問題もある．

これらの問題点は，パラメータ $\boldsymbol{\theta},\boldsymbol{\theta}'$ を明示的に推定せず，それらの差 $\boldsymbol{\alpha}=\boldsymbol{\theta}-\boldsymbol{\theta}'$ を直接推定することにより解決することができる (Liu et al., 2014)．パラメータの差 $\boldsymbol{\alpha}$ の推定は，対数線形密度比モデルの学習に対応する．

$$r(\boldsymbol{x};\boldsymbol{\alpha})=\frac{q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta}')}\propto\exp\left(\sum_{k\geq k'}\boldsymbol{\alpha}_{k,k'}^{\top}\boldsymbol{f}(x^{(k)},x^{(k')})\right)$$

対数線形密度比モデルは，密度比をカルバック・ライブラー距離のもとで適合する方法 (Tsuboi et al., 2009; Liu et al., 2014) によって効率良く学習できる．

$$\min_{\boldsymbol{\alpha}}\int p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})r(\boldsymbol{x};\boldsymbol{\alpha})}\mathrm{d}\boldsymbol{x}+\gamma\sum_{k\geq k'}\|\boldsymbol{\alpha}'_{k,k'}\|$$

5 まとめ

本論文では，共変量シフトとクラスバランス変化に対する半教師付き適応学習法，及び，変化検知のを紹介した．これらの手法の MATLAB による実装は，

`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/`

にて無償公開されている．より一般的な状況における適応学習法に関しては，Quiñonero-Candela et al. (2009) を参照せよ．

訓練入出力データ $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$ に加えてテスト入出力データ $\{(\boldsymbol{x}'_{i'},y'_{i'})\}_{i'=1}^{n'}$ も与えられれば，同時重要度 $p'(\boldsymbol{x},y)/p(\boldsymbol{x},y)$ を用いた重み付き学習により，$p(\boldsymbol{x},y)$ と $p'(\boldsymbol{x},y)$ が任意に異なる場合に対しても原理的には対応できる．このような状況では，訓練データからテストデータへの適応だけでなく，テストデータから訓練データへの適応も同時に行うことができる．これは**マルチタスク学習** (Caruana, 1997) の考え方であり，近年，機械学習分野において盛んに研究されている．

教師付き学習は，統計学や機械学習の分野において非常に盛んに研究されてきた．しかし実際には，訓練入出力データを収集するのにお金や時間がかかることがあるため，入力のみのテストデータも活用する半教師付き学習や，類似の学習タスクを利用する転移学習やマルチタスク学習など，訓練入出力データ以外の追加情報を用いる学習法の重要性が高まっている．更に，**クラウドソーシング** (Raykar et al., 2010) や**自己教示学習** (Raina et al., 2007) など，訓練入出力データを安価に収集するための新しい枠組みも開発されつつある．このような研究は，到来しつつある**ビッグデータ**時代の最重要課題の一つであり，今後の更なる発展が期待される．

ビッグデータ時代には，データサイエンティストは多種多様なデータに対して，様々なデータ解析を迅速に行うことを要請される．その際，データ解析タスクごとにアルゴリズムの開発・実装を行うのは効率が悪いため，様々なデータ解析タスクを統一的に解決するようなアプローチが，今後益々重要になってくると考えられる．

謝辞

杉山は科研費 25700022 と AOARD の助成を受けた．山田は科学技術振興機構さきがけ研究および東京工業大学 PLIP プログラムの助成を受けた．ドゥ・プレシは文部科学省奨学金を受けた．リウは日本学術振興会特別研究員の助成を受けた．

References

Agarwal, A., & Triggs, B. (2005). Monocular human motion capture with a mixture of regressors. *Proceedings of IEEE Workshop on Vision for Human Computer Interaction at Computer Vision and Pattern Recognition* (p. 72).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*, 716–723.

Akiyama, T., Hachiya, H., & Sugiyama, M. (2010). Efficient exploration through active learning for value function approximation in reinforcement learning. *Neural Networks*, *23*, 639–648.

Anderson, N., Hall, P., & Titterington, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*, 41–54.

Bickel, S., Bogojeska, J., Lengauer, T., & Scheffer, T. (2008). Multi-task learning for HIV therapy screening. *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)* (pp. 56–63).

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning (ICML2007)* (pp. 81–88).

Bickel, S., Sawade, C., & Scheffer, T. (2009). Transfer learning by distribution matching for targeted advertising. *Advances in Neural Information Processing Systems 21* (pp. 145–152).

Bickel, S., & Scheffer, T. (2007). Dirichlet-enhanced spam filtering based on biased samples. *Advances in Neural Information Processing Systems 19* (pp. 161–168). Cambridge, MA: MIT Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY, USA: Springer.

Bo, L., & Sminchisescu, C. (2010). Twin Gaussian processes for structured prediction. *International Journal of Computer Vision*, *87*, 28–52.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*, 200–217.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. Cambridge, MA, USA: MIT Press.

Chen, S.-M., Hsu, Y.-S., & Liaw, J.-T. (2009). On kernel estimators of density ratio. *Statistics*, *43*, 463–479.

Cheng, K. F., & Chu, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, *10*, 583–604.

Ćwik, J., & Mielniczuk, J. (1989). Estimating density ratio with application to discriminant analysis. *Communications in Statistics: Theory and Methods*, *18*, 3057–3069.

du Plessis, M. C., & Sugiyama, M. (2012). Semi-supervised learning of class balance under class-prior change by distribution matching. *Proceedings of 29th International Conference on Machine Learning (ICML2012)* (pp. 823–830). Edinburgh, Scotland.

du Plessis, M. C., & Sugiyama, M. (2014). Class prior estimation from positive and unlabeled data. *IEICE Transactions on Information and Systems*, *E97-D*.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 213–220).

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning* (pp. 131–160). Cambridge, MA, USA: MIT Press.

Hachiya, H., Akiyama, T., Sugiyama, M., & Peters, J. (2009). Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Networks*, *22*, 1399–1410.

Hachiya, H., Peters, J., & Sugiyama, M. (2011). Reward weighted regression with sample reuse. *Neural Computation*, *11*, 2798–2832.

Hachiya, H., Sugiyama, M., & Ueda, N. (2012). Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing*, *80*, 93–101.

Hall, P. (1981). On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B*, *43*, 147–156.

Hall, P., & Wand, M. P. (1988). On nonparametric discrimination using density differences. *Biometrika*, *75*, 541–547.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin, Germany: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY, USA: Springer.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, *26*, 309–336.

Kanamori, T. (2007). Pool-based active learning with optimal sampling distribution and its information geometrical interpretation. *Neurocomputing*, *71*, 353–362.

Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*, 1391–1445.

Kanamori, T., & Shimodaira, H. (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, *116*, 149–162.

Kanamori, T., Suzuki, T., & Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, *86*, 335–367.

Kanamori, T., Suzuki, T., & Sugiyama, M. (2013). Computational complexity of kernel-based density-ratio estimation: A condition number analysis. *Machine Learning*, *90*, 431–460.

Kawahara, Y., & Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, *5*, 114–127.

Kim, J., & Scott, C. (2010). L_2 kernel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 1822–1831.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Li, W., Guo, Q., & Elkan, C. (2011). A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, *49*, 717–725.

Li, Y., Kambara, H., Koike, Y., & Sugiyama, M. (2010). Application of covariate shift adaptation techniques in brain computer interfaces. *IEEE Transactions on Biomedical Engineering*, *57*, 1318–1324.

Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, *10*, 2295–2328.

Liu, S., Quinn, J., Gutmann, M. U., & Sugiyama, M. (2014). Direct learning of sparse changes in Markov networks by density ratio estimation. *Neural Computation*.

Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, *43*, 72–83.

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, *56*, 5847–5861.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*, 157–175.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*, 619–630.

Que, Q., & Belkin, M. (2013). *Inverse density as an inverse problem: The fredholm equation approach* (Technical Report 1304.5575). arXiv.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. (Eds.). (2009). *Dataset shift in machine learning*. Cambridge, Massachusetts, USA: MIT Press.

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. (2007). Self-taught learning: Transfer learning from unlabeled data. *Proceedings of the 24th Annual International Conference on Machine Learning (ICML2007)* (pp. 759–766). Corvallis, OR: Omnipress.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, *11*, 1297–1322.

Rifkin, R., Yeo, G., & Poggio, T. (2003). Regularized least-squares classification. In J. A. K. Suykens, G. Horvath, S. Basu, C. Micchelli and J. Vandewalle (Eds.), *Advances in learning theory: Methods, models and applications*, vol. 190 of *NATO Science Series III: Computer & Systems Sciences*, 131–154. Amsterdam, the Netherlands: IOS Press.

Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, *14*, 21–41.

Shakhnarovich, G., Viola, P., & Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. *Proceedings of International Conference on Computer Vision (ICCV2003)* (pp. 750–757).

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*, 227–244.

Sigal, L., & Black, M. J. (2006). *Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion* (Technical Report CS-06-08). Brown University.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, *36*, 111–147.

Sugiyama, M. (2006). Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, *7*, 141–166.

Sugiyama, M., & Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. Cambridge, Massachusetts, USA: MIT Press.

Sugiyama, M., Kawanabe, M., & Chui, P. L. (2010). Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, *23*, 44–59.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*, 985–1005.

Sugiyama, M., & Müller, K.-R. (2005). Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, *23*, 249–279.

Sugiyama, M., & Nakajima, S. (2009). Pool-based active learning in approximate linear regression. *Machine Learning*, *75*, 249–274.

Sugiyama, M., & Ogawa, H. (2001). Subspace information criterion for model selection. *Neural Computation*, *13*, 1863–1889.

Sugiyama, M., & Rubens, N. (2008). A batch ensemble approach to active learning with model selection. *Neural Networks*, *21*, 1278–1286.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012a). *Density ratio estimation in machine learning*. Cambridge, UK: Cambridge University Press.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012b). Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, *64*, 1009–1044.

Sugiyama, M., Suzuki, T., Kanamori, T., du Plessis, M. C., Liu, S., & Takeuchi, I. (2013a). Density-difference estimation. *Neural Computation*, *25*, 2734–2775.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.

Sugiyama, M., Yamada, M., & du Plessis, M. C. (2013b). Learning under non-stationarity: Covariate shift and class-balance change. *WIREs Computational Statistics*, *5*, 465–477.

Sugiyama, M., Yamada, M., von Bünau, P., Suzuki, T., Kanamori, T., & Kawanabe, M. (2011). Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, *24*, 183–198.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society series B*, *67*, 91–108.

Titterington, D. M. (1983). Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B*, *45*, 37–46.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., & Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, *17*, 138–155.

Ueki, K., Sugiyama, M., & Ihara, Y. (2011). Lighting condition adaptation for perceived age estimation. *IEICE Transactions on Information and Systems*, *E94-D*, 392–395.

Vapnik, V. N. (1998). *Statistical learning theory.* New York, NY, USA: Wiley.

Vapnik, V. N., Braga, I., & Izmailov, R. (2013). *Constructive setting of the density ratio estimation problem and its rigorous solution* (Technical Report 1306.0407). arXiv.

Wiens, D. P. (2000). Robust weights and designs for biased regression models: Least squares and generalized M-estimation. *Journal of Statistical Planning and Inference*, *83*, 395–412.

Yamada, M., Sigal, L., & Raptis, M. (2012). No bias left behind: Covariate shift adaptation for discriminative 3D pose estimation. *Proceedings of European Conference on Computer Vision (ECCV2012)* (pp. 674–687).

Yamada, M., & Sugiyama, M. (2009). Direct importance estimation with Gaussian mixture models. *IEICE Transactions on Information and Systems*, *E92-D*, 2159–2162.

Yamada, M., & Sugiyama, M. (2011). Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI2011)* (pp. 549–554). San Francisco, California, USA: The AAAI Press.

Yamada, M., Sugiyama, M., & Matsui, T. (2010a). Semi-supervised speaker identification under covariate shift. *Signal Processing*, *90*, 2353–2361.

Yamada, M., Sugiyama, M., Wichern, G., & Simm, J. (2010b). Direct importance estimation with a mixture of probabilistic principal component analyzers. *IEICE Transactions on Information and Systems*, *E93-D*, 2846–2849.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., & Sugiyama, M. (2013). Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, *25*, 1324–1370.

Yamanaka, M., Matsugu, M., & Sugiyama, M. (2013a). Detection of activities and events without explicit categorization. *IPSJ Transactions on Mathematical Modeling and Its Applications*, *6*, 86–92.

Yamanaka, M., Matsugu, M., & Sugiyama, M. (2013b). Salient object detection based on direct density-ratio estimation. *IPSJ Transactions on Mathematical Modeling and Its Applications*, *6*, 78–85.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, *68*, 49–67.

Zhang, B., & Wang, Y. (2010). Learning structural changes of Gaussian graphical models in controlled experiments. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010)* (pp. 701–708).

Zhao, T., Hachiya, H., Tangkaratt, V., Morimoto, J., & Sugiyama, M. (2013). Efficient sample reuse in policy gradients with parameter-based exploration. *Neural Computation*, *25*, 1512–1547.