ブートストラップに適した尤度比の保守的な直接推定法

菊地　真人† 　吉田　光男† 　梅村　恭司†

† 豊橋技術科学大学情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1–1
E-mail: †m143313@edu.tut.ac.jp, ††yoshida@cs.tut.ac.jp, †††umemura@tut.jp

あらまし　自然言語処理におけるブートストラップ法は，それぞれの繰り返しにおいて確実な分類結果を次の学習に使用することが求められる．ここで，尤度比に基づいて分類結果を得るとすると，尤度比の保守的な推定量を使用する方法が考えられる．本稿では，尤度比の直接推定法で導入される正則化項の扱いで保守的な推定量が得られることを指摘し，保守的な推定量とブートストラップ法を組み合わせることで分類性能が向上したことを報告する．

キーワード　ブートストラップ法，尤度比，uLSIF，保守的な推定

1. はじめに

ブートストラップ法は半教師有り学習でよく使用される方法の一つである．この方法は，少数のラベル付きデータをシードと呼ばれる初期の教師データとして与え，他のラベルなしデータを分類する．そして，その分類結果をも利用してラベルなしデータの分類を繰り返す．ブートストラップ法は，ウェブページの分類 [1]，語義曖昧性解消 [2], [3]，固有表現抽出・分類 [4], [5]，構文解析 [6]，機械翻訳 [7] や情報抽出 [8] など様々なタスクに応用されている．ブートストラップ法では，分類されたラベルなしデータを次回の学習に使用する観点から，次回の学習に有用な分類結果を多く得ること求められる．

尤度比は，二値分類によく用いられる統計的尺度である．近年，二値分類における順位付け関数として，尤度比の最適性が理論的に証明された [9]．そこで本稿では，尤度比を分類スコアとして利用するブートストラップ法について考える．尤度比は通常，母集団から得られた標本に基づいて推定されるが，標本サイズが小さい場合に推定誤差が大きくなりやすい問題を抱えている．特に，最尤推定量をブートストラップ法の分類スコアとして利用すると，高スコアの分類結果に誤ったデータが混入する原因となる．混入したデータは，現時点での分類性能の低下を招くのみならず，以降の繰り返しにおいても悪影響を与える．推定誤差の問題について，理解のために極端な例を示して説明する．全体の標本，目的の標本が従う確率密度をそれぞれ $p_{de}(x)$，$p_{nu}(x)$ とする．それぞれの確率密度から得た標本をもとに尤度比 $r(x) = p_{nu}(x)/p_{de}(x)$ を推定する．$p_{de}(x)$ が $p_{nu}(x)$ と等しい状況を仮定するとき，図 1 に示すように理想の尤度比 $r(x)$ は標本空間の全体で 1 となる．だが，実際の推定において，$p_{de}(x)$ の低い所は推定誤差が大きいと考えられる．なぜなら，このような所では，得られる全体の標本サイズが小さいと想像できるためである．我々は，全体の標本サイズが小さい場合に正確な尤度比の推定が難しいことに着目し，尤度比を低めに見積もることを提案する．本稿では，推定値を低めに見積もることを保守的な推定と定義する．

保守的な推定を実現する一方法として，尤度比の直接推定法を利用することを考える．直接推定法は連続的な標本空間から

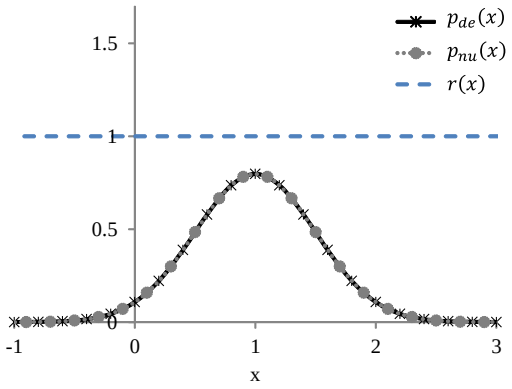


図 1: 確率密度と尤度比の例

尤度比を推定する手法であり，標本空間の構造を活用するため，基底関数を使用する．しかし，自然言語処理などでは離散的な標本空間を扱う場合が多い．そこで，空間毎に独立な基底関数を使用することで直接推定法を離散的なケースに応用する．この基底関数は，標本に含まれる異なる要素間の関係性を捉えられず，無意味なように思われるが，この方法で尤度比を推定すると，直接推定法における目的関数内部の正則化項を使用して推定値をスムージングする方式となることが分かった．この正則化項の扱いによって標本サイズの大きさを考慮した，尤度比の保守的な推定値が得られる．

提案する尤度比の保守的な推定法は，標本サイズが小さい場合に，推定値を保守的に見積もる．一方，ブートストラップ法は，分類結果を次回の学習に利用するため，確実な分類を行う必要がある．このことから，ブートストラップ法と尤度比の保守的な推定法は相性が良いと考えられる．提案手法の有効性を確認するため，ブートストラップ方式で科学ニュース記事から科学雑誌名を抽出する実験をした．雑誌名は多種多様であり，外国語の雑誌名を和名表記する際は記事の著者によって表記ゆれが生じる．この理由から，雑誌名を網羅する教師データを事前に用意することは困難であり，ブートストラップ法を適用する．また，雑誌名は特定の文脈に出現しやすい傾向にあり，雑誌名の特定に分布仮説 [10] が効果的である．それゆえ，雑誌名

の文脈パターンから雑誌名らしさのスコアを尤度比で測定できる．結果として，尤度比の保守的な推定法とブートストラップ法を組み合わせ，雑誌名抽出の性能が向上したことを報告する．

本稿の主な貢献は以下にまとめられる．(1) 尤度比の直接推定法を離散のケースに応用し，直接推定法における目的関数内部の正則化項の扱いによって保守的な推定量が得られることを示した．(2) 提案した保守的な推定法とブートストラップ法との相性の良さを実験的に示した．

2. 関 連 研 究

離散的な標本空間からの尤度比推定について，関連研究を述べる．母集団から十分な標本が観測できる場合は，最尤推定によってそれぞれの確率を求め，それらの比を取ることが単純な尤度比の推定法である．しかし，標本サイズが小さい場合，推定値にかかるバイアスが大きくなりやすく，最尤推定による方法は信頼できない．この場合の合理的な推定法として，大別して 3 つの方法を挙げる．

一つ目は，尤度比の信頼区間を構築し，その信頼区間を基に信頼できる標本サイズを求める．そして，それ以上の標本サイズが得られた場合のみ，尤度比を推定する方法である．しかし，この方法では標本サイズが小さい場合に尤度比を推定できないことが問題となる．

二つ目は，推定値にかかるバイアスを抑制する方法である．オッズ比は尤度比と同様，二つの確率から推定することができる．そのため，オッズ比推定のためのバイアス抑制法がしばしば尤度比推定にも応用できる．ここでは，次の 2 手法を挙げる．文献 [11] では，それぞれの確率推定に中位不偏推定量（MUE: Median Unbiased Estimator）を用い，それらの確率をもとにオッズ比を推定する．文献 [12] では，それぞれの確率推定に経験ベイズ法を適用し，それらの確率をもとにオッズ比を推定する．これらの方法は共通して，オッズ比を構成する確率推定を工夫するアプローチのため，最終的な推定対象をオッズ比から尤度比に変更するのみで応用可能である．

三つ目は，標本サイズの大きさに応じて推定値を低く（保守的に）見積もる方法である．自己相互情報量（PMI）の推定において，対数内部の尤度比計算に信頼区間を使用する方法がある [13]．具体的には，尤度比の分母を構成する確率推定に信頼区間の上限，分子を構成する確率推定に信頼区間の下限を採用する．この方法によって推定された PMI は，保守的な推定値となり，提案手法と似た性質を示す．この手法は二つの信頼区間を使用するため，調整すべきパラメータとして二つの信頼係数を持つ．条件付き確率の推定においては，保守的な推定法がいくつか提案されている．条件付き確率は，二つの確率の比で表現できるため，尤度比の特殊な場合と考えることもできる．文献 [14] では，信頼区間の下限を使用することで，条件付き確率の保守的な推定を実現する．文献 [15] では，条件付き確率を最尤推定し，推定値の分母に小さな正の定数を加算する．後者の手法は，本稿の提案手法と極めて似た推定法であるが，推定対象が条件付き確率に限定されており，導出の過程も異なる．提案手法は，条件付き確率を含む尤度比全般の推定に応用可能なため，文献 [15] の手法を尤度比の推定へと一般化したものと考えることもできる．

3. 尤度比の直接推定法

データの定義域を $D(\subset \mathbb{R}^d)$ で表す．$\mathbb{R}^d$ は実 d-次元空間である．いま，確率密度 $p_{\mathrm{de}}(x)$ を持つ確率分布に独立に従う i.i.d. 標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$，および $p_{\mathrm{nu}}(x)$ を持つ確率分布に独立に従う i.i.d. 標本 $\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ が与えられたとする．ただし，$p_{\mathrm{de}}(x)$ は

$$p_{\mathrm{de}}(x) > 0 \text{ for all } x \in D$$

を満たすと仮定する．二組の標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ と $\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ から次の尤度比を直接推定する問題を扱う．

$$r(x) = \frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}$$

尤度比の直接推定法としては，ロジスティック回帰を用いた方法 [16]，カルバック・ライブラー情報量を用いた方法 [17]，最小二乗法を用いた方法 [18] などがある．この節では，最小二乗法を用いた方法である LSIF とその近似である uLSIF について説明する．

3.1 LSIF

LSIF (Least-Squares Importance Fitting) [18] では，推定する尤度比 $r(x)$ を次の線形モデルで表現する．

$$\hat{r}(x) = \sum_{l=1}^{b} \alpha_l \varphi_l(x) \tag{1}$$

$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_b)^{\mathrm{T}}$ は標本から学習されるべきパラメータ，$\{\varphi_l(x)\}_{l=1}^{b}$ は非負の値を取る基底関数である．b および $\{\varphi_l(x)\}_{l=1}^{b}$ は標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$，$\{x_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ と独立である．

モデル $\hat{r}(x)$ のパラメータ $\boldsymbol{\alpha}$ は次の二乗誤差を最小にするように決定される．

$$J_0(\boldsymbol{\alpha}) = \frac{1}{2}\int (\hat{r}(x) - r(x))^2 p_{de}(x) dx$$

二乗誤差 $J_0(\boldsymbol{\alpha})$ は標本 $\{x_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ の確率のもとで期待値として定義される．$J_0(\boldsymbol{\alpha})$ の展開式を J とする．このとき，$\hat{r}(x)$ を含まない項は定数項であるため，無視できる．

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \frac{1}{2}\int \hat{r}(x)^2 p_{de}(x) dx - \int \hat{r}(x) p_{nu}(x) dx \\ &= \frac{1}{2}\sum_{l,l'=1}^{b} \alpha_l \alpha_{l'} \left(\int \varphi_l(x)\varphi_{l'}(x) dx\right) \\ &\quad - \sum_{l=1}^{b} \alpha_l \left(\int \varphi_l(x) p_{nu}(x) dx\right) \\ &= \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{h}^{\mathrm{T}} \boldsymbol{\alpha} \end{aligned} \tag{2}$$

ここで，$\boldsymbol{H}$ は (l, l') 番目の要素 $H_{l,l'}$ を持つサイズ $b \times b$ の行列である．要素 $H_{l,l'}$ は次のように定義される．

$$H_{l,l'} = \int \varphi_l(x)\varphi_{l'}(x) p_{de}(x) dx \tag{3}$$

$\boldsymbol{h}$ は l 番目の要素 h_l を持つ b 次元ベクトルである．要素 h_l は

次のように定義される．

$$h_l = \int \varphi_l(x) p_{nu}(x) dx \tag{4}$$

J における期待値の近似は経験的期待値（標本を用いて連続値 x を離散化すること）によって求められる．

$$\begin{aligned}\hat{J}(\boldsymbol{\alpha}) &= \frac{1}{2n_{de}} \sum_{i=1}^{n_{de}} \hat{r}(x_i^{\mathrm{de}})^2 - \frac{1}{n_{nu}} \sum_{j=1}^{n_{nu}} \hat{r}(x_j^{\mathrm{nu}}) \\ &= \frac{1}{2} \sum_{l,l'=1}^{b} \alpha_l \alpha_{l'} \left(\frac{1}{n_{de}} \sum_{i=1}^{n_{de}} \varphi_l(x_i^{\mathrm{de}}) \varphi_{l'}(x_i^{\mathrm{de}})\right) \\ &\quad - \sum_{l=1}^{b} \alpha_l \left(\frac{1}{n_{nu}} \sum_{j=1}^{n_{nu}} \varphi_l(x_j^{\mathrm{nu}})\right) \\ &= \frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \hat{\boldsymbol{H}} \boldsymbol{\alpha} - \hat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{\alpha}\end{aligned}$$

ここで，$\hat{\boldsymbol{H}}$ は (l, l') 番目の要素 $\hat{H}_{l,l'}$ を持つサイズ $b \times b$ の行列である．要素 $\hat{H}_{l,l'}$ は次のように定義される．

$$\hat{H}_{l,l'} = \frac{1}{n_{de}} \sum_{i=1}^{n_{de}} \varphi_l(x_i^{de}) \varphi_{l'}(x_i^{de}) \tag{5}$$

$\hat{\boldsymbol{h}}$ は l 番目の要素 $\hat{h}_l$ を持つ b 次元ベクトルである．要素 $\hat{h}_l$ は次のように定義される．

$$\hat{h}_l = \frac{1}{n_{nu}} \sum_{j=1}^{n_{nu}} \varphi_l(x_j^{nu}) \tag{6}$$

尤度比 $r(x)$ の非負性を考慮すると，次の最適化問題が得られる．

$$\begin{aligned}&\min_{\boldsymbol{\alpha} \in \mathbb{R}^b} [\frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \hat{\boldsymbol{H}} \boldsymbol{\alpha} - \hat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{\alpha} + \lambda \mathbf{1}_b^{\mathrm{T}} \boldsymbol{\alpha}] \\ &\text{subject to } \boldsymbol{\alpha} > \mathbf{0}_b\end{aligned} \tag{7}$$

$\mathbf{0}_b$ および $\mathbf{1}_b$ は要素がすべて 0，すべて 1 の b 次元ベクトルである．ベクトルの不等式 $\boldsymbol{\alpha} > \mathbf{0}_b$ は要素ごとに適用される．すなわち，$\alpha_l > 0$ for $l = 1, 2, ..., b$ となる．上式では $\boldsymbol{\alpha}$ に対する正則化のため，ペナルティ項 $\lambda \mathbf{1}_b^{\mathrm{T}} \boldsymbol{\alpha}$ を導入する．λ $(\geqq 0)$ は正則化パラメータ，$\mathbf{1}_b^{\mathrm{T}} \boldsymbol{\alpha}$ は l_1-正則化項である．上式は凸二次計画問題となり，標準的な最適化パッケージを用いれば大域的な最適解を一意に求めることができる．さらに，LSIF の原著論文では効率的に尤度比推定を行う正則化パス追跡アルゴリズムも提案されている．

3.2 uLSIF

LSIF は正則化パス追跡アルゴリズムを用いることで，効率的に尤度比を推定できる．しかし，このアルゴリズムは数値計算上，しばしば不安定な挙動をする．そこで，LSIF の近似解を計算する，計算効率が良く安定的な代替手法として uLSIF (unconstrained Least-Squares Importance Fitting) が提案されている [18].

uLSIF のアイディアはとても単純である．式 (7) で与えられる最適化問題について，パラメータの非負制約をなくすことである．これにより，次の拘束無し最適化問題が得られる．

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^b} [\frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{H}} \boldsymbol{\beta} - \hat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{\beta} + \frac{\lambda}{2} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}] \tag{8}$$

上式では，線形な正則化項 $\mathbf{1}_b^{\mathrm{T}} \boldsymbol{\beta}$ の代わりに二次正規化項 $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}/2$ を導入している．式 (8) は拘束無し凸二次計画問題であり，その解は次式で解析的に計算できる．

$$\tilde{\boldsymbol{\beta}}(\lambda) = (\hat{\boldsymbol{H}} + \lambda \mathbf{1}_b)^{-1} \hat{\boldsymbol{h}}$$

$\mathbf{1}_b$ は要素がすべて 1 の b 次元ベクトルである．パラメータの非負制約を取り除いたのでいくつかのパラメータは負の値となることが考えられる．この近似誤差を補うため，解を次のように修正する．

$$\hat{\boldsymbol{\beta}}(\lambda) = \max(\mathbf{0}_b, \tilde{\boldsymbol{\beta}}(\lambda))$$

上式の 'max' 操作はベクトルの要素毎に適用される．これが uLSIF の解となる．

この手法においても，尤度比の直接推定に標本空間の構造を利用する．そのため，基底関数を使用し，連続的な標本空間から得た標本をもとに尤度比推定を行う．最適化問題の枠組みでは，過学習を防ぐために二次の正則化項を導入し，パラメータが一様であるという事前知識を与える．

4. 提 案 手 法

uLSIF の原論文 [18] では，基底関数としてガウスカーネルを用いた．だが，本稿において標本の要素に相当するものは連続的な標本空間から得られる実数値ではなく，離散的な標本空間から得られる単語，バイグラムなどの特徴である．標本空間が連続から離散へと変化したため，従来のようにガウスカーネルを使用できない．離散的な標本空間を扱うためのカーネルとして，文字列カーネルや木カーネルなどが知られているが，これらのカーネルを使用すると推定モデルの定式化や尤度比の効率的な推定が難しい．そこで，次の単純な基底関数 $\{\varphi_l(x)\}_{l=1}^{v}$ を用いる．ここで，x は単語，バイグラムなどの特徴，v は存在しうる特徴の種類数であり，特徴ごとに対応する基底関数を一つ定義する．

$$\varphi_l(x) = \begin{cases} 1 & (x = w_{(l)}) \\ 0 & (x \neq w_{(l)}) \end{cases} \tag{9}$$

この基底関数は，異なる特徴間の関係性を捉えられないが，最終的に導出される推定式が単純な形式となって扱いやすい．添え字 l は v 種類存在する特徴から，特定の特徴を指定する．すなわち，$w_{(l)}$ は v 種類ある特徴のうち，l 種類目の特徴を指す．式 (9) の基底関数を uLSIF の枠組みに当てはめる．推定対象とする尤度比 $r(x)$ を次の線形モデルで表現する．

$$\begin{aligned}\hat{r}(x) &= \sum_{l=1}^{v} \beta_l \varphi_l(x) \\ &= \beta_l(\lambda)\end{aligned} \tag{10}$$

ただし，$x = w_{(l)}$ とする．このモデルは式 (1) で定義した線形

モデルに対応する．

式 (9) からわかるように，基底関数 $\varphi_l(x)$ は，x が $w_{(l)}$ と等しくないときに 0 となる．よって，式 (5) および式 (6) に対応する $\hat{H}_{l,l'}$，$\hat{h}_l$ は次式となる．

$$\begin{aligned}\hat{H}_{l,l'} &= \begin{cases} \frac{1}{n_{de}} \sum_{j=1}^{n_{de}} \varphi_l(x_j^{de}) \varphi_{l'}(x_j^{de}) & (l = l') \\ 0 & (l \neq l') \end{cases} \\ &= \frac{1}{n_{de}} c_{de}(w_{(l)}) \quad (l = l') \\ &= \hat{p}_{de}(w_{(l)}) \end{aligned} \tag{11}$$

$$\begin{aligned}\hat{h}_l &= \frac{1}{n_{nu}} \sum_{j=1}^{n_{nu}} \varphi_l(x_j^{nu}) \\ &= \frac{1}{n_{nu}} c_{nu}(w_{(l)}) \\ &= \hat{p}_{nu}(w_{(l)}) \end{aligned} \tag{12}$$

以上から，要素 $\hat{H}_{l,l'}$ を持つ $v \times v$ 行列 $\hat{\boldsymbol{H}}$ は対角成分のみが残り，それ以外の要素はすべて 0 となる．$c_{de}(w_{(l)})$ は $p_{de}(w_{(l)})$ を持つ確率分布から観測した $w_{(l)}$ の頻度，$c_{nu}(w_{(l)})$ は $p_{de}(w_{(l)})$ を持つ確率分布から観測した $w_{(l)}$ の頻度である．$\hat{p}_{de}(w_{(l)})$ は $p_{de}(w_{(l)})$ の最尤推定量，$\hat{p}_{nu}(w_{(l)})$ は $p_{nu}(w_{(l)})$ の最尤推定量である．

先に述べたように，行列 $\hat{\boldsymbol{H}}$ には対角成分のみが残る．これによって，$r(x)$ の推定量 $\hat{r}(x)$ は次式となり，簡単に求められる．

$$\begin{aligned}\hat{r}(x) &= \hat{\beta}_l(\lambda) \\ &= (\hat{H}_{l,l} + \lambda)^{-1} \hat{h}_l \\ &= \left(\frac{1}{n_{de}} c_{de}(w_{(l)}) + \lambda \right)^{-1} \cdot \frac{1}{n_{nu}} c_{nu}(w_{(l)}) \\ &= \frac{\hat{p}_{nu}(w_{(l)})}{\hat{p}_{de}(w_{(l)}) + \lambda} \end{aligned} \tag{13}$$

uLSIF では，パラメータ β_l の非負制約を取り除いたため，負の値となるパラメータを 0 に丸める必要がある．だが，式 (13) は負の値を取らないため，パラメータの推定量 $\hat{\beta}_l(\lambda)$ がそのまま $r(x)$ の推定量となる．ここで，正則化パラメータ λ は，推定量を保守的に見積もる効果を生む．この推定量は λ が 0 であった場合，尤度比の分母・分子にあたる確率を最尤推定し，それらの比を取った結果に等しい．式 (13) は単純な形をしているが，分母・分子の両方を補正する一般的なスムージング手法と異なり，分子に何も補正がないという特徴的な式になっている．

5. 雑誌名抽出の枠組み

実行する雑誌名抽出は以下の 3 手順からなる．

手順 1. 訓練データの作成： まず，データセットから辞書にある雑誌名を最長一致検索ですべて発見し，その左右にある文字バイグラム対を抽出する（図 2）．抽出した左右バイグラム対は種類ごとに観測頻度を数え上げる．同時に，雑誌名の左右バイグラム対について総頻度を数え上げる．次に，雑誌名を含む任意文字列の左右で観測したバイグラム対のうち，雑誌名の左右でも観測したものについて種類ごとに頻度を数え上げる．同時に，任意文字列の左右バイグラム対について総頻度を数え上げる．任意文字列の探索は，計算時間削減のため，あらかじめ指定した文字列長の範囲内で行う．本稿では，文字列長を 2 文字から 50 文字までとした．左右バイグラム対は，通常のバイグラムよりも低頻度になるため，これを上手く扱う必要がある．

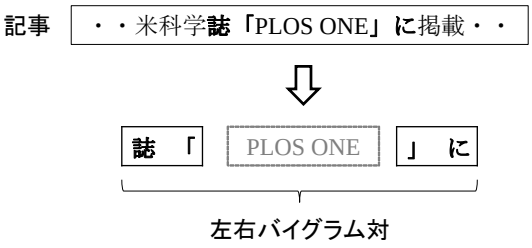


図 2: 左右バイグラム対の抽出例

手順 2. 雑誌名の抽出： 各記事から雑誌名の候補および左右バイグラム対を抽出する．この処理は，抽出の開始位置を固定し，候補の文字列長を範囲内で変化させて行う．この文字列長は，雑誌名か否かの判断が難しい 1 文字を除いて，2 文字から 50 文字までとした．記事の先頭から抽出を始め，候補の長さを一文字ずつ伸ばしていく．候補の長さが 50 文字に達した，あるいは右バイグラムが記事末尾に達した後は開始位置を一文字だけ右にシフトして再度抽出を始める．

抽出した候補について雑誌名らしさのスコアを計算する．スコアは下式に示す尤度比とした．

$$Score(x) = \frac{P(x \mid O_J)}{P(x \mid O_A)} \tag{14}$$

x は候補の左右に出現したバイグラム対であり，4 節において，離散的な標本空間から得られた特徴に対応する．O_J と O_A はバイグラム対がそれぞれ雑誌名，任意文字列の左右で出現することを表す．尤度比は，バイグラム対が雑誌名の左右に出現する確率と任意文字列の左右に出現する確率の比で表される．確率は本節の手順 1 で数え上げた観測頻度に基づいて推定される．尤度比が 1 より大きければバイグラム対は雑誌名の左右に現れやすい傾向にあり，逆に 1 より小さければ任意文字列の左右に現れやすい傾向にあるといえる．スコアが大きいほど，バイグラム対の間にある候補が雑誌名らしいと推定する．スコア降順から上位 N 件を抽出し，人手で正誤判定をする．そして，正解と判定した文字列のみを辞書に追加する．

手順 3. 繰り返し： 十分な数の雑誌名が得られるまでステップ 1 と 2 を繰り返す．

6. 評 価 実 験

提案手法の有効性を検証するため，日本語科学ニュース記事からの雑誌名抽出を試みる．この抽出は，5 節で述べたブートストラップ方式の枠組みに基づいて実行する．提案手法は，低頻度の観測事象から測定される尤度比の推定値を保守的に見積もる．それゆえ，高頻度に基づく，信頼できる尤度比を重点的に利用するタスクで効果がある．ブートストラップ法では，信頼できるインスタンスを多く集めることが求められ，分類スコ

表 1: シードとして使用する雑誌名

Scientific Reports
サイエンティフィック・リポーツ
サイエンティフィック・リポーツ（Scientific Reports）
サイエンティフィックリポーツ
サイエンティフィックリポーツ（Scientific Reports）
PLOS ONE
プロス・ワン
プロス・ワン（PLOS ONE）
プロスワン
プロスワン（PLOS ONE）

表 2: 雑誌名の表記例

英名
Neuron, Cell Research
和名
ニューロン, セル・リサーチ
英名・和名の併記
ニューロン（Neuron）, セル・リサーチ（Cell Research）
補足情報付き
ニューロン電子版, セル・リサーチ（電子版）

アとして使用する尺度によって性能が左右される．以上のことから，提案手法によって推定した尤度比を，分類スコアとしてブートストラップの枠組みに取り入れることで，雑誌名抽出の性能向上が期待できる．

6.1 実験概要

実験では，学術雑誌名を含む可能性の高いニュース記事をデータセットとして使用する．具体的には，複数の日本語ニュースサイトから過去およそ 10 年分のニュース記事を収集し，「学誌 OR 論文誌 OR 学術誌」という検索条件で絞り込んだ合計 30,076 記事を使用する．雑誌名抽出の際は，記事本文のみを参照する．初期の雑誌名辞書として，表 1 に示すシードを使用する．雑誌名の抽出元が日本語記事であるため，和文雑誌名および英文雑誌名が多く抽出される．雑誌名の表記は，大別して表 2 に示すパターンがある．雑誌名の抽出を 5 回実行し，それぞれでスコアの降順 1,000 件を抽出する．1,000 件には，シードおよび過去に抽出した候補を含めない．また，同じ候補であっても，左右にあるバイグラム対が異なれば雑誌名らしさのスコアも異なる．同じ候補で複数のスコアがある場合は，最高のスコアをその候補のスコアとして扱う．なお，繰り返しごとに辞書に加える雑誌名は，人手で正誤判定した際に正解と判定した文字列のみである．これは，スコアの推定方法による抽出性能の変化を正確に測定するためである．性能評価の指標として，ブートストラップの繰り返し毎に累積の適合率を算出する．

6.2 比較手法

二種類の試行に対し，表 3 に示す集計表を考えたとき，推定対象の尤度比を次式として定義する．本節で述べる手法は，5 節の式 (14) で示したスコアの推定に用いられる．

$$\mathrm{LR}(x) = \frac{p_1}{p_2} \tag{15}$$

表 3: 二種類の試行に対する集計表

	Outcome		
Trial	Success	Failure	Total
1	k_1	$n_1 - k_1$	n_1
2	k_2	$n_2 - k_2$	n_2

n_t，k_t，p_t は，試行 t ($\in \{1, 2\}$) の試行回数，成功回数，成功確率を示す．本稿で，試行 1 は雑誌名の左右バイグラム対を抽出すること，試行 2 は任意文字列の左右バイグラム対を抽出することを表す．特定の左右バイグラム対 x が与えられたもとで，抽出した左右バイグラム対が x と一致することを成功，それ以外を失敗とする．このとき，n_t，k_t，p_t は，抽出した左右バイグラム対の総頻度，x の観測頻度，x の出現確率にそれぞれ置き換えられる．以下では，これらの記号を用いて，比較手法を簡潔に説明する．手法 1 は，ベースラインとなる最も単純な尤度比推定法である．手法 2 と手法 3 は，推定量にかかるバイアスを抑制するアプローチである．手法 4 と手法 5 は，観測頻度の低さに応じて推定値を保守的に見積もるアプローチである．

手法 1：最尤推定量を用いた方法（MLE）

尤度比を構成する確率を最尤推定で求め，その比を取る．次に定義式を示す．

$$\mathrm{LR}_{\mathrm{MLE}}(x) = \frac{\hat{p}_1}{\hat{p}_2} \tag{16}$$

ただし，

$$\hat{p}_t = \frac{k_t}{n_t} \qquad (t = 1, 2) \tag{17}$$

である．

手法 2：経験ベイズによる方法（EB）

まず，p_t に関する事前分布としてベータ分布 $\beta(a_t, b_t)$ を仮定する．そして，事後分布の期待値 p_t^* を p_t の推定量として求めて，その比を取る．次に定義式を示す．本手法は，低頻度の観測事象から尤度比を推定する場合に有効である [12]．

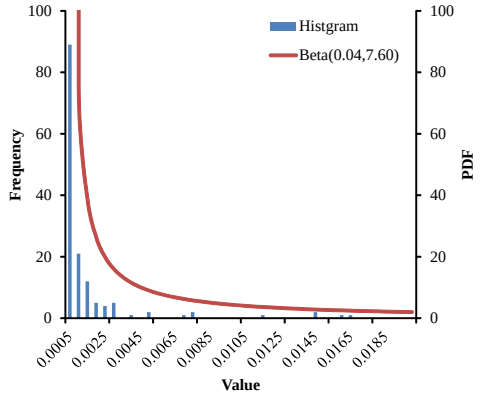
$$\mathrm{LR}_{\mathrm{EB}}(x) = \frac{p_1^*}{p_2^*} \tag{18}$$

p_t^* は次式で表される．

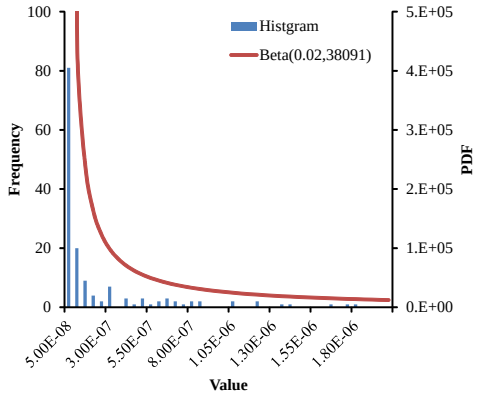
$$p_t^* = \frac{k_t + \hat{a}_t}{n_t + \hat{a}_t + \hat{b}_t} \qquad (t = 1, 2) \tag{19}$$

本手法の原論文 [12] では，ハイパーパラメータ $\hat{a}_t, \hat{b}_t$ を最尤推定によって求めている．しかし，本稿では原論文と異なり，n_t が大きい．この場合，原論文の方法では最尤推定解を解析的に求めることが困難である．それゆえ，ハイパーパラメータをデータから近似的に推定することにした．具体的には，特定位置（雑誌名・任意文字列の左右）における左右バイグラム対の出現割合について，その平均・分散をデータから求め，ベータ分布のそれと一致するようにハイパーパラメータを推定した．

ハイパーパラメータの推定結果を図 3 に示す．各グラフの横軸は左右バイグラム対の出現割合に関する区間，縦軸は区間に対応する左右バイグラム対の種類数，ベータ分布の確率密度



(a) 雑誌名の周囲



(b) 任意文字列の周囲

図 3: ハイパーパラメータの推定結果

である．ヒストグラムの形状はべき乗則に従う．推定したハイパーパラメータを持つベータ分布は，概ねヒストグラムの形状を近似できている．

手法 3：Modified MUE を用いた方法（MMUE）

p_t を中位不偏推定量（MUE: Modian Unbiased Estimator）$\tilde{p}_t$ によって推定し，その比を取る．次に定義式を示す．本手法は，標本のサイズが小さい場合に，推定量にかかるバイアスを抑制できることが報告されている [11]．

$$\mathrm{LR}_{\mathrm{MMUE}}(x) = \frac{\tilde{p}_1}{\tilde{p}_2} \tag{20}$$

ただし，

$$\tilde{p}_t = \frac{\tilde{p}_t^L + \tilde{p}_t^U}{2} \qquad (t = 1, 2) \tag{21}$$

$$\tilde{p}_t^L = F^{-1}(.5 \mid \hat{a}_t = k_t, \hat{b}_t = n_t - k_t + 1)$$

$$\tilde{p}_t^U = F^{-1}(.5 \mid \hat{a}_t = k_t + 1, \hat{b}_t = n_t - k_t + 2)$$

である．$F^{-1}(Q \mid \hat{a}_t, \hat{b}_t)$ はハイパーパラメータ $\hat{a}_t$，$\hat{b}_t$ を持つベータ分布の Q 分位点である．

手法 4：信頼区間を用いた方法（CI）

尤度比を構成する分母・分子の確率推定に信頼区間の上限値・下限値を使用する．本手法は，提案手法と同様に，低頻度の観測事象から推定される尤度比を保守的に見積もる作用がある．次に定義式を示す．

$$\mathrm{LR}_{\mathrm{CI}}(x) = \frac{\theta_{1-}}{\theta_{2+}} \tag{22}$$

θ_{t+} は信頼区間の上限値，θ_{t-} は信頼区間の下限値である．信頼区間の構築には，二項分布を正規近似する方法がよく用いられる．しかし，正規近似が上手く働く状況は二項分布の母平均が 0.5 に近い場合であり，本稿で扱うのは母平均が 0 に近い場合である．このときに正規分布による近似を用いた場合，下限が負の値となることが知られている．そこで，信頼区間の両端が 0 や 1 に近い場合に上手く働くとされる Wilson score interval を用いる．この区間は次式で定義される．

$$\theta_{t\pm} = \frac{\hat{p}_t + \frac{z_{\alpha/2}^2}{2n_t} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{n_t} + \frac{z_{\alpha/2}^2}{4n_t^2}}}{1 + z_{\alpha/2}/n_t} \tag{23}$$

$$(t = 1, 2)$$

ここで，$z_{\alpha/2}$ は標準正規分布の上側 $100(\alpha/2)\%$点を表す．

この方法は二つの信頼区間を構築する必要があるため，二つの信頼係数をパラメータとして持つ．それぞれの信頼係数を個別に変化させて最適値を探すこともできるが，提案手法と対等な比較を行うため，二つの信頼係数に同じ値を採用することにした．実験では，信頼区間の幅が片側 95%，片側 99%，両側 95%，両側 95%となるように信頼係数を設定した．

手法 5：提案手法

提案手法は，最尤推定した尤度比である式 (16) の分母に正則化パラメータ λ を加算した式で表される． 次に定義式を示す．

$$\mathrm{LR}_{\mathrm{Proposed}}(x) = \frac{\hat{p}_1}{\hat{p}_2 + \lambda} \tag{24}$$

本手法は定数 λ をパラメータとして持つ．λ の値が大きくなるにつれて標本サイズの大きさを重視し，標本サイズが小さいとき，尤度比を保守的に見積もる．λ の値を 10^{-9} から 10^{-1} まで 10 倍ずつ変化させ，シードから作成した訓練データを使用して左右バイグラム対の尤度比を推定した．λ の値を 10^{-8} および 10^{-5} としたとき，尤度比の降順上位 10 件を表 4，5 に示す． 表 4 の下線部分に着目すると，観測頻度が 10 以下のバイグラム対が上位に位置している．この場合，λ の値が小さすぎるため，低頻度の影響を抑制できていない．次に，表 5 の下線部分に着目すると，低頻度の影響を低減できているが，任意文字列の左右における観測頻度が大きく異なるにもかかわらず，雑誌名の左右における観測頻度が近いバイグラム対が類似した推定値を持つ．これは，λ の値が大きすぎるゆえ，観測頻度を過剰評価した結果と考える．以上から，適切な λ の値が 10^{-6} あるいは 10^{-7} 付近になると予想し，この二つの値を，それぞれパラメータの値として採用した．

6.3 評価尺度

ブートストラップ法の繰り返し毎に，抽出結果の上位 1,000

表 4: $\lambda = 10^{-8}$（推定値の降順 10 件）

バイグラム対		観測頻度		推定値
左	右	雑誌名	任意文字列	
誌「	（P	59	162	139,819.30
ル「	」に	11	26	125,843.00
典：	）[	4	4	120,155.90
も（	提供	3	5	83,915.85
、米	誌提	2	2	70,496.91
】「	」発	2	2	70,496.91
誌「	（サ	3	10	62,434.71
学誌	で 1	38	257	57,955.54
】「	」の	2	6	52,342.16
E（	)」	2	8	46,371.27

表 5: $\lambda = 10^{-5}$（推定値の降順 10 件）

バイグラム対		観測頻度		推定値
左	右	雑誌名	任意文字列	
誌「	」に	601	5,590	16,155.26
学誌	に発	431	9,366	9,270.70
学誌	に掲	239	7,074	5,850.36
学誌	電子	151	5,820	3,998.16
誌「	」で	90	788	3,544.913
誌「	」電	64	682	2,546.98
誌「	（P	59	162	2,473.94
学誌	（電	60	1,703	2,170.81
誌「	」(	54	749	2,135.02
学誌	で発	56	1,707	2,025.37

件について人手で正誤判定し，累積の適合率を算出する．適合率の式を次に示す．

$$\text{適合率} = \frac{\text{これまでに正解とした文字列数}}{\text{これまでに正誤判定した文字列数}} \tag{25}$$

本来ならば，再現率や F 値も用いた性能評価が望ましい．だが，再現率や F 値の算出には，雑誌名を網羅した，偏りのない正解リストが必要となる．現状として，そのようなリストを用意することは難しいため，今回は適合率のみで性能評価をする．

6.4 実験結果と考察

各手法において得られた雑誌名の数および適合率（ともに累積）を表 6 に示す．表中の ♣ は各繰り返しにおいてトップの適合率，$*$ は 2 番目に良い適合率を表している．また，各繰り返しにおける適合率の変化を図 4 に示す．

表 6 と図 4 から，$\lambda = 10^{-6}$ としたときの提案手法は繰り返しの 1 回目，2 回目，および 4 回目でトップの適合率，繰り返しの 3 回目と 5 回目でも 2 番目に良い適合率となった．次いで，$\lambda = 10^{-7}$ としたときの提案手法は，繰り返しの 3 回目と 5 回目でトップ，繰り返しの 2 回目で 2 番目に良い適合率となった．片側 95%信頼区間を使用した手法は，繰り返しの 4 回目で 2 番目に良い適合率，片側 99%，両側 95%，両側 99%信頼区間を用いた手法も繰り返しの 1 回目で 2 番目に良い適合率となった．それに対し，MLE，EB，MMUE は互いに近い適合率となったが，信頼区間を用いた手法と提案手法よりも低い性能を示した．

このタスクでは，低頻度の左右バイグラム対が多く観測され，

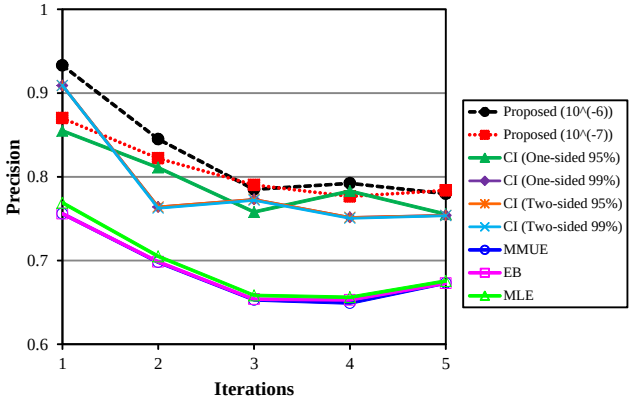


図 4: 各手法における適合率の変化

これらのバイグラム対をどう扱うかが性能差に大きな影響を与える．単純に観測頻度の比を取る MLE や，バイアスを取り除くアプローチである MMUE は，低頻度のバイグラム対から推定される尤度比を不当に高く見積もり，雑誌名の抽出性能に悪影響を与えたと考えられる．EB は，確率推定において低頻度を上手く活用する手段として知られている．しかし，個々の確率を経験ベイズ法で推定し，その比を取る推定方法は，MLE や MMUE と同様に，低頻度のバイグラム対から推定される尤度比を高く見積もってしまう．

一方で，CI と提案手法は良い性能を示した．これらの手法は，低頻度の左右バイグラム対から推定される尤度比を低めに（保守的に）見積もるアプローチである．このことから，本実験においては，提案手法の保守的な推定による有効性が確認できた．CI の適合率が提案手法よりも低い理由として，CI は低頻度から推定される尤度比を十分低く見積もれていないことが原因と考えている．全バイグラム対の総頻度に対し，特定のバイグラム対の観測頻度は非常に小さい．結果として，尤度比を構成する確率の分散は小さくなり，信頼区間の下限・上限が最尤推定値と近くなる．提案手法はパラメータ λ を大きく変化させたため，CI よりも良い性能を示したと考えられる．また，EB，MMUE，CI は尤度比の分母・分子両方に補正を加えているのに対し，提案手法は分母のみを考慮する特性がある．そのため，提案手法は推定すべきパラメータが一つでよいという利点がある．

7. まとめと今後の課題

ブートストラップ法では，得られた分類結果を繰り返して使い回す観点から，それぞれの繰り返しで多くの信頼できる分類結果を得ることが求められる．尤度比を利用して，分類結果を得るとすると，尤度比の保守的な推定法を用いることが考えられる．本稿では，尤度比の直接推定法を標本空間が離散である場合に応用し，尤度比の直接推定法で導入される正則化項の扱いで保守的な推定量が得られることを指摘した．そして，科学ニュース記事から雑誌名を抽出する実用的なタスクにおいて，保守的な推定量とブートストラップ法の組み合わせが良い結果をもたらすことを報告した．

表 6: 各手法の抽出雑誌数および適合率（ともに累積）

Methods	# of iterations				
	1	2	3	4	5
MLE	769 (.769)	1,410 (.705)	1,975 (.658)	2,625 (.656)	3,378 (.676)
EB	756 (.756)	1,397 (.699)	1,962 (.654)	2,612 (.653)	3,365 (.673)
MMUE	756 (.756)	1,397 (.699)	1,959 (.653)	2,597 (.649)	3,366 (.673)
CI (One-sided 95%)	855 (.855)	1,622 (.811)	2,274 (.758)	∗ 3,132 (.783)	3,777 (.755)
CI (One-sided 99%)	∗ 909 (.909)	1,528 (.764)	2,319 (.773)	3,006 (.752)	3,770 (.754)
CI (Two-sided 95%)	∗ 909 (.909)	1,528 (.764)	2,319 (.773)	3,006 (.752)	3,769 (.754)
CI (Two-sided 99%)	∗ 909 (.909)	1,525 (.763)	2,316 (.772)	3,002 (.751)	3,768 (.754)
Proposed ($\lambda = 10^{-6}$)	♣ 933 (.933)	♣ 1,690 (.845)	∗ 2,355 (.785)	♣ 3,169 (.792)	∗ 3,899 (.780)
Proposed ($\lambda = 10^{-7}$)	870 (.870)	∗ 1,644 (.822)	♣ 2,371 (.790)	3,106 (.777)	♣ 3,919 (.784)

今後の課題として，ブートストラップ法の繰り返しの回数や一回での抽出件数を変化させての検証を行う．加えて，より評価の容易なベンチマークタスクでの実験を行い，詳細な定量的評価をする必要がある．さらに，提案手法のパラメータである正則化パラメータを自動的に決定する方法も検討する．

文　献

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 92–100, 1998.

[2] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, 1995.

[3] Rada F. Mihalcea and Dan I. Moldovan. A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal on Artificial Intelligence Tools*, Vol. 10, No. 1-2, pp. 5–21, 2001.

[4] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 100–110, 1999.

[5] Zornitsa Kozareva. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 15–21, 2006.

[6] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 152–159, 2006.

[7] Nicola Ueffing. Self-training for machine translation. In *NIPS workshop on Machine Learning for Multilingual Information Access*, 2006.

[8] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the 3rd ACM international conference on Web search and data mining*, pp. 101–110, 2010.

[9] 中西健太郎, 田中利幸, 上田修功. 尤度比に基づく順位づけ関数による受信者操作特性曲線下面積の漸近的性質. 電子情報通信学会技術研究報告, 第 114 巻, pp. 101–110, 2015.

[10] Zellig S. Harris. Distributional structure. *Word*, Vol. 10, No. 2-3, pp. 146–162, 1954.

[11] Michael Parzen, Stuart Lipsitz, Joseph Ibrahim, and Neil Klar. An estimate of the odds ratio that always exists. *Journal of Computational and Graphical Statistics*, Vol. 11, No. 2, pp. 420–436, 2002.

[12] Kobkun Raweesawat, Yupaporn Areepong, Katechan Jampachaisri, and Saowanit Sukparungsee. Odds ratios estimation of rare event in binomial distribution. *Journal of Probability and Statistics*, 2016.

[13] Mark Johnson. Confidence intervals on likelihood estimates for estimating association strengths. Unpublished technical report, 1999.

[14] 菊地真人, 山本英子, 吉田光男, 岡部正幸, 梅村恭司. 条件付き確率の保守的な推定. 電子情報通信学会論文誌, Vol. 100, No. 4, pp. 544–555, 2017.

[15] Cynthia Rudin, Benjamin Letham, and David Madigan. Learning theory analysis for association rules and sequential event prediction. *Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 3441–3492, 2013.

[16] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[17] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pp. 1433–1440, 2008.

[18] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, Vol. 10, pp. 1391–1445, July 2009.