素性に重みを付ける Self-training 手法を用いた文書分類の領域適応

國井 慎也　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
{13nm707s@hcs, shinnou@mx, msasaki@mx, kkomiya@mx}.ibaraki.ac.jp

1 はじめに

自然言語処理の多くのタスクで教師付きの機械学習手法が利用されるが、そこには領域適応の問題が存在する [1] [2]。領域適応の問題とは、学習の際に訓練データとして利用するソースデータの領域と、学習により得られた分類器を適用する先のターゲットデータの領域が異なる問題であり、近年活発に研究が行われている。

一般に、領域適応の手法は事例ベースの手法と素性ベースの手法に分けられる [3]。事例ベースの手法とは訓練事例に重みをつけて学習する手法であり、共変量シフト下の学習が代表的研究である [4]。共変量シフトとは $P_S(\boldsymbol{x}) \neq P_T(\boldsymbol{x})$ であるが、$P_S(y|\boldsymbol{x}) = P_T(y|\boldsymbol{x})$ という仮定である。共変量シフト下の学習は期待損失最小化から確率密度比 $P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$ を重みとした損失ベースの学習に帰着される。一方、素性ベースの手法とはソース領域の素性空間とターゲット領域の素性空間を共通の素性空間に写影する手法である。概略、領域間の違いを減少させ、本来の領域の重要な性質を保持するような次元縮約法となる。Blitzer はソース領域とターゲット領域の両方に頻出する素性を Pivot Features と呼び、それを使って Structural Correspondence Learning (SCL) と呼ばれる次元縮約法を提案した [5]。Pan はソース領域を写影した空間とターゲット領域を写影した空間との距離を Maximum Mean Discrepancy (MMD) によって評価し、これを最小にするような MMD Embedding (MMDE) と呼ばれる変換法を提案した [6]。さらに Sinno は MMDE を改良した Transfer Component Analysis (TCA) と呼ばれる手法を提案した [7]。また素性に重みをつけて学習させる手法も、素性ベースの手法の一種である。Daumé は簡易な素性の重み付け手法 [8] を提案した。そこではソース領域の訓練データのベクトル $\boldsymbol{x}_s$ を $(\boldsymbol{x}_s, \boldsymbol{x}_s, \boldsymbol{0})$ と連結した 3 倍の長さのベクトルに直し，ターゲット領域の訓練データのベクトル $\boldsymbol{x}_t$ を $(\boldsymbol{0}, \boldsymbol{x}_t, \boldsymbol{x}_t)$ と連結した 3 倍の長さのベクトルに直す．この 3 倍にしたベクトルを用いて，通常の分類問題として解く．この手法はソース領域とターゲット領域に共通している素性が重なることで，共通している素性に重みをつけている形になる。

一方、領域適応の問題は、訓練データのスパース性の問題とも見なせる。このため、Self-Training、半教師付き学習 [9] あるいは能動学習も領域適応 [10][11] に利用できる。特に Self-Training はターゲット領域のデータにラベル付けを必要としない教師なし領域適応が可能であるため有用性が高い。ここでは Self-Training を利用する。

Chen は、領域適応に Self-Training を利用する際に、領域間の違いを考慮して素性に重みをつけることで学習の効果を高める手法を提案した [12]。本論文では Chen が提案した重み付け手法を改良した新たな重み付け手法を提案する。提案した重み付け手法による Self-Training を利用することで領域適応の問題を解決する。

実験では 20 Newsgroups data set[1] から 6 つのカテゴリの文書群を取り出し、文書分類の領域適応を行った。

2 重み付け Self-training 手法

2.1 Self-training

Self-Training は、ターゲット領域のデータにラベル付けを必要としない教師なし領域適応の手法として利用できる。今、ラベル付きのデータを L、ラベルなしデータを U とする。初期の状態では、L はソース領域のラベル付きデータであり、U はターゲット領域のラベルなしデータである。

Self-Training の各ステップでは、現在の訓練データ L から分類器 h が学習され、それをラベルなしデータ U に適用する。このとき U の各データにラベルに対

[1] http://qwone.com/~jason/20Newsgroups/

する確信度を与える。確信度が高い上位 c 個のデータに、h によって識別されたラベルを付与し、c 個のデータを L に追加する。このステップを U が空になるか、全てのデータの確信度がある閾値以下になったときに終了する。

Self-Training は確信度の与え方と c の設定が必要である。ここでは学習アルゴリズムとして SVM を利用する。SVM により得られた分離超平面までの距離によって確信度を与える。また $c = 10$ とする。

2.2 従来手法

素性への重みとして、Chen は素性とクラスラベルとの相関係数を利用した [12]。

データ $\boldsymbol{x}$ の素性 f の値を x_f、データ $\boldsymbol{x}$ のクラスを y_x とする。ソース領域のラベル付きデータに対して x_f と y_x の相関係数を $\rho_S(x_f, y_x)$ とおく。ターゲット領域のデータ $\boldsymbol{x}$ については、そのクラスが不明であるが、その時点の分類器で識別されたクラス y'_x を代用することで、x_f と y'_x の相関係数 $\rho_T(x_f, y'_x)$ が得られる。

そして素性 f の重み係数 $w(f)$ を以下で定義する。

$$w(f) = \frac{1 + \rho_S(x_f, y_x)\rho_T(x_f, y'_x)}{2} \quad (1)$$

$w(f)$ を利用して、x_f は以下のように更新される。

$$x_f \leftarrow x_f + \gamma_n w(f)$$

ここで n は Self-Training のステップ数を表し、γ_n は $\gamma_n \to 0$ を満たす。具体的にここでは γ_n を以下で定義した。

$$\gamma_n = 0.01 \cdot \frac{N - n}{N}$$

ここで N は Self-Training のステップ数の最大値である。

また各素性 f に対して x_f を更新した後に $\boldsymbol{x}$ の大きさを 1 に正規化する。

2.3 提案手法

本論文ではクラス分布に対する素性の重みの分布を利用した新たな重み付け法を提案する。この重みは、ソース領域とターゲット領域の両領域において、ある素性が同じような使われ方をしているなら、大きくし、逆に異なる使われ方をしている場合は小さくするように設定される。従来手法は多クラス分類のタスクにおいては拡張が出来ないため、提案手法では多クラス分類においても適応できることを目指す。

本論文では以下の手続きによって $w(f)$ を定義する。

$$\hat{i} = \arg\max_i P_S(x, y_i)$$

$$\hat{j} = \arg\max_j P_T(x, y_j)$$

もし、$\hat{i}$ と $\hat{j}$ が等しいならば、

$$w(f) = \max(P_S(x, y_{\hat{i}}), P_T(x, y_{\hat{j}})$$

等しくないならば、

$$w(f) = 0$$

と定義する。

ここでソース領域 S のデータ $\boldsymbol{x}$ の中で、クラスが i となっているデータ $\boldsymbol{x}$ の集合を S_i と書くことにし、P_S を以下で定義する。

$$P_S(x, y_i) = \frac{\sum_{\boldsymbol{x} in S_i} x_f}{\sum_{\boldsymbol{x} in S} x_f}$$

ターゲット領域 T のデータ $\boldsymbol{x}$ に関しては、クラスが未知である。そのためその時点での分類器によって与えられたクラスを $\boldsymbol{x}$ のクラスと考え、T の中でクラスが i となっているデータ $\boldsymbol{x}$ の集合を T_i と書くことにし、P_T を以下で定義する。

$$P_T(x, y_i) = \frac{\sum_{\boldsymbol{x} in T_i} x_f}{\sum_{\boldsymbol{x} in T} x_f}$$

3 実験

3.1 文書分類の領域適応

本論文では文書分類の領域適応の問題を扱う。文書分類とは与えられた文書がカテゴリ A のものかカテゴリ B のものかを識別するタスクである。通常、扱う文書群のセット S が領域 A の文書群と領域 B の文書群からなっており、領域 A と領域 B の文書群からそれぞれ少量の文書集合 D_A と D_B を取り出し、それらを訓練データとして、分類器 h を作成することで文書分類が解決できる。ただし現実の文書分類では、文書群のセットが S から T に変化する。具体的には領域 A が領域 C へ、領域 B が領域 D へ変化する。ここで領域 A と領域 C 及び領域 B と領域 D は類似しているために、分類器 h は文書群 T においてもある程度は有効に機能するが、文書群セット S における精度は得られない。文書分類の領域適応とは分類器 h を文書群セット T に合うように調整することである。また文書分類の領域適応では領域 A と領域 B の組をソース領域 S と考え、領域 C と領域 D の組をターゲット領域 T と考える。

3.2 実験データ

実験では 20 Newsgroups data set[2]から以下の 6 つのカテゴリの文書群を取り出した。

```
A: comp.sys.ibm.pc.hardware
B: rec.sport.baseball
C: sci.electronics
D: comp.sys.mac.hardware
E: rec.sport.hockey
F: sci.med
```

実験は 3 クラスの多クラス分類を行う。ソース領域 S を (A,B,C)、ターゲット領域 T を (D,E,F) とした領域適応と、ソース領域 S を (D,E,F)、ターゲット領域 T を (A,B,C) とした領域適応の 2 つの文書分類の領域適応の実験を行う。

各文書群の文書数（データ数）を表 1 に示す。

表 1: 各領域のデータ数

	Labeled data	Unlabeled data	Test data
A	100	578	(300,200,100)
B	100	590	(300,200,100)
C	100	594	(300,200,100)
D	100	557	(300,200,100)
E	100	595	(300,200,100)
F	100	587	(300,200,100)

通常の領域適応のタスクにおいてはソース領域とターゲット領域でのクラス分布が異なることが多い。そのためターゲット領域のテストデータの比率を変更し、実験を行う。例えば、ソース領域 S を (A,B,C)、ターゲット領域 T を (D,E,F) とした場合、ターゲット領域の T のテストデータの文書数を (D,E,F) =(300,200,100),(300,100,200),(200,300,100),(200,100,300),(100,300,200),(100,200,300) のように変更して実験を行い、各正解率の平均を各手法の評価値とする。

3.3 実験結果

ソース領域 S を (A,B,C)、ターゲット領域 T を (D,E,F) とした領域適応の結果を図 1 に示す。またソース領域 S を (D,E,F)、ターゲット領域 T を (A,B,C) とした領域適応の結果を図 2 に示す。

図 1、図 2 ともに、横軸は Self-Training において unlabeled data を追加した数、縦軸はターゲット領域

[2]http://qwone.com/~jason/20Newsgroups/

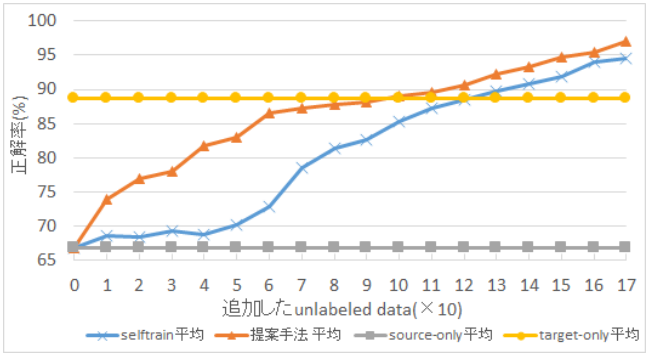

図 1: 提案手法による領域適合の実験

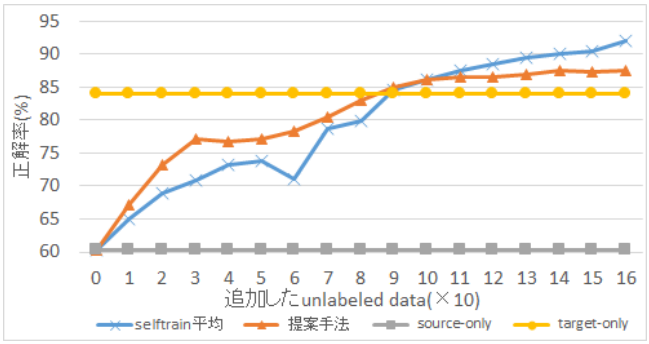

図 2: 提案手法による領域適合の実験

の Labeled data をテストデータとした分類器の正解率の平均である。また直線の source-only はソース領域の Labeled data のみから学習した分類器の正解率の平均であり、target-only はターゲット領域の Labeled data のみから学習した分類器の正解率の平均である。

図 1 では、提案手法が既存手法である self-training よりも上回っている。図 2 では、学習の初期段階、つまり、unlabeled data の追加数が少ない場合は提案手法が既存手法よりも上回っているが、学習の後半になるにつれ提案手法の効果が減少していることがわかる。最終的には提案手法は領域適応において学習の初期段階において特に効果があることがわかる。

4 考察

4.1 テストデータにおけるクラス分布の影響

図 1、図 2 では、テストデータの比率を変更した 6 つのデータセットを利用し、それらの正解率の平均を評価値としていた。ここでは平均を取らず、その 6 つのデータセットのそれぞれの正解率を図 3、図 4 に示す。手法は Self-training である。

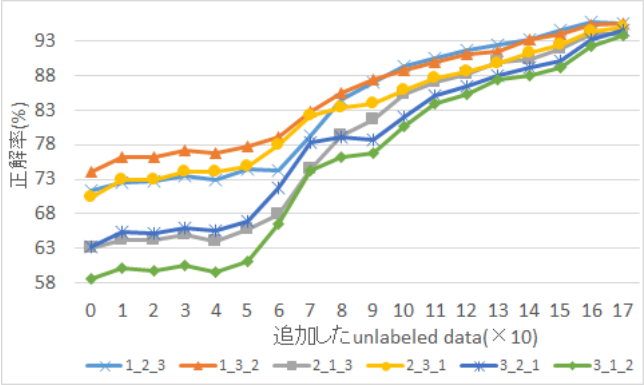

図 3: テストデータにおけるクラス分布を変更したときの実験

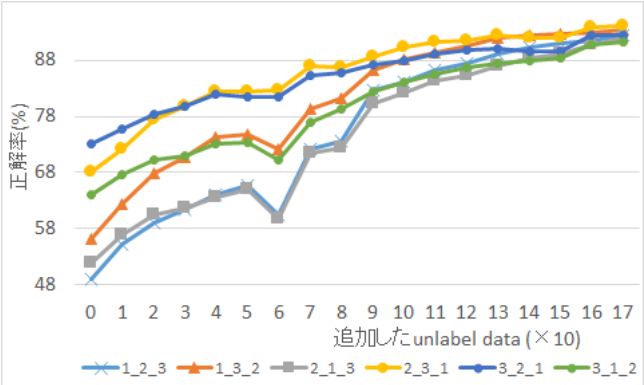

図 4: テストデータにおけるクラス分布を変更したときの実験

図 3 は S を (A,B,C)、ターゲット領域 T を (D,E,F) とした領域適応の結果であり、図 4 は S を (D,E,F)、ターゲット領域 T を (A,B,C) とした領域適応の結果である。また、ターゲット領域のテストデータのクラス分布の比率を $(1:2:3), (1:3:2), (2:1:3), (2:3:1), (3:1:2), (3:2:1)$ のように変更している。Self-training の初期段階、つまり学習データが少ないときは、テストデータのクラス分布の影響があり、正解率が図 3 では最大で 15 %、4 では最大で 20 %程度の差が見受けられる。また、Self-training の学習の後半になると、各正解率の差は減少していき、ある点に向かって収束していくことがわかる。この実験結果から、学習データの量によってテストデータのクラス分布の影響の度合いが変わることがわかる。そのため、十分に学習データがあれば、ターゲット領域のクラス分布を推定する必要はないと考えられる。

5 おわりに

本論文では文書分類の教師なし領域適応の問題に対して、素性に重みをつける Self-training の手法を利用した。従来、素性に重みをつけるために、ソース領域およびターゲット領域における素性とクラスとの相関係数を求め、それを基に重みを算出していた。ここではソース領域およびターゲット領域におけるクラス分布に対する素性の重みの分布を利用した多クラス分類に拡張できる新たな重み付け法を提案した。

20 Newsgroups data set の 6 つのカテゴリ文書 群（comp.sys.ibm.pc.hardware,rec.sport.baseball, sci.electronics,comp.sys.mac.hardware,rec.sport.hockey, sci.med）を利用した文書分類の領域適応の実験を行い、提案手法を評価した。

参考文献

[1] Shinsuke Mori. Domain adaptation in natural language processing (in japanese). *The Japanese Society for Artificial Intelligence*, Vol. 27, No. 4, pp. 365–372, 2012.

[2] Anders Sogaard. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool, 2013.

[3] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 22, No. 10, pp. 1345–1359, 2010.

[4] Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation* . MIT Press, 2011.

[5] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128, 2006.

[6] Sinno Jialin Pan, James T Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *AAAI*, Vol. 8, pp. 677–682, 2008.

[7] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, Vol. 22, No. 2, pp. 199–210, 2011.

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

[9] Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. *Semi-supervised learning*, Vol. 2. MIT press Cambridge, 2006.

[10] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[11] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pp. 27–32, 2010.

[12] Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *NIPS*, pp. 2456–2464, 2011.