領域適応のためのサポートベクトルを用いた訓練事例の反復的選択

小林　優稀（茨城大学工学部　情報工学科）
古宮　嘉那子（茨城大学工学部 情報工学科）
佐々木　稔（茨城大学工学部 情報工学科）
新納　浩幸（茨城大学工学部 情報工学科）
奥村　学（東京工業大学 精密工学研究所）

Iterative Selection of Training Data Using Support Vectors for Domain Adaptation

Yuma Kobayashi (Department of Computer and Information Sciences, Ibaraki University)
Kanako Komiya (Department of Computer and Information Sciences, Ibaraki University)
Minoru Sasaki (Department of Computer and Information Sciences, Ibaraki University)
Hiroyuki Shinnou (Department of Computer and Information Sciences, Ibaraki University)
Manabu Okumura (Precision and Intelligence Laboratory, Tokyo Institution of Technology)

要旨

テストの対象となるドメインではなく、異なるドメインのデータ（ソースデータ）で学習を行い、それをターゲットのドメインのデータ（ターゲットデータ）に適応することを領域適応といい、近年様々な手法が研究されている。

語義曖昧性解消のタスクについて領域適応を行った場合、ソースデータ全体を学習に用いるよりも、確信度と LOO-bound という指標を利用して、自動的に選択したソースデータの部分集合を用いたほうが、正解率が上昇することが先行研究により指摘されている。本稿では、自動的に選択したソースデータの部分集合にさらにサポートベクトルを利用して反復的にソースデータを追加することを繰り返す、という手法を試みた。その結果、ベースラインよりも正解率は劣るものの、それほど正解率を落とさずに、訓練事例の数を大幅に減らすことに成功した。

１．はじめに

テストの対象となるドメインではなく、異なるドメインのデータ（ソースデータ）で学習を行い、それをターゲットのドメインのデータ（ターゲットデータ）に適応することを領域適応といい、近年様々な手法が研究されている。

語義曖昧性解消のタスクについて領域適応を行った場合、ソースデータ全体を学習に用いるよりも、確信度と LOO-bound という指標を利用して、自動的に選択したソースデータの部分集合を用いたほうが、正解率が上昇することが先行研究により指摘されている（古宮, 小谷, 奥村(2013)）。本稿では、自動的に選択したソースデータの部分集合にさらにサポートベクトルを利用して反復的にソースデータを追加することを繰り返す、という手法を試みた。

２．関連研究

領域適応は、学習に使用する情報により、supervised，semi-supervised，unsupervised の三種に分けられる。本研究で扱うのは、semi-supervised の領域適応、つまりラベルつきのソースデータとラベルなしのターゲットデータを利用するものである。

文献(Komiya, Okumura (2012))、(古宮, 奥村, 小谷(2013))では、訓練データの選択に分類器の確信度を用いて訓練事例を自動的に選択している。用例ごとに訓練事例を自動的に選択している。

また、文献(古宮 小谷 奥村(2013))は、semi-supervised な領域適応において、あるターゲットデータに対して複数のジャンルのソースデータが混在した場合、確信度と

LOO-bound という指標を利用して、領域適応のための訓練事例の部分集合を WSD の対象単語タイプごとに自動的に選択する手法について述べている。訓練データをいくつかのグループに分け分類器を作り、分類した時の各分類器の確信度と、SVM に対し、leave-one-out-estimation を行った場合の期待値の上限である LOO-bound という指標を用いて、訓練データを選択する手法である。この研究では、確信度と LOO-bound を組み合わせたスコアを用いることで、ベースラインよりも精度が向上することを報告している。本稿でも、確信度と LOO-bound を利用した、このスコアを利用する。また、先行研究と同じくラベルなしターゲットデータが手に入ると仮定して、語義曖昧性解消についての領域適応を行った。

２．１確信度と LOO-bound

本稿では、分類器のスコアとして確信度と LOO-bound をもとにした数値を掛け合わせたスコアを使用している。

確信度とは、テストデータに対し、どの程度自信を持って分類したのかを表す。つまり、テストデータと同じドメインのコーパスをどの程度正確に分類できるかを示している。確信度は用例ごとに算出されるので、全用例の平均を分類器のスコアとした。

LOO-bound は SVM に対し、Leave-One-Out-Estimation を行った時のエラーの期待値の上限であり、サポートベクトルの数を訓練事例の数で割った値である。この値はエラー率であるため、分類器のスコアとする際に 1 からこの値を引いた。

$$\text{LOO} - \text{bound のスコア} = 1 - \frac{\text{サポートベクトルの数}}{\text{訓練事例の数}} \quad \cdots (1)$$

３．領域適応のためのサポートベクトルを用いた訓練事例の反復的選択

あるドメインのターゲットデータに対して WSD を行う。このターゲットデータのラベルは未知とする。ソースデータとして複数ドメインのコーパスが利用可能であるとし、ソースデータの全体集合から、ターゲットデータに適した訓練事例を自動的に選択することを試みる。以下で、具体的な手順を示す。

(1) ソースデータの全体集合から訓練事例をランダムに選択して、訓練事例集合を複数個作成する。

(2) それぞれの訓練事例集合で分類器を学習し、ターゲットデータに適用する。

(3) 分類器が出力する値をもとに分類器ごとにスコアを計算する。

(4) スコアの最も高い分類器を作成した訓練事例集合を選択する。

SVM では分離平面を決定する際に、サポートベクトルからの距離を最大にするという性質がある。そこで、サポートベクトルを残し、反復的に訓練事例を増加させるために、以下の処理を追加した。

(5) 選択した訓練事例集合のサポートベクターの集合（SV 集合）を作成する。

(6) SV 集合にソースデータの全体集合から訓練事例をランダムに選択して加え、訓練事例集合を複数個作成する。

(7) 有限回、(2)～(6) を繰り返す。

４．実験

４．１データセット

実験には、マルチクラス対応の分類器として SVM(libsvm)(Chih-Chung Chang, Chih-Jen Lin(2001)) を使用した。また、現代日本語書き言葉均衡コーパス(Maekawa(2008))の YAHOO！知恵袋（OC）、白書(OW)、YAHOO！ブログ(OY)、新聞(PN)、書籍(PB)、雑誌(PM)のコアデータ 6 種と YAHOO！知恵袋(YAHOO)、白書(BCCWJ)非コアデータ 2 種、RWC コーパス(Hashida, Isahara, Tokunaga, Hashimoto, Ogino, and Kashino(1998))を用いた。YAHOO 知恵袋と白書のコーパスは２種あるが、内容はほぼ同一のものなので、より用例数が少なかったコアデータの方をソースデータから除いた。

また、ソースデータにテストデータのドメインと同一のドメインのコーパスを含まないようにした。テストデータには１単語あたり 50 用例以上のものを使用した。コーパスごとの単語数とデータ数の平均値を表１に示す。

また、実験には岩波国語辞典の中分類の語義を採用した。単語の語義は、岩波国語辞典(西尾、 岩淵、 水谷 (1994))の小分類の語義を採用した。語義事の単語の内訳は、1 語義（新語義を入れると 2 語義）：可能、2 語義：生きる、一般、生まれる、書く、考える、技術、経済、現在、現場、子供、自分、情報、高い、作る、強い、電話、場合、早い・速い、文化、ほか、見せる、3 語義：相手、与える、言う、今、入れる、大きい、教える、買う、関係、聞く、市場、市民、社会、進む、地方、出来る、出る、入る、初め・始め、始める、場所、開く、前、求める、訴える、4 語義：時間、時代、出す、乗る、計る、一つ、見える、認める、持つ、進める、5 語義：やる、良い、6 語義：合う・会う、立つ・建つ、見る、もの、7 語義：手、8 語義：する、取る、上げるであった。

また、本実験で使用する素性として、次の 24 の素性を使用した。

- 対象単語と前後２つの形態素の表記　5 種類
- 対象単語と前後２つの形態素の品詞　5 種類
- 対象単語と前後２つの形態素の品詞の細分化　5 種類
- 係り受け　1 種類
- 前後２つの形態素の５桁の分類コード　4 種類
- 前後２つの形態素の４桁の分類コード　4 種類

ここで用いている分類コードとは国立国語研究所が発行している「分類語彙表」（秀英出版 (1964))に記載されている分類番号、段落番号からなる、語を意味によって分類した番号のことである。

４．２．ベースライン

本実験のベースラインとして、以下の３つの実験を行った。

・すべてのコーパス

利用できるコーパス全てを使用する

・最大のコーパス

利用できるコーパスのうち、単語ごとに用例数が最大のものを使用する

・平均的なコーパス

利用できるコーパスについて、それぞれ分類器を作成し、正解率を平均する

４．３．サポートベクトルを用いた反復的手法実験

提案手法は次の手順で行う。

(1) ソースデータの全体集合から訓練事例をすべての語義を含むようにランダムに 100 件もしくは 200 件（データ件数がこの数に満たない際にはそれ以下の件数となる）選択して、訓練事例集合を 10 個作成する

(2) それぞれの訓練事例集合で分類器を学習し、ターゲットデータに適用する

(3) 分類器が出力する値をもとに分類器ごとにスコアを計算する

(4) スコアの最も高い分類器を作成した訓練事例集合を選択する

(5) 選択した訓練事例集合のサポートベクターの集合（SV 集合）を作成する

(6) SV 集合にソースデータの全体集合から訓練事例をランダムに選択して加え、訓練事例集合を複数個作成する

(7) 10 回（10 ステージ）、(2)～(6) を繰り返す

訓練事例の部分集合は 1 単語あたり 10 個作成した。また、初期事例数を 100 件または 200 件とし、すべての語義を含むようにランダムに選択した。　予備実験の結果、繰り返し回数は 10 回程度でスコアはほぼ収束することが分かったので、本実験では(7)の繰り返し回数は 10 回とする。また、この実験はランダム性が高いので、10 セット行いそれぞれの正解率を平均した。その他、前者ではすべての語義を含むように初期訓練事例集合を作成しているが、語義数にかかわらずランダムに 100 件選択したものを用いた実験も 2 回行なった。

表 1　コーパスの単語数の内訳

	単語数	テストデータ数 平均	ソースデータ数 平均
コア　Yahoo！知恵袋	22	157.77	1630.50
コア　白書	5	79.20	508.80
コア　Yahoo！ブログ	9	245.22	10226.56
コア　書籍	35	158.91	6068.54
コア　雑誌	26	7408.00	5806.08
コア　新聞	25	92.28	7586.60
非コア　白書	38	2069.11	3564.87
非コア　Yahoo！知恵袋	42	3986.83	2901.45
RWC　新聞	66	473.79	3903.88

５．結果

ベースラインとアッパーバウンドの結果を表２に示す。Self はタグつきターゲットデータが手に入ったと仮定して、supervised の学習を 5 分割交差検定を用いて行った結果であり、アッパーバウンドである。また、表 3 に提案手法による繰り返し回数が 10 回目（ステージ 10）の 10 セット（ランダムだけ 2 セット）の平均の正解率を表す。表中の「macro」と「micro」はそれぞれマクロ平均、マイクロ平均を表している。表中では各コーパスはそれぞれコアデータの YAHOO 知恵袋（OC）、コアデータの白書(OW)、YAHOO ブログ(OY)、新聞(PN)、書籍(PB)、雑誌(PM)、 非コアデータの YAHOO 知恵袋、(YAHOO)、非コアデータの白書(BCCWJ)コアデータ２種、RWC コーパス(RWC)となっている。図 1 中の、「all_senses_100」は初期事例集合にすべての語義を含む 100 件のデータを使用したもの、「all_senses_200」は初期事例集合にすべての語義を含む 200 件のデータを使用したもの、「random_100」は初期事例集合に完全にランダムな 100 件のデータを使用したものである。図 1 は、全体のマクロ平均と訓練事例を示している。図の「average」は「平均的なコーパス」、「big」は「最大のコーパス」、「all」は「すべてのコーパス」をそれぞれ示す。

表 2　ベースラインとアッパーバウンド

	最大のコーパス		平均的なコーパス		すべてのコーパス		Self	
(%)	macro	micro	macro	micro	macro	micro	macro	micro
OC	68.42	64.22	60.47	53.81	76.09	74.13	79.80	84.02
OW	70.22	73.74	54.48	53.87	65.14	68.18	85.29	90.43
OY	77.04	75.99	67.00	57.86	82.51	86.23	77.22	82.81
PB	76.08	78.73	62.01	62.05	77.58	80.44	79.68	84.76
PM	77.45	78.70	65.16	59.07	78.32	87.61	71.98	87.67
PN	77.73	77.94	64.41	63.04	80.79	81.75	72.77	76.85
BCCWJ	83.06	86.26	64.18	70.78	84.45	86.82	90.47	95.30
YAHOO	75.26	71.22	62.18	55.17	79.28	74.70	89.73	89.23
RWC	79.08	66.12	55.29	51.31	79.92	68.05	82.34	89.20
平均	**77.26**	**75.02**	**61.12**	**59.16**	**79.58**	**77.75**	**80.27**	**90.40**

表 3　各ドメイン別正解率と全体の正解率

	all_senses_100		all_senses_200		random_100	
(%)	macro	micro	macro	micro	macro	micro
OC	72.84	70.12	74.55	72.61	76.23	73.87
OW	65.46	65.81	77.17	80.59	67.13	69.82
OY	76.41	78.20	69.07	72.32	78.23	84.75
PB	73.30	72.55	73.88	73.26	79.05	77.57
PM	74.52	82.87	77.92	76.78	76.60	86.94
PN	76.67	73.84	75.59	83.22	81.06	80.56
BCCWJ	75.57	83.24	75.43	71.64	82.14	86.78
YAHOO	73.33	68.69	77.77	84.21	80.15	78.53
RWC	76.17	71.83	76.96	66.93	77.66	67.34
平均	**74.69**	**73.39**	**76.05**	**74.89**	**78.47**	**75.14**

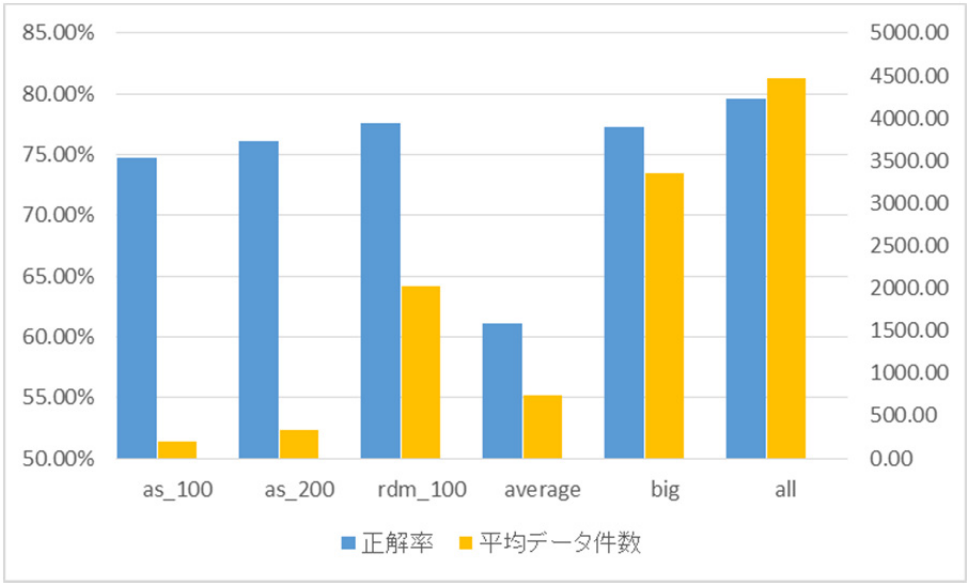

図１　正解率のマイクロ平均と訓練事例数

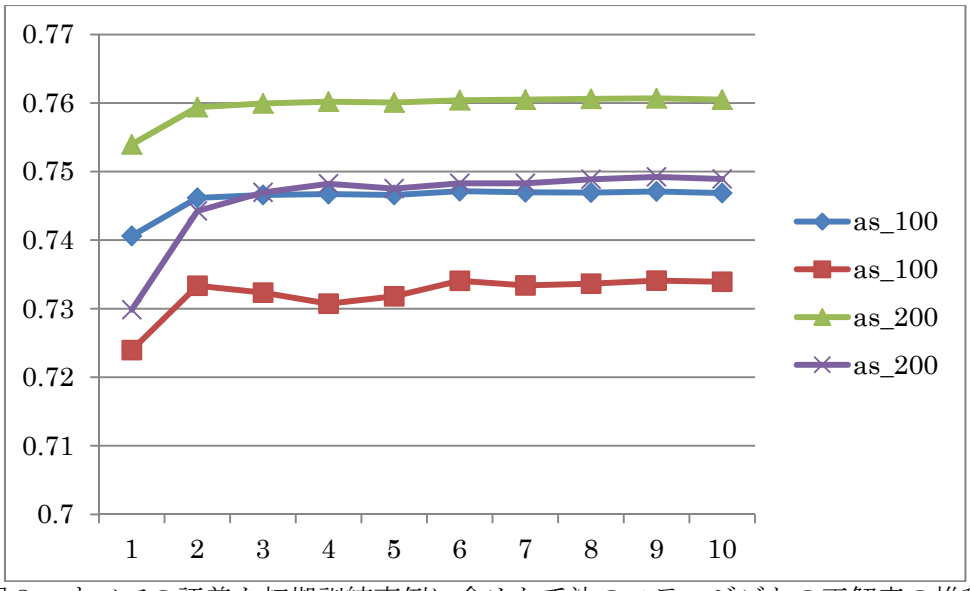

図２　すべての語義を初期訓練事例に含めた手法のステージごとの正解率の推移

６．考察

図１から、提案手法はベースラインよりも、少ないデータ数でベースラインに近い正解率を出していることが分かる。特に、「最大のコーパス」と「random_100」を比較した際、「random_100」の方が、訓練事例数が少ないのにもかかわらず、正解率はわずかながら上回っている。また、「as_100」や「as_200」、そして「random_100」を「平均的なコーパス」と比較すると、「as_100」、「as_200」、「random_100」の方が、訓練事例数が少ないのにもかかわらず、正解率が「平均的なコーパス」を上回っている。このことから、実験で用いた確信度と LOO-bound を用いたスコアが初期事例を選択する際に有効にであったと考えられる。

しかし、表２、表３からベースラインを上回ったのはドメイン別に見ると「白書」のコアデータのみで、全体の平均では、すべてのコーパスの結果に届かなかったことが読み取れる。また、図２を見ると正解率が３回目からはほとんど増加していない。そのため、サ

ポートベクトルを継承することで、分離平面の更新が起こりにくくなり、局所解に陥ってしまったと考えられる。このため、もっとサポートベクトルが入れ替わるような設定をするなどの改良をしたほうがよいと思われる。

次に、図 1 から、「all_senses_100」と「random100」を比較すると、正解率こそ「random_100」の方が優れているが、「all_senses_100」の方がより少ない事例数で分類できていることが分かる。訓練事例数は、「all_senses_100」は 189 件だったのに対し「random_100」は 2030 件であった。このことから、確信度と LOO-bound を用いたスコアが、訓練事例集合に最初から全ての語義を含むことで、より小数の訓練事例で正解率が収束することが分かる。また、「all_senses_100」や「all_senses_200」は、「平均的なコーパス」に比べ、訓練事例数を格段に少なくしながら、正解率を上昇させている。そのため、「all_senses_100」は、少量のデータを使用しながらも比較的、正解率を落とさないことが分かった。

また、「all_senses_100」の結果ステージ 10 の訓練事例が 189 件だったため、「all_senses_100」と、189 件よりも少々多めの 200 件をランダムに選択して、確信度などのスコアを使わずに分類器を作成した場合（すべての語義を含む。また、10 回の平均値）を比較した。その結果、「all_senses_100」はマイクロ平均が 73.39%、マクロ平均が 74.69%だったのに対して、ランダムの 200 件では、マイクロ平均が 72.87%、マクロ平均が 75.16%となった。このうち、マイクロ平均の結果はカイ二乗検定により有意であった。このことから、マクロ平均は、わずかに下がってしまう（有意ではない）が、マイクロ平均は確信度と LOO-bound を用いて上昇したことが分かった。このことから、局所解には陥ったものの、確信度と LOO-bound を用いたスコアにより、サポートベクトルを残して反復的に訓練事例集合を増やしていく手法は、マイクロ平均においては、語義曖昧性解消の学習に有効な訓練事例を選択するのに有効な手法であることが分かった。

7．おわりに

本稿では、semi-supervised な領域適応において、ソースデータに複数ドメインからなるデータを用いた場合に、確信度と LOO-bound を用いて部分集合を選択し、そのサポートベクトルのみを継承し反復的に訓練事例集合を選択する手法について述べた。正解率こそ全てのデータを利用するというベースラインを下回ってしまったが、正解率を大幅には落とさずに、訓練事例数を大幅に減らすことに成功した。また、その際、訓練事例数がより多かった「平均的なコーパス」の正解率を上回った。このことから、提案手法は、学習に有効な訓練事例を選択するという点において有効であることが分かった。

また、サポートベクトルの継承については局所解に陥るという問題があり、この点はもっとサポートベクトルが入れ替わるようにしたほうがよいと思われる。半面、このように反復的な訓練事例の選択を行うことで、微小ながらも正解率を上昇させるということが分かった。今後は、サポートベクトルを継承しないランダムな訓練事例集合を比較対象に含むなど、局所解に陥らないような工夫を施せば、正解率を上げることができるかもしれない。

謝辞

本研究は、文部科学省科学研究費補助金[若手 B（No : 24700138）]の助成により行われました。 ここに、謹んで御礼申し上げます。

参考文献

Chih-Chung Chang and Chih-Jen(2001),「Lin.LIBSVM: a library for support vectormachines」. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Koichi Hashida, Hitoshi Isahara, TakenobuTokunaga, Minako Hashimoto, Shiho Ogino,and Wakako Kashino(1998).「The rwc text databases」*In LREC 1998*, pp.

457-461.
Kanako Komiya and Manabu Okumura(2012). 「Au-tomatic domain adaptation for word sense dis-ambiguation based on comparison of multipleclassiers」*In PACLIC 2012*, pp. 77-85.
Kikuo Maekawa (2008). 「Balanced corpus of contempo-rary written japanese」 *In ALR 2008*, pp. 101-102.
古宮嘉那子、奥村学、小谷善行. 「分類器の確信度を用いた合議制による語義曖昧性解消の semi-supervised な領域適応」 *第三回コーパス日本語学ワークショップ予稿集*, pp. 1-6, 2013.
古宮嘉那子、小谷善行、奥村学(2013). 「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」 *第十九回言語処理学会年次大会予稿集*, pp.940-943
国立国語研究所(1964). *『分類語彙表』*. 秀英出版.
西尾実，岩淵悦太郎，水谷静夫(1994). *『岩波国語辞典第五版』*. 岩波書店.