ベイズ規則による確率密度比の推定を用いた語義曖昧性解消の領域適応

菊池 裕紀 † 新納 浩幸 ‡ 佐々木 稔 ‡ 古宮 嘉那子 ‡
† 茨城大学大学院理工学研究科
‡ 茨城大学工学部情報工学科
{13nm705g@hcs, shinnou@mx, msasaki@mx, kkomiya@mx}.ibaraki.ac.jp

1 はじめに

本論文では, 語義曖昧性解消（Word Sense Disambiguation, WSD）の教師なし領域適応をタスクに設定し，新納ら [7] によって提案されたベイズ規則を用いて推定した確率密度比を重みとした重み付け学習がSVM でも機能するどうかを調べることを目的としている．分類器に SVM を利用した時に，WSD の領域適応の問題をソース領域の訓練データに対して事例ごとに重みを付ける重み付け学習を利用して解決していく．

自然言語処理の多くのタスクでは，訓練データとテストデータのコーパスが属する領域が異なる領域適応の問題が生じている．領域が異なる場合，訓練データから得られた学習規則がテストデータに上手く適応できない問題が出てくる．このような領域適応の問題を解決する手法は，大きく分けて事例ベースと素性ベースの手法が存在する．素性ベースの手法では，ソース領域の素性空間をターゲット領域に合うように拡張する [1][2]．事例ベースの手法では，訓練事例に対して重み付け学習を行う．本論文では，事例ベースの手法を用いる．

事例ベースの手法ではソース領域とターゲット領域の間に共変量シフト [6] を仮定することが多い．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．共変量シフトでは $P_S(c|\boldsymbol{x}) = P_T(c|\boldsymbol{x})$ だが $P_S(\boldsymbol{x}) \neq P_T(\boldsymbol{x})$ と問題を仮定する．これはある領域で出現した事例が他の領域で出現しても，その事例が示す意味は変わらないとすることを意味する．自然言語処理のタスクでは，これは通常は成立している仮定である．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て $P_T(c|\boldsymbol{x})$ を構築するアプローチが取られる．確率密度比 $w(\boldsymbol{x})$ の算出手法は，$P_T(\boldsymbol{x})$ と $P_S(\boldsymbol{x})$ をそれぞれモデル化して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．前者の研究として新納らによって行われたベイズ規則を用いて確率密度比を算出する手法が挙げられる [7]．新納らは，$P_T(\boldsymbol{x})$ と $P_S(\boldsymbol{x})$ をベイズ規則を用いて求め，その比を確率密度比として用いている．また先の論文では，WSD において推定される確率密度比の値が低くなってしまう問題に対して，$P_S(\boldsymbol{x})$ を求める際に，ソース領域とターゲット領域のデータを合わせたものを新たなソース領域のデータ S とみなすアプローチをとっている．本論文では，この手法が重み付き SVM で上手く機能するかを調べることとする．

実験には現代書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese, BCCWJ[4]）における 3 つの領域である Yahoo!知恵袋（OC），書籍（PB），新聞（PN）を利用する．SemEval-2 の日本語 WSD タスクでは，これらのコーパスの一部に語義タグを付けたデータが公開されており，そのデータを使う．すべての領域適応である程度頻度のある多義語 16 単語を対象として，WSD の領域適応を行う．領域適応の種類は，${}_3C_2$ の 6 通りである．つまり，16×6 より合計 96 通りの実験を行う．その結果，ベイズ規則によって算出した確率密度比を重みとした重み付き SVM の効果を確認することができた．

2 共変量シフト下における領域適応

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 と識別した時の損失関数を $l(\boldsymbol{x}, c, d)$ 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，本タスクにおける損失関数 L_0 は以下で表すことができる．

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

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．

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

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．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 を最小化する d を求めればよい．

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

ここで分類器として 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ていく．

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

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とる場合，共変量シフトの仮定の下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．

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

3 確率密度比の算出

確率密度比 $w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ の算出の手法は大きく 2 つに分類できる．1 つは $P_T(\boldsymbol{x})$ と $P_S(\boldsymbol{x})$ をそれぞれ推定してその比を取る方法，もう 1 つは $w(\boldsymbol{x})$ を直接モデル化する方法である [3]．ここでは前者の手法として論文 [7] で提案された手法を利用する．

3.1 ベイズ規則による算出

対象単語 w の用例 $\boldsymbol{x}$ の素性リストを $\{f_1, f_2, \cdots, f_n\}$ とする．求めたいのは領域 $R \in \{S,T\}$ 上の $\boldsymbol{x}$ の分布 $P_R(\boldsymbol{x})$ である．ここで Naive Bayes で使われるモデルを用いている．Naive Bayes のモデルでは以下を仮定する．

$$P_R(\boldsymbol{x}) = \prod_{i=1}^n P_R(f_i)$$

領域 R のコーパス内の w の全ての用例について素性リストを作成する．ここで用例数を $N(R)$ ，全ての用例中で素性 f_i が現れた用例数を $n(R,f_i)$ とおく．MAP 推定でスムージングを行い，$P_R(f_i)$ を以下で定義する．

$$P_R(f_i) = \frac{n(R,f_i)+1}{N(R)+2}$$

以上より，ソース領域 S の用例 $\boldsymbol{x}$ に対して，確率密度比 $w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ を計算する．

$$\begin{aligned} w(\boldsymbol{x}) &= \frac{P_T(\boldsymbol{x})}{P_S(\boldsymbol{x})} \\ &= \prod_{i=1}^n \left(\frac{n(T,f_i)+1}{N(T)+2} \cdot \frac{N(S)+2}{n(S,f_i)+1} \right) \end{aligned}$$

3.2 $P_S(\boldsymbol{x})$ の補正

WSD のタスクでは算出される確率密度比が小さい値を取る傾向があり，実際に重みとして適用する際には多少上方修正した値を採用したほうが識別結果が改善されることが多い．この理由には，S には $\boldsymbol{x}$ が必ず入っているが T に入っているかは確率的であることと，$P_S(\boldsymbol{x})$ を推定する際に $\boldsymbol{x} \in S$ を用いるため，訓練データに過学習した結果 $P_S(\boldsymbol{x})$ が $P_T(\boldsymbol{x})$ に比べて高く見積もられてしまうことの 2 点が考えられる．

このため，新納らは確率密度比を上方修正するために，ソース領域とターゲット領域のデータを合わせた

ものを新たなソース領域のデータとみなしてベイズ規則によって $P_S(\boldsymbol{x})$ を補正している．この手法は，確率密度 $P_S(\boldsymbol{x})$ が真の値よりも低く見積もられる原因が S のスパース性からくるものだと考え，そのスパース性を S にデータを加えることで解消する手法である．ただしこのとき，追加するデータは S と似ているデータが望ましいとされる．この点は，WSD のタスクでは扱う領域は類似してることがほとんどであるため満たされるとしている．

つまり，新たなソース領域のデータを $S+T$ で表すとすると，$P_{S+T}(\boldsymbol{x})$ を以下のベイズ規則を利用した式で算出することとなる．

$$\begin{aligned} P_{S+T}(f_i) &= \frac{n(S+T,f_i)+1}{N(S+T)+2} \\ &= \frac{n(S,f_i)+n(T,f_i)+1}{N(S)+N(T)+2} \end{aligned}$$

上記の $P_{S+T}(f_i)$ を利用して，$w(\boldsymbol{x})$ を新たに以下で定義する．

$$\begin{aligned} w(\boldsymbol{x}) &= \frac{P_T(\boldsymbol{x})}{P_{S+T}(\boldsymbol{x})} \\ &= \prod_{i=1}^{n}\left(\frac{n(T,f_i)+1}{N(T)+2}\cdot\frac{N(S)+N(T)+2}{n(S,f_i)+n(T,f_i)+1}\right) \end{aligned}$$

4 SVM による重み付け学習

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，損失関数ベースの手法であれば重み付け学習を行うことができる．

ここでは，不均衡データに対する SVM の手法 [5] 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．

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

ここで

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

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

5 実験

実験には BCCWJ コーパスの Yahoo!知恵袋 (OC), 書籍 (PB), 新聞 (PN) を異なる領域として使用する．SemEval-2 の日本語 WSD タスクでは，これらの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 の対象単語に設定する．領域適応の種類は，OC → PB，OC → PN，PB → OC，PB → PN，PN → OC，PN → PB の 6 種類である．

本実験では 8 種類の素性を利用している．(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．ここでは対象単語の直前の単語を $w_{(-1)}$，直後の単語を $w_{(1)}$ で表している．

ここで，対象単語 w に関するソース領域 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．分類器の学習には SVM を用いる．その際，重みとして与える確率密度比の算出方法により比較する手法を分類する．Base は重みを考慮しない（重みを 1 で固定）手法，NB はベイズ規則による重みを付けた手法，NB(S+T) はソース領域とターゲット領域を合わせたデータを新たなソース領域のデータとしてベイズ規則により確率密度比を算出し重み付けする手法とする．学習した分類器により w に対する正解率を求める．16 種類の対象単語 $w_1,w_2,\cdots,w_{16}$ に対する正解率の平均をソース領域からターゲット領域に対する各手法の正解率とする．つまり，各手法に対して 6 種類の領域適応の正解率が求まることとなる．それらの平均を各手法の平均正解率とする．また本論文では分類器の学習に scikit-learn で提供されている SVM[1] を利用した．カーネルは線形カーネルを利用した．結果を表 1 に示す．

結果から分かるように，算出した確率密度比を用いた SVM での重み付け学習の効果が確認できた．平均正解率は NB(S+T) が一番良い値を示している．NB の場合，Base の平均正解率よりも下がってしまっている．単純にベイズ規則により確率密度比を算出し，それを重みとして SVM での学習に適用しただけでは効果がないことが分かる．

[1] http://scikit-learn.org/stable/modules/svm.html

表 1: 実験結果

	OC → PB	OC → PN	PB → OC	PB → PN	PN → OC	PN → PB	平均正解率
Base	0.7172	0.7006	0.7008	0.7173	0.7123	0.7174	0.7109
NB	0.7039	0.6900	0.6893	0.7128	0.7022	0.6928	0.6985
NB(S+T)	**0.7175**	**0.7026**	**0.7025**	**0.7198**	**0.722**1	**0.7253**	**0.7150**

6 考察

論文 [7] では，$P_S(\boldsymbol{x})$ を下方修正し，確率密度比を上方修正する手法であった．ここで，$P_T(\boldsymbol{x})$ の推定時に補正をかけることで確率密度比の上方修正を試みてみる．$P_T(\boldsymbol{x})$ を補正する式を以下の式 2 に示す．

$$P_T(f_i) = \frac{n(T, f_i) + 2}{N(T) + 4} \tag{2}$$

ここでは，スムージングの値を $P_S(\boldsymbol{x})$ よりも高くすることで $P_T(\boldsymbol{x})$ の補正を行っている．T-NB は NB における $P_T(\boldsymbol{x})$ を補正して推定した確率密度比を重みとして適用する手法，T-NB(S+T) は NB(S+T) における $P_T(\boldsymbol{x})$ を補正して推定した確率密度比を重みとして適用する手法である．結果を表 2 に示す．

表 2: $P_T(\boldsymbol{x})$ を補正した結果

	平均正解率
Base	0.7109
NB	0.6985
NB(S+T)	**0.7150**
T-NB	0.7123
T-NB(S+T)	0.7110

実験結果から $P_S(\boldsymbol{x})$ に補正をかけた NB(S+T) よりも平均正解率が向上することはなかった．ソース領域とターゲット領域の両方の確率密度に補正をかけた場合，その平均正解率は Base の値とほぼ変わらない．しかし，NB(S+T) の次に良い精度を示しているのは T-NB である．どちらも Base の平均正解率よりも良い値を示している．つまり，$P_S(\boldsymbol{x})$ もしくは $P_T(\boldsymbol{x})$ のどちらか一方を補正することで精度の向上が見込めることがいえる．

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，$P_S(\boldsymbol{x})$ と $P_T(\boldsymbol{x})$ をそれぞれベイズ規則を用いて推定し、その比を取る手法を採用した．またソース領域とターゲット領域のデータを合わせたものを新たなソース領域のデータとして用いることで値の上方修正を図る手法により確率密度比を求め，算出された確率密度比を SVM に重みとして適用した際の効果を示した．

また考察において，$P_T(\boldsymbol{x})$ の補正に関してスムージングの値を $P_S(\boldsymbol{x})$ よりも高くすること利用して試みたが，効果はなかった．しかし，ベースとなる平均正解率よりも良い精度を示していることから $P_S(\boldsymbol{x})$ もしくは $P_T(\boldsymbol{x})$ のどちらかを適当に補正することができればある程度の効果を得られることが分かった．今後は，より適切に確率密度比を補正するための手法を考案することが課題となってくる．

参考文献

[1] Daumé III, Hal. Frustratingly Easy Domain Adaptation. In *ACL-2007*, pp. 256–263, 2007.

[2] Daumé III, Hal. Frustratingly Easy Semi-Supervised Domain Adaptation. In *ACL-2010*, p. 2359, 2010.

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

[4] Kikuo Maekawa. Design of a Balanced Corpus of Contemporary Written Japanese. In *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58, 2007.

[5] Y. Tang. Svms modeling for highly imbalanced classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, Vol. 39, No. 1, pp. 281–288, 2009.

[6] 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.

[7] 新納浩幸, 佐々木稔. 共変量シフト下における語義曖昧性解消の教師なし領域適応. 自然言語処理, Vol. 21, No. 5, pp. 1011–1035, 2014.