Vol.2014-NL-218 No.2
2014/9/1

uLSIF を用いた事例への重み付けによる語彙曖昧性解消の領域適応

新納 浩幸[1,a)]　菊池 裕紀[1,b)]　佐々木 稔[1,c)]　古宮 嘉那子[1,d)]

概要：本論文では語義曖昧性解消の教師なし領域適応の問題に対して，ソース領域の訓練事例に重みをつけた重み付き学習を利用する．事例の重みを算出するために，ターゲット領域に対するソース領域の確率密度比を直接モデル化する uLSIF を用いる．また uLSIF では，通常，基底関数にガウスカーネルを利用するが，本論文では線形カーネルを利用する．また重み付き学習では，通常，ロジスティック回帰や最大エントロピー法を用いるが，ここでは SVM を利用する．その上で確率密度比が極端に小さい，あるいは大きい事例のみに重みを与える方法を提案する．

Shinnou Hiroyuki[1,a)]　Kikuchi Hironori[1,b)]　Sasaki Minoru[1,c)]　Komiya Kanako[1,d)]

1. はじめに

本論文では語義曖昧性解消の教師なし領域適応の問題に対して，ソース領域の訓練事例に重みをつけた重み付き学習を利用する．

自然言語処理の多くのタスクで教師付き学習が利用されているが，そこでは訓練データとテストデータの領域が異なるという領域適応の問題が生じている [13][8]．例えば「ゴルフ」という単語には sport と car の意味があり，その語義曖昧性解消を考えた場合，学習元のコーパスがスポーツ記事であれば主に sport と判定される規則が学習されるが，その規則を車の記事に適用すると誤る場合が多いという問題である．このような領域適応の問題に対する手法には，大きく分けて事例ベースの手法と素性ベースの手法が存在する [12]．素性ベースの手法では，概略，ソース領域の素性空間をターゲット領域に合うように変換する [3][1][10][11][4]．また事例ベースの手法では，概略，訓練事例に重みをつけて重み付き学習を行う [5][18]．本論文では事例ベースの手法を利用する．

事例ベースの手法ではソース領域とターゲット領域間に共変量シフトを仮定することが多い．今，事例を $\boldsymbol{x}$，クラスを c としたとき，領域適応は $P_T(c|\boldsymbol{x})$ を求めることで解決できる．共変量シフトとは $P_S(c|\boldsymbol{x}) = P_T(c|\boldsymbol{x})$ であるが $P_S(\boldsymbol{x}) \neq P_T(\boldsymbol{x})$ という仮定である．自然言語処理ではある領域で出現した文が他の領域で出現しても，その文自体の意味が変化することは稀であるため，通常，自然言語処理のタスクでは成立している仮定である．そして共変量シフト下では $\boldsymbol{x}$ の確率密度比 $w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ を重みとして，重み付き対数尤度を最大化するパラメータを求めることで，$P_T(c|\boldsymbol{x})$ を構築するアプローチが取られる [14]．

確率密度比 $w(\boldsymbol{x})$ の算出方法は，$P_T(\boldsymbol{x})$ と $P_S(\boldsymbol{x})$ をそれぞれモデル化して求めその比を取る方法と，$w(\boldsymbol{x})$ を直接モデル化する方法が存在する [20]．前者の研究としては Jiang の研究 [5] と齋木の研究 [21] があるが，自然言語処理のタスクでは後者の手法は試みられていない．本論文では後者の手法として uLSIF を利用する [6]．uLSIF は，概略，b 個の基底関数 $\psi_l(\boldsymbol{x})$ を設定し，その線形和で $w(\boldsymbol{x})$ をモデル化する．

$$w(\boldsymbol{x}) = \sum_{l=1}^{b} \alpha_l \psi_l(\boldsymbol{x})$$

また基底関数 $\psi_l(\boldsymbol{x})$ には，通常，ガウスカーネルが利用される．このためガウスカーネルの幅 σ がパラメータとして増える．本論文では基底関数に自然言語処理では一般に利用される線形カーネルを利用する．これによってパラメータ σ が省かれ，より適切な確率密度比が推定できる．

[1] 茨城大学工学部情報工学科
Ibaraki University, Nakanarusawa 4-12-1, Hiachi, Ibaraki 316–8511, Japan
[a)] shinnou@mx.ibaraki.ac.jp
[b)] 13nm705g@hcs.ibaraki.ac.jp
[c)] msasaki@mx.ibaraki.ac.jp
[d)] kkomiya@mx.ibaraki.ac.jp

本論文で扱うタスクは語義曖昧性解消の領域適応である．語義曖昧性解消の場合，対象単語毎に $w(\boldsymbol{x})$ をモデル化する必要がある．またターゲット領域の対象単語毎の用例の数は少ないため，確率密度比の算出はより困難である．このため重み付き学習には通常の重み付き対数尤度の最大化ではなく，重み付き SVM を利用する．更に，事例毎の細かな重みではなく，重みが極端に小さい事例と極端に大きい事例だけに重みを与える手法を提案する．

実験では現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese, BCCWJ [7]）における 3 つの領域 OC (Yahoo! 知恵袋)，PB (書籍) 及び PN (新聞) を利用する．SemEval-2 の日本語 WSD タスク [9] ではこれらのコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．すべての領域である程度の頻度が存在する多義語 16 単語を対象にして，WSD の領域適応の実験を行う．領域適応としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りが存在する．結果 $16 \times 6 = 96$ 通りの WSD の領域適応の問題に対して実験を行った．その結果，線形カーネルと重み付き SVM を利用する効果および重みが極端に小さい事例と極端に大きい事例だけに重みを与える提案手法の効果を確認できた．

2. 共変量シフト下の領域適応

対象単語 w の語義の集合を C，また w の用例 $\boldsymbol{x}$ 内の w の語義を c と識別したときの損失関数を $l(\boldsymbol{x}, c, d)$ で表す．d は w の語義を識別する分類器である．$P_T(\boldsymbol{x}, c)$ をターゲット領域上の分布とすれば，本タスクにおける期待損失 L_0 は以下で表せる．

$$L_0 = \sum_{\boldsymbol{x},c} l(\boldsymbol{x}, c, d) P_T(\boldsymbol{x}, c)$$

また $P_S(\boldsymbol{x}, c)$ をソース領域上の分布とすると以下が成立する．

$$L_0 = \sum_{\boldsymbol{x},c} l(\boldsymbol{x}, c, d) \frac{P_T(\boldsymbol{x}, c)}{P_S(\boldsymbol{x}, c)} P_S(\boldsymbol{x}, c)$$

ここで共変量シフトの仮定から

$$\frac{P_T(\boldsymbol{x}, c)}{P_S(\boldsymbol{x}, c)} = \frac{P_T(\boldsymbol{x}) P_T(c|\boldsymbol{x})}{P_S(\boldsymbol{x}) P_S(c|\boldsymbol{x})} = \frac{P_T(\boldsymbol{x})}{P_S(\boldsymbol{x})}$$

となり，$w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ とおくと以下が成立する．

$$L_0 = \sum_{\boldsymbol{x},c} w(\boldsymbol{x}) l(\boldsymbol{x}, c, d) P_S(\boldsymbol{x}, c)$$

訓練データを $D = \{(\boldsymbol{x}_i, c_i)\}_{i=1}^N$ とし，$P_S(\boldsymbol{x}, c)$ を経験分布で近似すれば，

$$L_0 \approx \frac{1}{N} \sum_{i=1}^N w(\boldsymbol{x}_i) l(\boldsymbol{x}_i, c_i, d)$$

となるので，期待損失最小化の観点から考えると，共変量シフトの問題は以下の式 L_1 を最小にする d を求めればよいことがわかる．

$$L_1 = \sum_{i=1}^N w(\boldsymbol{x}_i) l(\boldsymbol{x}_i, c_i, d) \tag{1}$$

分類器 d として以下の事後確率最大化推定に基づく識別を考える．

$$d(\boldsymbol{x}) = \arg\max_c P_T(c|\boldsymbol{x})$$

また損失関数として対数損失 $-\log P_T(c|\boldsymbol{x})$ を用いれば，式 (1) は以下となる．

$$L_1 = -\sum_{i=1}^N w(\boldsymbol{x}_i) \log P_T(c_i|\boldsymbol{x}_i)$$

つまり，分類問題の解決に $P_T(c|\boldsymbol{x}, \boldsymbol{\lambda})$ のモデルを導入するアプローチを取る場合，共変量シフト下での学習では，確率密度比を重みとした以下に示す重み付き対数尤度 $L(\boldsymbol{\lambda})$ を最大化するパラメータ $\boldsymbol{\lambda}$ を求める形となる．

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^N w(\boldsymbol{x}_i) \log P_T(c_i|\boldsymbol{x}_i, \boldsymbol{\lambda}) \tag{2}$$

モデルとしては以下の式で示される最大エントロピー法が用いられる．

$$P_T(c|\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{\lambda})} \exp\left(\sum_{j=1}^M \lambda_j f_j(\boldsymbol{x}, c)\right) \tag{3}$$

$\boldsymbol{x} = (x_1, x_2, \cdots, x_M)$ が入力，c がクラスである．関数 $f_j(\boldsymbol{x}, c)$ は素性関数であり，実質 $\boldsymbol{x}$ の真のクラスが c のときに x_j を返し，そうでないとき 0 を返す関数に設定される．$Z(\boldsymbol{x}, \boldsymbol{\lambda})$ は正規化項であり，以下で表せる．

$$Z(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum_{c \in C} \exp\left(\sum_{j=1}^M \lambda_j f_j(\boldsymbol{x}, c)\right) \tag{4}$$

そして $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_M)$ が素性に対応する重みパラメータとなる．

3. 確率密度比の算出

確率密度比 $w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ の算出法は大きく 2 つに分類できる．1 つは $P_S(\boldsymbol{x})$ と $P_T(\boldsymbol{x})$ を各々推定し，その比を取る手法であり，もう 1 つは $w(\boldsymbol{x})$ を直接モデル化する手法である [20].

ここでは論文 [6] において提案された拘束無し最小二乗重要度適合法 (unconstrained Least-Squares Importance Fitting, uLSIF) を利用する．

3.1 uLSIF による算出

ソース領域内のデータを $\{\boldsymbol{x}_i^s\}_{i=1}^{N_s}$，ターゲット領域内のデータを $\{\boldsymbol{x}_i^t\}_{i=1}^{N_t}$ とする uLSIF では確率密度比 $w(\boldsymbol{x})$ を以

下の式でモデル化する．

$$\begin{aligned} w(\boldsymbol{x}) &= \sum_{l=1}^{b} \alpha_l \psi_l(\boldsymbol{x}) \\ &= \boldsymbol{\alpha} \cdot \boldsymbol{\psi}(\boldsymbol{x}) \end{aligned}$$

ただしここで，$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_b)$，$\boldsymbol{\psi}(\boldsymbol{x}) = (\psi_1(\boldsymbol{x}), \psi_2(\boldsymbol{x}), \cdots, \psi_b(\boldsymbol{x}))$ である．また α_l は正の実数であり，$\psi_l(\boldsymbol{x})$ は基底関数と呼ばれるソース領域のデータ $\boldsymbol{x}$ から正の実数値への関数である．uLSIF では，概略，自然数 b と基底関数 $\boldsymbol{\psi}(\boldsymbol{x})$ を定めた後に，パラメータ $\boldsymbol{\alpha}$ を推定する手順をとる．

説明の都合上，b と $\boldsymbol{\psi}(\boldsymbol{x})$ が定まった後の $\boldsymbol{\alpha}$ の推定を先に説明する．$w(\boldsymbol{x})$ のモデルを $\hat{w}(\boldsymbol{x})$ とおくと，パラメータ α_l を推定するには，$w(\boldsymbol{x})$ と $\hat{w}(\boldsymbol{x})$ の平均 2 乗誤差 $J_0(\boldsymbol{\alpha})$ を最小にするような $\boldsymbol{\alpha}$ を求めれば良い．$w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ に注意すると，$J_0(\boldsymbol{\alpha})$ は以下のように変形できる．

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &= \frac{1}{2}\int (\hat{w}(\boldsymbol{x}) - w(\boldsymbol{x}))^2 P_S(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{1}{2}\int \hat{w}(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} - \int \hat{w}(\boldsymbol{x}) w(\boldsymbol{x}) P_S(\boldsymbol{x}) d\boldsymbol{x} \\ &\quad + \frac{1}{2}\int w(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{1}{2}\int \hat{w}(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} - \int \hat{w}(\boldsymbol{x}) P_T(\boldsymbol{x}) d\boldsymbol{x} \\ &\quad + \frac{1}{2}\int w(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

3 項目の式は定数なので，$J_0(\boldsymbol{\alpha})$ を最小にするには，以下の $J(\boldsymbol{\alpha})$ を最小にすればよい．

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\int \hat{w}(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} - \int \hat{w}(\boldsymbol{x}) P_T(\boldsymbol{x}) d\boldsymbol{x}$$

$J(\boldsymbol{\alpha})$ を経験分布で近似した $\widehat{J}(\boldsymbol{\alpha})$ は以下となる．

$$\begin{aligned} \widehat{J}(\boldsymbol{\alpha}) &= \frac{1}{2N_s}\sum_{i=1}^{N_s} \widehat{w}(\boldsymbol{x}_i^s)^2 - \frac{1}{N_t}\sum_{j=1}^{N_t} \widehat{w}(\boldsymbol{x}_j^t) \\ &= \frac{1}{2}\sum_{l,l'=1}^{b} \alpha_l \alpha_{l'} \left(\frac{1}{N_s}\sum_{i=1}^{N_s} \psi_l(\boldsymbol{x}_i^s)\psi_{l'}(\boldsymbol{x}_i^s) \right) \\ &\quad - \sum_{l=1}^{b} \alpha_l \left(\frac{1}{N_t}\sum_{j=1}^{N_t} \psi_l(\boldsymbol{x}_j^t) \right) \\ &= \frac{1}{2}\boldsymbol{\alpha}^T \widehat{H} \boldsymbol{\alpha} - \widehat{h}^T \boldsymbol{\alpha} \end{aligned} \tag{5}$$

ここで $\widehat{H}$ は $b \times b$ の行列であり，その l 行 l' 列の要素 $\widehat{H}_{l,l'}$ は以下である．

$$\widehat{H}_{l,l'} = \frac{1}{N_s}\sum_{i=1}^{N_s} \psi_l(\boldsymbol{x}_i^s)\psi_{l'}(\boldsymbol{x}_i^s)$$

また $\widehat{h}$ は b 次元のベクトルであり，その l 次元目の要素 $\widehat{h}_l$ は以下である．

$$\widehat{h}_l = \frac{1}{N_t}\sum_{j=1}^{N_t} \psi_l(\boldsymbol{x}_j^t)$$

$\widehat{J}(\boldsymbol{\alpha})$ の最小値を求める際に正則化を行う．このとき付加する正則化項を L2 ノルムに設定し，$\boldsymbol{\alpha} > 0$ の条件を外して，以下の最小化問題を解く．ここでパラメータ λ が導入されることに注意する．λ は基底関数を設定する際に決められる．

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2}\boldsymbol{\alpha}^T \widehat{H} \boldsymbol{\alpha} - \widehat{h}^T \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^T \boldsymbol{\alpha} \right]$$

この最小化問題は制約のない凸 2 次計画問題であるために，唯一の大域解が得られる．その解は以下である．

$$\tilde{\boldsymbol{\alpha}} = (\widehat{H} + \lambda I_b)^{-1} \widehat{h}^T \tag{6}$$

最後に $\boldsymbol{\alpha} > 0$ の条件に合わせるように，以下の調整を行う．

$$\begin{aligned} \widehat{\boldsymbol{\alpha}} &= ((\max(0, \tilde{\alpha_1}), \max(0, \tilde{\alpha_2}), \cdots, \max(0, \tilde{\alpha_b})) \\ &= \max(0_b, \tilde{\boldsymbol{\alpha}}) \end{aligned} \tag{7}$$

パラメータ b と基底関数の設定であるが，まず，b については以下で設定する．

$$b = \min(100, N_t)$$

次にターゲット領域のデータから重複を許さずに b 個の点をランダムに取り出す．それらの点を $\{\boldsymbol{x}_j^t\}_{j=1}^b$ とおく．そして基底関数 $\psi_l(\boldsymbol{x})$ を以下のガウスカーネルで定義する．

$$\psi_l(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_l^t) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}_l^t||^2}{\sigma^2} \right)$$

以上より，確率密度比を求めるために残されているパラメータは正則化項の係数 λ とガウスカーネルの幅 σ の 2 つである．これらのパラメータはグリッドサーチの交差検定で求める．まずソース領域のデータとターゲット領域のデータをそれぞれ交わりのない R 個の部分集合に分割する．それらの部分集合の中で r 番目の部分集合を除き，残りを結合した集合を作る．それらを新たなソース領域のデータとターゲット領域のデータと見なす．そして λ と σ をある値に設定し，式 (6) と式 (7) より $\boldsymbol{\alpha}$ を求め，式 (5) より $\widehat{J}(\boldsymbol{\alpha})^{(r)}$ の値を求める．r を 1 から R まで変化させることで，R 個の $\widehat{J}(\boldsymbol{\alpha})^{(r)}$ の値が求まり，それらを平均した値を λ と σ に対する $\widehat{J}(\boldsymbol{\alpha})$ の値とする．次に λ と σ を変化させ，上記手順で得られる $\widehat{J}(\boldsymbol{\alpha})$ の値が最小となる $\hat{\lambda}$ と $\hat{\sigma}$ を求め，これを λ と σ の推定値とする．

3.2 線形カーネルの利用

本論文では，uLSIF の基底関数にガウスカーネルではなく線形カーネルを利用することを提案する．

$$\psi_l(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_l^t) = \boldsymbol{x} \cdot \boldsymbol{x}_l^t$$

表 1 対象単語

単語	辞書上の語義数	OC での頻度	OC での語義数	PB での頻度	PB での語義数	PN での頻度	PN での語義数
言う	3	666	2	1114	2	363	2
入れる	3	73	2	56	3	32	2
書く	2	99	2	62	2	27	2
聞く	3	124	2	123	2	52	2
子供	2	77	2	93	2	29	2
時間	4	53	2	74	2	59	2
自分	2	128	2	308	2	71	2
出る	3	131	3	152	3	89	3
取る	8	61	7	81	7	43	7
場合	2	126	2	137	2	73	2
入る	3	68	4	118	4	65	3
前	3	105	3	160	2	106	4
見る	6	262	5	273	6	87	3
持つ	4	62	4	153	3	59	3
やる	5	117	3	156	4	27	2
ゆく	2	219	2	133	2	27	2
平均	3.44	148.19	2.94	199.56	3.00	75.56	2.69

一般に，高次元への非線形空間への写像はガウスカーネルなどが利用されるが，本論文で扱うタスクでは次元数に比べて事例数が極端に小さいために，高次元に写影する必要がなく線形カーネルで十分である．

また事例数が少ないために，uLSIF で必要となるパラメータ b も $b = \min(100, N_t)$ ではなく，$b = N_t$ として問題ない．

3.3 特殊事例のみへの重み付け

語義曖昧性解消の領域適応では，確率密度比のモデルを対象単語毎に構築する必要がある．また対象単語の用例はソース領域，ターゲット領域の両方において，その次元数に比べると非常に少ない．このため推定された確率密度比は真の値よりも小さくなる傾向がある．このため推定された確率密度比を 1 に近づけることが提案されている．杉山は確率密度比 r に p $(0 < p < 1)$ 乗した r^p を重みにすることを提案している [19]．また Yamada は relative density ratio として確率密度比を以下の相対確率密度比で求めることを提案してる [16]

$$\frac{P_T(\boldsymbol{x})}{\alpha P_T(\boldsymbol{x}) + (1-\alpha)P_S(\boldsymbol{x})}.$$

ここではほとんどの事例の重み 1 を 1.1 にし，推定された確率密度比が極端に小さいものに 0.1，極端に大きいものに 2.1 を与えることを提案する．

「極端に小さいあるいは大きい」の判定は以下の手順に従った．まず得られている確率密度比の集合を $W = \{w_i\}_{i=1}^N$ とする．W の平均を μ，分散を σ^2 とし，w_i を以下で正規化する．

$$w_i' = \frac{w_i - \mu}{\sigma}$$

W を正規分布と仮定し，w_i' が上位 20% の分位点 0.84 よりも大きな場合に w_i を 2.1 とし，下位 20% の分位点 -0.84 よりも小さな場合に w_i を 0.1 とし，それ以外の場合に w_i を 1.1 とする．

4. SVM による重み付き学習

共変量シフト下の学習では確率密度比を重みとした重み付き学習を行う．通常はロジスティック回帰や最大エントロピー法が用いられるが，損失関数ベースの学習法であれば利用することができる．

ここでは不均衡データに対する SVM の手法 [15] を利用する．訓練データを $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ $(\boldsymbol{x}_i \in R^d, y_i \in \{1, -1\})$ とするとき，SVM は通常，以下の式からパラメータ $\boldsymbol{w}, b, \boldsymbol{\zeta}$ を求めて識別器を学習する．

$$\min_{\boldsymbol{w},b,\boldsymbol{\zeta}} \left\{ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^N \zeta_i \right\} \tag{8}$$

ここで

$$y_i(\boldsymbol{w}^T\phi(x_i) + b) \geq 1 - \zeta_i\ , \quad \zeta_i \geq 0$$

である．上記の式で $\boldsymbol{x}_i$ に対して C の代わりに $w(\boldsymbol{x}_i)C$ を用いることで，重み付き学習が可能となる [2]．

5. 実験

BCCWJ の PB(書籍)，OC(Yahoo! 知恵袋) 及び PN (新

聞) を異なった領域として実験を行う．SemEval-2 の日本語 WSD タスク [9] ではこれら領域のコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．この 3 つの領域からある程度頻度のある多義語 16 単語を WSD の対象単語とする．これら単語と辞書上での語義数及び各コーパスでの頻度と語義数を表 2 に示す[*1]．領域適応の方向としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りの方向が存在する．

本実験で利用した素性は以下の 8 種類である．(e0) w の表記，(e1) w の品詞，(e2) w_{-1} の表記，(e3) w_{-1} の品詞，(e4) w_1 の表記，(e5) w_1 の品詞，(e6) w の前後 3 単語までの自立語の表記，(e7) e6 の分類語彙表の番号の 4 桁と 5 桁．なお対象単語の直前の単語を w_{-1}，直後の単語を w_1 としている．

対象単語 w についてソース領域 S からターゲット領域 T への領域適応の実験について説明する．ソース領域 S の訓練データのみを用いて，手法 A により分類器を学習し w に対する正解率を求める．16 種類の各対象単語 $(w_1, w_2, \cdots, w_{16})$ に対する正解率の平均，つまりマクロ平均をソース領域 S からターゲット領域 T に対する手法 A の正解率とする．結果，手法 A について 6 種類の各領域適応に対しての正解率が得られる．それらの平均を手法 A の平均正解率とする．

本論文で扱う手法 A は 2 つの要素から構成される．確率密度比の算出法と重み付き学習の種別である．以下のような組み合わせが得られる．

表 2 重み付けと学習手法

手法	重み付け	学習法
Base-M	1	ME
Base-S	1	SVM
Mtd-G-M	Gauss	ME
Mtd-G-S	Gauss	SVM
Mtd-L-M	Linear	ME
Mtd-L-S	Linear	SVM
Ours-G-M	Gauss → 0.1, 1.1, 2.1	ME
Ours-G-S	Gauss → 0.1, 1.1, 2.1	SVM
Ours-L-M	Linear → 0.1, 1.1, 2.1	ME
Ours-L-S	Linear → 0.1, 1.1, 2.1	SVM

Base-M は重み付けなしで学習法に最大エントロピー法を用いたもの．Base-S は重み付けなしで学習法に SVM を用いたもの．Mtd-G-M はガウスカーネルを利用した uLSIF により確率密度比を推定し，重み付き学習には最大エントロピー法を用いたもの，Mtd-G-S は学習法に SVM を用いたもの．Mtd-L-M は線形カーネルを利用した uLSIF により確率密度比を推定し，重み付き学習には最大エントロピー法を用いたもの，Mtd-L-S は学習法に SVM を用いたもの．Ours-G-M はガウスカーネルを利用した uLSIF により確率密度比を推定し，それを提案手法により 0.1, 1.1 および 2.1 に変換し，重み付き学習には最大エントロピー法を用いたもの，Ours-G-S は学習法に SVM を用いたもの．Ours-G-M は線形カーネルを利用した uLSIF により確率密度比を推定し，それを提案手法により 0.1, 1.1 および 2.1 に変換し，重み付き学習には最大エントロピー法を用いたもの，Ours-L-S は学習法に SVM を用いたものである．本論文の提案手法は Ours-L-S である．

実験の結果を表 3 に示す．`Base-M < Base-S`，`Mtd-G-M < Mtd-G-S`，`Mtd-L-M < Mtd-L-S`，`Our-G-M < Our-G-S`，`Our-L-M < Our-L-S` が成立しているので，最大エントロピー法を用いるよりも SVM を用いる方が効果があることがわかる（図 1 参照）．また `Mtd-G-M < Mtd-L-M`，`Mtd-G-S = Mtd-L-S`，`Ours-G-S < Mtd-L-S` が成立し，`Ours-G-M` と `Ours-L-M` はそれぞれ 0.7160 と 0.7159 でありほぼ等しい．このため uLSIF ではガウスカーネルを利用するよりも線形カーネルを利用する方が効果があることがわかる（図 2 参照）．さらに本論文の提案手法 `Ours-L-S` は最も高い平均正解率である．また各領域適応においても，`PN` → `PB` を除いて最も高い正解率を示した．

6. 考察

6.1 重みのスケール

最大エントロピー法における事例の重みは，そのスケールとは無関係である．最大エントロピー法では重み付き対数尤度を最大化するパラメータを求めるため，全ての事例の重みを定数倍しても，最大化する式に変化がないからである．一方，SVM における事例の重みは式 (8) の C の値であるために，重みのスケールの問題がある．通常，uLSIF で求まる確率密度比は 10^{-2} 程度のスケールになっているために，ここではその 100 倍の値を重みとして利用した．

100 倍という値は，この問題に特有の値であるために，問題に応じてこの値を変えられるようにする方が適切である．ここでは以下の方法を試みた．

(a) 重みの集合の平均を m とし，重み x を x/m に変更する．
(b) 重みの集合の最小値を m とし，重み x を x/m に変更する．
(c) 重みの集合の中央値を m とし，重み x を x/m に変更する．

(a) (b) あるいは (c) を行った場合の Mtd-L-S と Mtd-G-S を試した結果を表 4 に示す．表の *-100 は重みを一律

[*1] 語義は岩波国語辞書がもとになっている．そこでの中分類までを対象にした．また「入る」は辞書上の語義が 3 つだが，OC や PB では 4 つの語義がある．これは SemEval-2 の日本語 WSD タスクでは新語義も許しているからである．

表 3　各手法の平均正解率

	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC	平均正解率
Base-M	0.7163	0.7700	0.6920	0.6778	**0.7474**	0.6991	0.7171
Base-S	0.7141	0.7676	0.6907	0.6880	0.7452	0.7011	0.7178
Mtd-G-M	0.7008	0.7289	0.6854	0.6840	0.7110	0.6760	0.6977
Mtd-G-S	0.7143	0.7692	0.6903	0.6900	0.7455	0.7034	0.7189
Mtd-L-M	0.7145	0.7339	0.6907	0.6887	0.7144	0.7008	0.7055
Mtd-L-S	0.7134	0.7699	0.6905	0.6898	0.7450	0.7045	0.7189
Ours-G-M	0.7145	0.7670	0.6907	0.6787	0.7446	0.7008	0.7160
Ours-G-S	0.7129	0.7707	0.6911	0.6884	0.7451	0.7021	0.7184
Ours-L-M	0.7145	0.7665	0.6907	0.6787	0.7445	0.7008	0.7159
Ours-L-S (提案手法)	**0.7197**	**0.7723**	**0.6971**	**0.6936**	0.7416	**0.7062**	**0.7218**

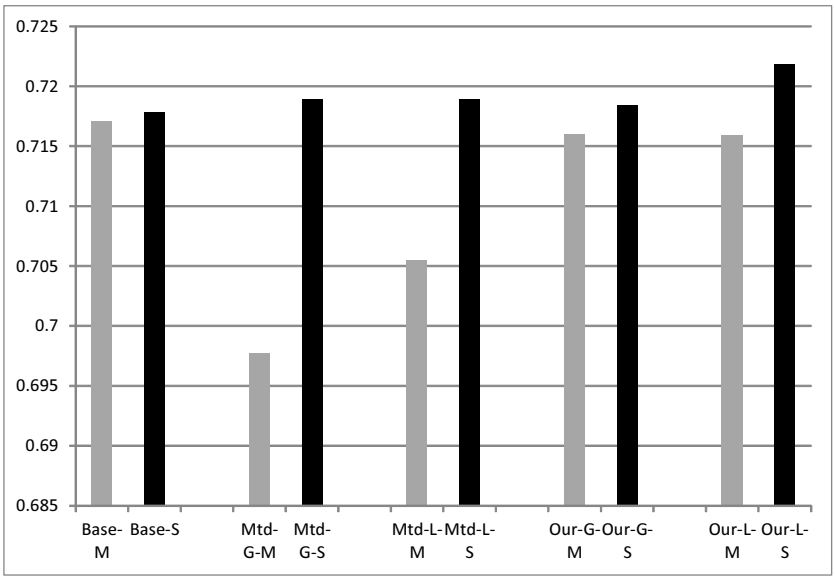

図 1　最大エントロピー法と SVM の比較

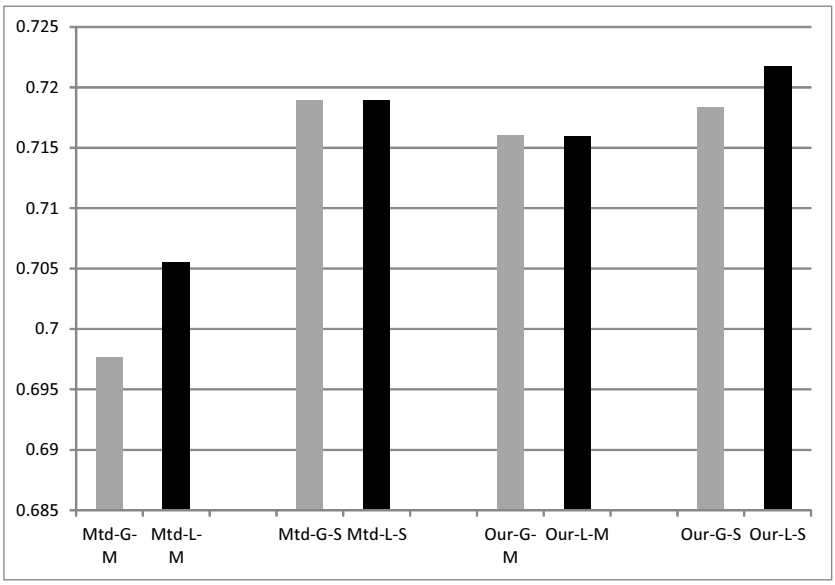

図 2　ガウスカーネルと線形カーネルの比較

に 100 倍した本論文の手法であり，*-mean は (a)，*-min は (b) そして *-mid は (c) による重みに対応する．

どの手法を用いても大きな違いは見られなかったが，本論文で行ったように単純に 100 倍する方法が最も安定していた．

表 4 重みのスケール

	平均正解率		平均正解率
Mtd-G-S-100	0.7189	Mtd-L-S-100	0.7189
Mtd-G-S-mean	0.7178	Mtd-L-S-mean	0.7192
Mtd-G-S-min	0.7178	Mtd-L-S-min	0.7178
Mtd-G-S-mid	0.7179	Mtd-L-S-mid	0.7176

6.2 片側だけの重みの適用

提案手法は uLSIF で求めた重みが極端に小さい場合と極端に大きな場合にその値を 0.1 と 2.1 にし，それ以外を 1.1 にしている．これを変更し，(a) 重みが極端に小さい場合だけその値を 0.1 としそれ以外を 1.1 にする，あるいは (b) 重みが極端に大きい場合だけその値を 2.1 としそれ以外を 1.1 にする，という片側だけの重みの調整も考えられる．

ここではこの片側だけの重みの調整の実験を行った．結果を表 5 に示す．表の Ours-L-S-small は上記の (a) のケースに相当し，Ours-L-S-large は上記の (b) のケースに相当する．

重要な事例のみ重みを大きくした場合，ベースの平均正解率より悪くなるため，重要な事例の重みを大きくする効果よりも，重要でない事例の重みを小さくする効果の方が高いことがわかる．

表 5 片側だけの重みの適用

	平均正解率
Base-S	0.7178
Ours-L-S	0.7218
Ours-L-S-small (小さい重みだけ適用)	0.7183
Ours-L-S-large (大きい重みだけ適用)	0.7176

6.3 Misleading データの削除

本タスクの領域適応では，重要でない事例の重みを小さくする効果の方が高い．これは訓練データ中に Misleading データが存在するからだと考えられる．Misleading とは訓練データの中で，識別器の精度を悪化させるようなデータであり，領域適応ではしばしば現れる [5]．

ここではまず Misleading の存在を確認する．そのために論文 [17] で行ったように，しらみつぶしに Misleading を見つけ出す．領域 S から領域 T の領域適応において，対象単語 w の S 上のラベル付きデータ D が存在する．まず D で学習した識別器の T に対する正解率 p_0 を測る．次に D から 1 つデータ x を取り除き，$D-\{x\}$ から学習した識別器の T に対する正解率 p_1 を測る．$p_1 > p_0$ となった場合，データ x を Misleading データと見なす．これを D 内のすべてのデータに対して行い，S から T の領域適応における対象単語 w の Misleading データを見つける．この処理によって見つけ出された Misleading データの個数を表 6 示す．括弧内の数値は全データ数である．

求めた Misleading データを除いてから SVM により識別した結果を表 7 に示す．表の Mislead がそれに当たる．本論文の実験で得られている平均正解率よりもかなり高い．Misleading データを除く学習は Misleading データの重みを 0 とした重み付き学習と見なせる．つまり重みの設定のみでもベースライン (Base-S) の正解率 0.7178 を 0.7542 まで改善可能である．

前節での実験では，重要でない事例の重みのみを小さくすることを行ったが，重要でない事例の重みを 0，つまりそのデータを Misleading データとして除いた結果を示す．表 7 の Mislead2 がそれに当たる．この場合，効果はなかった．確率密度比だけでは適切に Misleading データを検出できていないからである．今後は適切に Misleading データを検出する方法を考えたい．

7. おわりに

本論文では語義曖昧性解消の教師なし領域適応の問題に対して，共変量シフト下の学習を試みた．確率密度比 $w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ の算出に，$P_T(\boldsymbol{x})$ と $P_S(\boldsymbol{x})$ を推定しその比を求める手法でなく，直接 $w(\boldsymbol{x})$ をモデル化した手法である uLSIF を用いた．またその際，通常のガウスカーネルではなく線形カーネルを利用した．また重み付き学習には通常使われる最大エントロピー法ではなく SVM を利用した．BCCWJ の 3 つの領域のコーパスと 16 単語を利用した領域適応の実験を行い，線形カーネルの利用と重み付き SVM の利用の効果を示した．さらに極端に小さいあるいは大きい重みを持つデータだけに重みを反映した手法も提案した．この提案手法は，先の実験で最も高い平均正解率を出し，その有効性が示された．今後は Misleading の検出方法を考案し，教師なし領域適応の問題に対処したい．

参考文献

[1] Blitzer, J., McDonald, R. and Pereira, F.: Domain adaptation with structural correspondence learning, *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128 (2006).
[2] Cortes, C. and Vapnik, V.: Support-vector networks, *Machine learning*, Vol. 20, No. 3, pp. 273–297 (1995).
[3] Daumé III, Hal: Frustratingly Easy Domain Adaptation, *ACL-2007*, pp. 256–263 (2007).
[4] Glorot, X., Bordes, A. and Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learn-

表 6 Misleading データの個数

単語	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC
言う	159 (666)	75 (1114)	82 (363)	158 (666)	35 (363)	127 (1114)
入れる	6 (73)	15 (56)	3 (32)	28 (73)	1 (32)	19 (56)
書く	21 (99)	2 (62)	12 (27)	39 (99)	15 (27)	0 (62)
聞く	26 (124)	0 (123)	4 (52)	21 (124)	27 (52)	26 (123)
子供	5 (77)	1 (93)	12 (29)	0 (77)	13 (29)	12 (93)
時間	1 (53)	0 (74)	0 (59)	8 (53)	5 (59)	0 (74)
自分	13 (128)	0 (308)	0 (71)	25 (128)	1 (71)	0 (308)
出る	14 (131)	32 (152)	22 (89)	10 (131)	10 (89)	39 (152)
取る	6 (61)	18 (81)	12 (43)	5 (61)	22 (43)	10 (81)
場合	0 (126)	13 (137)	14 (73)	0 (126)	9 (73)	7 (137)
入る	36 (68)	27 (118)	27 (65)	11 (68)	42 (65)	38 (118)
前	8 (105)	1 (160)	15 (106)	5 (105)	2 (106)	10 (160)
見る	10 (262)	12 (273)	8 (87)	3 (262)	28 (87)	3 (273)
持つ	8 (62)	11 (153)	1 (59)	0 (62)	1 (59)	2 (153)
やる	0 (117)	0 (156)	0 (27)	0 (117)	0 (27)	0 (156)
ゆく	17 (219)	1 (133)	3 (27)	0 (219)	3 (27)	15 (133)

表 7 Misleading データの削除

	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC	平均正解率
Base-S	0.7141	0.7676	0.6907	0.6880	0.7452	0.7011	0.7178
Ours-L-S	0.7197	0.7723	0.6971	0.6936	0.7416	0.7062	0.7218
Mislead	0.7459	0.7927	0.7450	0.7213	0.7869	0.7334	0.7542
Mislead2	0.7117	0.7627	0.6833	0.6920	0.7399	0.6984	0.7146

ing approach, *ICML-11*, pp. 513–520 (2011).
[5] Jiang, J. and Zhai, C.: Instance weighting for domain adaptation in NLP, *ACL-2007*, pp. 264–271 (2007).
[6] Kanamori, T., Hido, S. and Sugiyama, M.: A least-squares approach to direct importance estimation, *The Journal of Machine Learning Research*, Vol. 10, pp. 1391–1445 (2009).
[7] Maekawa, K.: Design of a Balanced Corpus of Contemporary Written Japanese, *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58 (2007).
[8] Mori, S.: Domain Adaptation in Natural Language Processing (in Japanese), *The Japanese Society for Artificial Intelligence*, Vol. 27, No. 4, pp. 365–372 (2012).
[9] Okumura, M., Shirai, K., Komiya, K. and Yokono, H.: SemEval-2010 Task: Japanese WSD, *The 5th International Workshop on Semantic Evaluation*, pp. 69–74 (2010).
[10] Pan, S. J., Kwok, J. T. and Yang, Q.: Transfer Learning via Dimensionality Reduction, *AAAI*, Vol. 8, pp. 677–682 (2008).
[11] Pan, S. J., Tsang, I. W., Kwok, J. T. and Yang, Q.: Domain adaptation via transfer component analysis, *Neural Networks, IEEE Transactions on*, Vol. 22, No. 2, pp. 199–210 (2011).
[12] Pan, S. J. and Yang, Q.: A survey on transfer learning, *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 22, No. 10, pp. 1345–1359 (2010).
[13] Sogaard, A.: *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*, Morgan & Claypool (2013).
[14] Sugiyama, M. and Kawanabe, M.: *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation* , MIT Press (2011).
[15] Tang, Y., Zhang, Y.-Q., Chawla, N. V. and Krasser, S.: SVMs modeling for highly imbalanced classification, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, Vol. 39, No. 1, pp. 281–288 (2009).
[16] Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H. and Sugiyama, M.: Relative density-ratio estimation for robust distribution comparison, *Neural Computation*, Vol. 25, No. 5, pp. 1370–1370 (2011).
[17] 吉田拓夢, 新納浩幸：外れ値検出手法を利用した Misleading データの検出，第 5 回コーパス日本語学ワークショップ，pp. 49–56 (2014).
[18] 古宮嘉那子，小谷善行，奥村学：語義曖昧性解消の領域適応のための訓練事例集合の選択，言語処理学会第 19 回年次大会，pp. C6–2 (2013).
[19] 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， Vol. 13, No. 3, pp. 111–118 (2006).
[20] 山田誠：私のブックマーク　確率密度比に基づく機械学習，人工知能学会誌，Vol. 27, No. 4, pp. 449–452 (2012).
[21] 齋木陽介，高村大也，奥村学：文の感情極性判定における事例重み付けによるドメイン適応 (情報抽出・評判分析)，情報処理学会研究報告. 自然言語処理研究会報告，Vol. 2008, No. 33, pp. 61–67 (2008).