半教師あり学習と素性の重み付け学習の交互適用による文書分類の領域適応

新納 浩幸　古宮 嘉那子　佐々木 稔
茨城大学 工学部 情報工学科
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp,
{kanako.komiya.nlp, minoru.sasaki.01}@vc.ibaraki.ac.jp

1 はじめに

本論文では文書分類の領域適応の問題に対して，半教師あり学習と素性の重み付け学習を交互に適用させることで，分類精度を徐々に向上させていく手法を提案する．

自然言語処理の多くのタスクにおいて，教師あり学習は大きな成功を収めている．ただし教師あり学習を現実の問題に利用する場合，領域適応の問題が生じることが多い．一般に，教師あり学習ではラベル付きの訓練データから SVM などの学習アルゴリズムを用いて分類器を作成し，その分類器を用いてテストデータのラベルを識別する．この際，訓練データとテストデータの領域が異なる問題が領域適応の問題である [10]．

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 [8]．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 [11] が代表的研究である．共変量シフトとは $P_S(y|\boldsymbol{x}) = P_T(y|\boldsymbol{x})$ であるが，$P_S(\boldsymbol{x}) \neq P_T(\boldsymbol{x})$ 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 [1]．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を提案した [4]．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る形になる．

これらの領域適応適応の手法はうまく機能することも多いが，領域間の違いが小さいときは，このような手法を利用することが逆効果になることもある．領域間の違いが小さいときは，領域適応の問題は単にデータスパースネスの問題と捉えた方が現実的である．その場合は，従来の半教師あり学習 [2] や能動学習 [9] の手法がそのまま利用できる．

本論文では文書分類の領域適応の問題を扱う．ここで扱う領域適応では領域間の違いが小さいために，前述したように，半教師あり学習が利用できる．特に文書分類の半教師あり学習としては，Naive Bayes 法を基本に EM アルゴリズムを用いる手法 [7]（本論文ではこの手法を NBEM と呼ぶ）が効果的であることが知られており，ここでも NBEM を用いる．ただし NBEM は訓練データとテストデータの領域の違いを利用していないために改良の余地がある．ここでは Chen が示した素性の重み付け学習を改良した手法 [3]（本論文ではこの手法を STFW と呼ぶ）を利用する．Chen が示した手法はターゲット領域上の素性のラベル分布を推定するのに自己学習を用いるが，STFW では NBEM を用いる．更に Chen の手法の重み付けは２値分類の問題にしか適用できないが，我々の STFW では多値分類における重み付けを行えるようにする．最後に，NBEM と STFW を交互に利用することで，文書分類の領域適応の精度を徐々に向上させていく．

実験では 20 Newsgroups のデータ[1]を利用して領域 A の文書分類のタスクと領域 B の文書分類のタスクを構築し，領域 A から領域 B への領域適応と領域 B から領域 A への領域適応の実験を行った．結果，提案手法の有効性を確認できた．

2 NBEM と STFW の交互適用

2.1 NBEM

NBEM は少量のラベル付き訓練データと大量のラベルなしデータから分類器を学習する半教師あり学習の１つである．概略，ラベル付き訓練データから Naive Bayes の分類器を学習し，その分類器を大量のラベルなしデータと EM アルゴリズムを用いて改善していく手法である．

まず Naive Bayes 法を示す．$C = \{c_1, c_2, \cdots, c_m\}$ をラベルの集合，$\boldsymbol{x} = (f_1, f_2, \cdots, f_n)$ を事例（データ）とする．$\boldsymbol{x}$ のラベル c_x はベイズの定理から以下で推定できる．

$$c_x = arg\max_{c \in C} P(c)P(\boldsymbol{x}|c).$$

ここで Naive Bayes では $P(\boldsymbol{x}|c) = \prod_{i=1}^{n} P(f_i|c)$ を仮定する．これにより $P(c)P(\boldsymbol{x}|c)$ の計算が可能なり，識別が可能になる．

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

次に Naive Bayes をベースにした EM アルゴリズム [7] を示す．ポイントは以下の式により $P(f_i|c_j)$ の推定である．

$$P(f_i|c_j) = \frac{1 + \sum_{k=1}^{|D|} N(f_i, d_k) P(c_j|d_k)}{|F| + \sum_{m=1}^{|F|} \sum_{k=1}^{|D|} N(f_m, d_k) P(c_j|d_k)}. \quad (1)$$

D: ラベル付きデータとラベル無しデータを合わせたデータの集合
d_k: D 内のデータ
F: 全素性の集合
f_m: F 内の素性
$N(f_i, d_k)$: d_k 内の f_i の頻度

式 1 を利用して，以下の分類器が構築できる．

$$P(c_j|d_i) = \frac{P(c_j) \prod_{f_n \in K_{d_i}} P(f_n|c_j)}{\sum_{r=1}^{|C|} P(c_r) \prod_{f_n \in K_{d_i}} P(f_n|c_r)}. \quad (2)$$

ここで K_{d_i} は d_i 内の素性の集合である．また $P(c_j)$ を以下で与える．

$$P(c_j) = \frac{1 + \sum_{k=1}^{|D|} P(c_j|d_k)}{|C| + |D|}. \quad (3)$$

EM アルゴリズムでは式 2 を利用して $P(c_j|d_i)$ を計算し (E-step)，次に式 1 を利用して $P(f_i|c_j)$ を計算する (M-step)．$P(f_i|c_j)$ と $P(c_j|d_i)$ が収束するまで E-step と M-step を繰り返す．ここでは現状の $P(f_i|c_j)$ と次のステップでの $P(f_i|c_j)$ との差が $8 \cdot 10^{-6}$ 以下になるか，繰り返しが 10 回に達するかで収束の判定を行った．

2.2 STFW

本論文では素性に重みを付ける手法として Chen が提案した手法を改良する．この手法を Self-Training Feature Weight，略して STFW と呼ぶことにする．

領域適応では素性ベースの手法が効果的である．本質的には，素性ベースの手法はソース領域の空間とターゲット領域の空間を共通の素性空間に写影する手法と見なせる．操作的には素性に重みを付けていることに対応するので，直感的には，ソース領域とターゲット領域の両領域において識別に効果のある素性に重みを付ける手法とも見なせる．

Chen は以下のような方法で素性に重みを付けた．まずデータ $\boldsymbol{x}$ 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，そのクラスを自己学習により推定したクラス y'_x で代用し，x_f と y'_x の相関係数 $\rho_T(x_f, y'_x)$ を得る．そして素性 f の重み $w(f)$ を以下で定義した．

$$w(f) = \frac{1 + \rho_S(x_f, y_x)\rho_T(x_f, y'_x)}{2} \quad (4)$$

新たな素性の値は以下により更新される．

$$v_{t+1} = (1 + w(f))v_t$$

Chen の手法は重みの定義に相関係数 $\rho_S(x_f, y_x)$ と $\rho_T(x_f, y'_x)$ を利用している．ラベルがカテゴリカルな値なので，実質，2 値分類しか対象にできない．ここでは Chen の手法を基本として，多値分類にも対応した重みつけの定義を行う．Chen の重みは，ソース領域の素性 f のラベル分布 P_s とターゲット領域の素性 f のラベル分布 P_t の類似性を測っていると見なせる．そこで本論文はまず P_s と P_t の距離 $d(f)$ を以下で定義した．

$$d(f) = |P_s - P_t| \quad (5)$$

次に $d(f)$ を用いて重みを設定する．ただし，ここでのタスクは文書分類であり，学習アルゴリズムとして Naive Bayes を用いるので，素性の値は頻度となる．そのため重みを与えた後の素性の値も 0 以上の整数値であるのが望ましい．このことから $d(f) < \theta_1$ のときは素性 f の値を 1 プラスし，$d(f) > \theta_2$ のときは素性 f の値を 1 マイナスする．ただしマイナスすることで負の数になる場合は，その値は 0 とする．また本論文では $\theta_1 = 0.2$ と $\theta_2 = 1.5$ とした[2]．

またターゲット領域のデータにはラベルがないために P_t は未知である．Chen は自己学習を利用してターゲット領域のデータにラベルを付け，その信頼度の高いデータだけを対象にして P_t を推定している．本論文でも自己学習を用いるが NBEM から学習された分類器を用いる．また信頼度の高いものだけに限定せず，すべてのデータを利用して P_t を推定する．

2.3 NBEM と STFW の交互適用

本論文では NBEM と STFW を交互に適用する手法を提案する．提案手法は以下の手順で行われる．

(1) ソース領域のラベル付き訓練データ L_S とターゲット領域のラベルなしデータ U_T に対して NBEM を用いて分類器を学習する．

(2) この分類器を利用して U_T のラベルを推定する．

(3) 推定したラベルを利用して，STFW により L_s の素性に重みを付け，新たな訓練データ L'_S を構築する．

(4) L'_s を L_S に設定し (1) に戻る

i 回目の繰り返しの (1) から作成される分類器を NBEM+STFW-(i) と表記する．ここでは上記の繰り返しを 20 回行い NBEM+STFW-(1) から NBEM+STFW-(20) を作成する．また NBEM+STFW-(1) は NBEM と同じであることに注意する．

3 実験

実験では 20 Newsgroups data set[3]から以下の 6 つのカテゴリの文書群を取り出した．括弧内の記号はクラス名を意味する．

[2] これらの閾値はクラス数に依存する．本論文の実験は全てクラス数が 3 であることを考慮して，これらの値を設定している．
[3] http://qwone.com/~jason/20Newsgroups/

表 1: 実験結果 (正解率 %)

	NB (S-only)	NBEM	NBEM+STFW-(20)	NB (T-only)
X → Y	72.83	90.00	93.33	94.67
Y → X	81.17	82.67	89.17	90.00

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

(A,B,C) のデータセットを領域 X, (D,E,F) のデータセットを領域 Y とする．各領域は C = {comp, rec, sci}をクラスラベルセットとする文書分類ののデータセットとなっている．本論文では領域 X から領域 Y 及び領域 Y から領域 X の文書分類の領域適応の問題を扱う．

各文書群の文書数（データ数）を表 2 に示す．どちらの領域でもラベル付き訓練データのクラス分布は一様だが，現実の問題に合うようにテストデータのクラス分布が領域毎に異なるように設定した．

領域 X から領域 Y の領域適応では，A, B, C のラベル付きデータが訓練データ (計 300 文書) となり，D, E, F のラベルなしデータが利用できるラベルなしデータ (計 900 文書) である．そして D, E, F のテストデータがテストデータ (計 600 文書) となる．逆に領域 Y から領域 X の領域適応では，D, E, F のラベル付きデータが訓練データ (計 300 文書) になり，A, B, C のラベルなしデータが利用できるラベルなしデータ (計 900 文書) である．そして A, B, C のテストデータがテストデータ (計 600 文書) となる．

表 2: 各領域のデータ数

	Labeled data	Unlabeled data	Test data
A	100	400	300
B	100	300	200
C	100	200	100
D	100	200	100
E	100	400	300
F	100	300	200

実験の結果を表 1 に示す．

NB (S-Only) の列にはソース領域の訓練データのみから Naive Bayes で分類器を学習し，テストデータを識別した正解率が記されている．NBEM の列は訓練データとラベルなしデータを用いた NBEM による正解率，NBEM+STFW-(2) の列は STFW を最初に適用した後の NBEM により構築された分類器の正解率，そして NBEM+STFW-(20) の列は提案手法により得られた最終的な分類器の正解率である．表 1 より提案手法の効果が確認できる．また参考としてターゲット領域の訓練データのみから Naive Bayes で分類器を学習した場合の正解率を NB (T-Only) に示した．この値が通常の領域適応の問題が生じていない場合の教師あり学習の正解率を示している．

また NBEM+STFW-(i) の正解率の変化を図 1（X → Y）と図 2（Y → X）に示す．大まかに 20 回の繰り返しでほぼ上限の値に達していると見なせる．

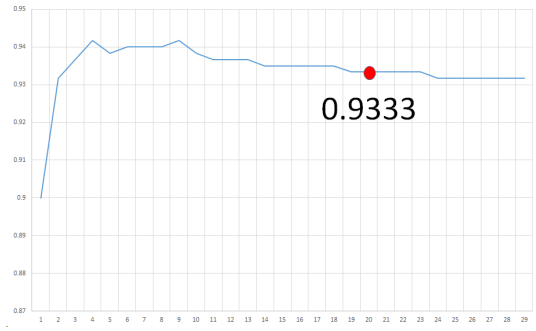

図 1: NBEM+STFW-(i) の正解率 (X → Y)

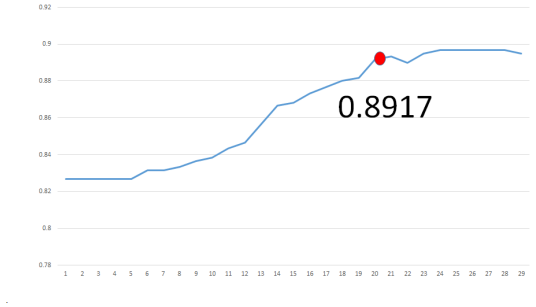

図 2: NBEM+STFW-(i) の正解率 (Y → X)

4 考察

4.1 Transductive-SVM との比較

分類器の精度を向上させるためにラベルなしデータを利用する手法は半教師あり学習の他に Transductive Learning も存在する．そして Transductive Learning の代表的手法として，Transductive-SVM (TSVM) がある [5].

本論文では半教師あり学習の NBEM を利用したが，NBEM ではなく TSVM を利用することも可能である．実験のデータに対して TSVM を用いた結果を表 3 に示す．

表 3: TSVM との比較

	NB	NBEM	SVM	TSVM
X → Y	72.83	90.00	75.83	66.50
Y → X	81.17	82.67	71.16	70.83

一般には SVM は NB よりも精度が高いが，文書分類のタスクに限ると NB の方が精度が高い場合もある．実際に Y → X の領域適応では NB の方がよかった．文書分類のタスクに NB を用いる場合，文書は単に Bag of words で表現すれば良いが，SVM では加工が必要なるためだと考える．上記の SVM を利用した実験では TF*IDF によりベクトル値を修正し，最後にベクトルのサイズを 1 に正規化する処理を行っている．

また TSVM は SVM の精度を改善せず，逆に精度を下げてしまっている．これは TSVM では訓練データのクラス分布とテストデータのクラス分布が同じであることを仮定しており，本実験のデータではその仮定が崩れているからだと思われる．

4.2 他の領域適応の手法との比較

領域適応の手法は素性ベースの手法と事例ベースの手法に大別できる．本節では素性ベースの手法と事例ベースの手法を本タスクに適用し，本論文の提案手法と比較する．

素性ベースの手法としては，素性ベースの代表的手法といえる SCL を利用する．また事例ベースの代表的手法は共変量シフト下の学習である．共変量シフト下の学習では確率密度比の算出がキーとなるが，ここでは Unconstrained Least Squares Importance Fitting (uLSIF) [6] という密度算出法を利用する．

実験の結果を表 4 に示す．表中の NBEM+STFW-(20) が提案手法であり，SCL が素性ベースの手法の SCL，uLSIF が事例ベースの手法の uLSIF を意味する．

表 4: 他の領域適応手法との比較

	NBEM+STFW-(20)	SVM	SCL	uLSIF
X → Y	93.33	75.83	74.33	73.67
Y → X	89.17	71.16	71.83	72.17

SCL も uLSIF もベースの SVM の結果とほとんど変化はなく，提案手法の方が圧倒的に精度が高い．これはベースの学習アルゴリズムが SVM と NB であるという違いが大きかったと考えられる．本タスクに限れば NB の方が SVM よりも精度が高い．また SCL も uLSIF もトランスダクティブな手法であり，学習の過程で，ターゲット領域のテストデータは利用するが，ラベルなしデータを利用していないことも影響していると考えられる．提案手法は学習の過程で，ラベルなしデータを利用している．本実験ではラベルなしデータの量がテストデータの量の 1.5 倍であり，ラベルなしのデータの量の多い方が有利であったためと考えられる．

4.3 パラメータの調整

本論文の SFTW では領域適応において識別に有効そうな素性に重みを与え，識別に悪影響を与えそうな素性の重みを減じた．この重みの与え方や大きさには，いくつかのバリエーションが考えられる．

紙面の都合上，結果だけ述べるが，パラメータの θ_1 や θ_2 の値，及び重みの大きさより精度が変化した．適切なパラメータの設定方法が今後の課題といえる．

5 おわりに

本論文では文書分類の領域適応の問題に対して，半教師あり学習の NBEM と領域適応の重み付き学習である STFW を交互に繰り返し適用する手法を提案した．概略，NBEM を利用して分類器を学習し，その分類器を利用して STFW から訓練データに重みをつけて再構築し，その再構築された訓練データを使ってこの手順を繰り返す．20 Newsgroups の一部のデータを利用して実験した結果，本手法の効果を確認できた．本手法ではいくつかのパラメータがあるが，それらパラメータの適切な設定方法が今後の課題である．

参考文献

[1] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning.

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

[3] Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation.

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

[5] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, Vol. 99, pp. 200–209, 1999.

[6] 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.

[7] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134, 2000.

[8] 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.

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

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

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