外れ値検出手法を利用した Misleading データの検出

吉田 拓夢 (茨城大学 工学部 情報工学科) [1]
新納 浩幸 (茨城大学 工学部 情報工学科) [2]

Detection of Misleading Data by Outlier Detection Methods

Hiromu Yoshida (Ibaraki University, Department of Computer and Information Sciences)
Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) の領域適応の問題に対して，識別精度を低下させる Misleading データを検出するために，外れ値検出手法を利用する．

自然言語処理のタスクにおいて帰納学習手法を用いる際，訓練データとテストデータは同じ領域のコーパスから得ていることが通常である．ただし実際には異なる領域である場合も存在する．そこである領域（ソース領域）の訓練データから学習された分類器を，別の領域（ターゲット領域）のテストデータに合うようにチューニングすることを領域適応という (Sogaard (2013)) [3]．領域適応の問題の一つは負の転移である (Rosenstein et al. (2005))．これはソース領域のデータを使いすぎるとターゲット領域での識別精度が下がる現象である．我々は負の転移の原因を Misleading データの存在だと考えている．Misleading データとは分類器の学習に悪影響を与えるデータであり，Misleading データを検出，削除しておくことは分類器の精度向上に寄与する (Jiang and Zhai (2007))．

本論文では Misleading データはターゲット領域に対して外れ値になっていると予想し，この予想を確認する．まず訓練データ D から分類器を作成し，テストデータでその正解率 p_0 を測る．次に D の中の各データ x に対して，$D-x$ から分類器を作成し，テストデータでその正解率 p_1 を測る．$p_1 > p_0$ のとき x は Misleading データと考えられる．このようにして予め Misleading データを検出しておき，それを正解データと考えて，次に外れ値検出手法を利用して，どの程度 Misleading データを検出できるかを調べる．外れ値検出手法としては (1) 最近傍法，(2) LOF および (3) 確率密度比を試した．

実験では現代日本語書き言葉均衡コーパス (BCCWJ コーパス (Maekawa (2007))) 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 (Okumura et al. (2010)) 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．

実験の結果，Misleading データの存在自体は確認できたが，外れ値検出手法による Misleading データの検出精度は低かった．外れ値検出手法では本論文で設定したような Misleading データの検出は困難であるが，負の転移の有無を判定することは，ある程度可能であることが判明した．外れ値ではない Misleading データがどのような特徴を持っていたかを調べることが今後の課題である．

2　外れ値検出手法

本論文では外れ値検出手法を利用して，Misleading データの検出を試みる．利用した外れ値検出手法は (1) 最近傍法，(2) LOF および (3) 確率密度比である．以下，それぞれの手法を説明する．

[2] 10t4067l@hcs.ibaraki.ac.jp
[2] shinnou@mx.ibaraki.ac.jp
[3] 領域適応は機械学習の分野では転移学習 (神嶌敏弘 (2010)) の一種と見なされている．

2.1 最近傍法 (Erk の手法)

Erk による外れ値検出の手法 (Erk (2006)) を示す．外れ値の度合いを測るデータ点を点 x とする．この点 x に対して，対象データの中で最近傍となる点 t_n と，その点 t に対する最近傍点 t_n' を定める．これらの 3 つの点について，以下の距離を求める．

点 x と点 t_n の距離 dxt
点 t_n と点 $t_n{}'$ の距離 dtt'

この 2 つの距離を用いて，以下のように外れ値 $p_{NN}(x)$ を定める．

$$p_{NN}(x) = \frac{d_{xt}}{d_{tt'}}$$

2.2 LOF

LOF(local outlier factor) は密度をベースとした外れ値検出手法である (Breuning et al. (2000))．ある点のまわりの密度が他の点と比べて小さいほど，LOF の値は大きくなる．LOF の値を測る点を x としたとき，x の k 距離近傍集合 $N_k(x)$ を以下の様に定める．

$$N_k(x) = \{y \in D \setminus \{x\} | d(x,y) \leq kdist(x)\}$$

ここで，$kdist(x)$ は以下の条件を満たす $d(x,o)$ である．

1. 少なくとも k 個のデータ $o' \in D \setminus \{x\}$ に対して $d(x,o') \leq d(x,o)$ が成立する
2. 高々 $k-1$ 個のデータ $o' \in D \setminus \{x\}$ に対してのみ $d(x,o) < d(x,o)$ が成立する

すなわち，簡単には k 距離近傍集合 $N_k(x)$ は点 x から k 番目に近い点 o_k までの距離 $kdist(x) = dist(x,o_k)$ の範囲内にある点の集合である．LOF の算出に先立ち，まずは以下を求める．

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} rd_k(x,y)}$$

これは局所到達可能密度 (local reachability density，lrd) と呼ばれる値で，x の k 近傍内にあるデータの到達可能距離 (reachability distance，rd) の平均の逆数となっている．到達可能距離 rd は以下で定める値である．

$$rd_k(x,y) = \max\{d(x,y), kdist(y)\}$$

つまり，点 x と y の距離が y の k 距離よりも近い場合には y の k 距離に置き換えて到達可能距離 rd としている．以上をもって，LOF は次式により定められる．

$$LOF(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}$$

上式に示されるとおり，LOF は点 x の局所到達可能密度と点 x の k 近傍点の局所到達可能密度との平均を取っている．

2.3 確率密度比

確率密度比も外れ値検出手法として利用できる．確率密度比の算出は困難であるが，論文 (新納浩幸・佐々木稔 (2014)) では以下の簡易な手法を提案している．

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．求めるのは領域 $R \in \{S,T\}$ 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しておく．ここで用例の数を $N(R)$ とおく．また $N(R)$ 個の用例の中で，素性 f が現れた用例数を $n(R,f)$ とおく．MAP 推定でスムージングを行い，$P_R(f)$ を以下で定義する (高村大也 (2010))．

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

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

3　実験

BCCWJ コーパスの 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 (Okumura et al. (2010)) 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 1 に示す．[4]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．

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

3.1　正解データの構築

語義曖昧性解消の領域適応において Misleading データの存在を確かめるため，以下のような実験を行った．実験のタスクにおいて 6 つの領域適応が行われるが，それぞれの領域適応のための機械学習で用いるソースデータを $\mathbf{x}^{(\mathrm{S})} = \{x_1, x_2, ..., x_n\}$ とする．このソースデータに対して，任意の i 番目のデータ x_i1 つを取り除いた新たなソースデータを $\mathbf{x}^{(\mathrm{S})}{}_i$ とする．ここで新たなソースデータ $\mathbf{x}^{(\mathrm{S})}{}_i$

[4]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 タスクでは新語義のタグも許しているからである．

で学習を行った場合に，元のソースデータ $\mathbf{x}^{(S)}$ で学習を行った場合よりも分類器の精度が向上したならば，$\mathbf{x}^{(S)}$ に含まれるデータ x_i は学習の精度を下げる Misleading データだったと考えられる．ソースデータ $\mathbf{x}^{(S)}$ に対して ${\mathbf{x}^{(S)}}_1$ から ${\mathbf{x}^{(S)}}_n$ までの新しい n 個のソースデータを作成，学習し，データ $\mathbf{x}^{(S)}$ に含まれる n 個全てのデータ点 x_i が Misleading データであるかどうかをそれぞれ 1 つずつ判別する．このようにして判別された Misleading データの集合を，その領域適応における検出すべき Misleading データの正解集合とした．結果を表 2 に示す.

表 2: 検出した Misleading データ

単語	OC		PB		PN	
	PB	PN	OC	PN	OC	PB
言う	159/666	158/666	127/1114	75/1114	82/363	35/363
	23.87	23.72	11.40	6.730	22.59	9.640
入れる	6/73	28/73	19/56	15/56	3/32	1/32
	8.220	38.36	33.93	26.79	9.380	3.130
書く	21/99	39/99	0/62	2/62	12/27	15/27
	21.21	39.40	-	3.230	44.44	55.56
聞く	26/124	21/124	26/123	0/123	4/52	27/52
	20.97	16.94	21.14	-	7.700	51.92
子供	5/77	0/77	12/93	1/93	12/29	13/29
	6.490	-	12.90	1.080	41.38	44.83
時間	1/53	8/53	0/74	0/74	0/59	5/59
	1.890	15.09	-	-	-	8.470
自分	13/128	25/128	0/308	0/308	0/71	1/71
	10.16	19.53	-	-	-	1.410
出る	14/131	10/131	39/152	32/152	22/89	10/89
	10.69	7.630	25.66	21.05	24.72	11.24
取る	6/61	5/61	10/81	18/81	12/43	22/43
	9.840	8.200	12.35	22.22	27.91	51.16
場合	0/126	0/126	7/137	13/137	14/73	9/73
	-	-	5.110	9.490	19.18	12.33
入る	36/68	11/68	38/118	27/118	27/65	42/65
	52.94	16.18	32.20	22.88	41.54	64.62
前	8/105	5/105	10/160	1/160	15/106	2/106
	7.620	4.760	6.250	0.625	14.15	1.890
見る	10/262	3/262	3/273	12/273	8/87	28/87
	38.18	1.150	1.100	4.400	9.200	32.18
持つ	8/62	0/62	2/153	11/153	1/59	1/59
	12.90	-	1.310	7.190	1.690	1.690
やる	0/117	0/117	0/156	0/156	0/27	0/27
	-	-	-	-	-	-
ゆく	17/219	0/219	15/133	1/133	3/27	3/27
	7.760	-	11.29	0.752	11.11	11.11

3.2 外れ値検出手法による検出評価

16 単語の 6 つの領域適応において，提案手法により 3 手法の外れ値の計算を行った．LOF では算出した値を正規化し，閾値 $\theta = 1.96$ より大きな値を Misleading データとみなした．Erk, 密度比の手法はそれぞれ Misleading データの検出正答率の平均が大きくなるような閾値を探し，結果，Erk の手法では閾値 $\theta = 1.9$ より大きな値を，密度比の手法では閾値 $\theta = 0.005$ より小さな値を Misleading データとした．

これらの提案手法による Misleading データの検出正答率を，6 領域適応ごとに 16 単語の平均を取った．これを表 3 に示す．検出正答率は検出した Misleading データの数でそのうちの Misleading データの正答集合に含まれる数を割ったものである．いずれの手法，領域適応においても検出正答率は著しく低い．

表 3: 提案手法による検出正答率 (%)

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	3.590	9.920	14.36	19.07	10.64	29.32	14.48
LOF	3.67	10.19	6.990	5.260	18.48	19.62	10.70
密度比	10.49	10.03	8.880	11.93	17.02	22.10	13.41

また，提案手法による misleading データを除いた場合の領域適応の正答率を表 4 に示す．LOF の手法のみが僅かに通常の領域適応の正答率を上回った．

表 4: 提案手法で misleading を除いた場合の正答率 (%)

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	70.91	68.27	69.45	75.38	69.45	73.33	70.10
LOF	70.71	67.11	70.09	75.57	68.70	73.44	70.94
密度比	69.63	67.34	68.38	76.37	59.72	66.42	67.98
NORMAL	70.77	66.96	70.29	75.56	68.49	73.26	70.89

4 考察

4.1 最近傍距離と Misleading データとの相関

外れ値検出の手法では Misleading データの検出能力が極めて低いと言える．外れ値検出はいずれの手法もデータ点の距離の差異を利用するものであるが，Misleading データと非 Misleading データの間にはその差異が認められなかったと考えられる．そこで，Misleading データと非 Misleading データについて，以下を調べる．

- ターゲットデータに対する最近傍点への距離についての差異の有無

実験で使用した 16 単語 6 種計 96 ケースの領域適応のデータについて，外れ値検出同様にそれぞれのソースデータ点のターゲットデータへの最近傍距離を算出した．これを Misleading データの正答集合により，Misleading データと非 Misleading データに分けて平均を取り，それらの相関を t 検定により判定した．なお，有意水準は 5% とした．その結果，有意差が認められたケースは 96 ケース中 12 ケースであった．また，その有意差の有無と外れ値の 3 手法による Misleading データの検出結

果を照らし合わせたところ，有意差が認められた 12 ケースの領域適応において Misleading データの検出率が特段高い訳でもなかった．Misleading データはターゲットデータへの最近傍距離において非 Misleading データとの差異があるとは言えず，そのため，距離を利用する外れ値検出の手法をもって Misleading データを判別することは難しい．

4.2　負の転移が生じない対象単語の検出

外れ値検出手法を用いても，ここで設定したような Misleading データを検出することは困難であった．本節では Misleading データと関連の深い負の転移の有無を外れ値検出手法で判定できるかを調べてみる．

まず論文 (新納浩幸・佐々木稔 (2014)) では本論文と同じデータを利用して，負の転移が生じなかった対象単語を選出している．その結果は以下の表 5 にまとめられる．表 5 でチェックが付いているものが，負の転移が生じなかった単語である．

表 5: 負の転移が生じていない領域適応

単語	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC
言う		✓	✓	✓	✓	
入れる		✓	✓	✓	✓	✓
書く	✓			✓	✓	
聞く	✓					
子供			✓		✓	
時間	✓		✓		✓	
自分	✓	✓				
出る				✓		✓
取る			✓		✓	✓
場合		✓	✓		✓	✓
入る		✓	✓		✓	✓
前		✓				
見る	✓					
持つ	✓	✓				✓
やる		✓		✓		✓
ゆく		✓		✓	✓	

次に本論文で行った外れ値検出手法で検出された Misleading データの割合が，全体のデータの 1 割以下である場合に，負の転移が生じないという判定を行う．これによって外れ値検出手法を利用して，負の転移が生じない対象単語の検出評価を行うことができる．検出の正解率，再現率，F 値をそれぞれ表 6，表 7，表 8 に示す．Mislead は本実験で用いた Misleading の正解データを利用した検出を示す．

表 6: 外れ値検出手法を利用した負の転移が生じない単語の検出・正解率

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	0.286	0.333	0.500	0.533	0.250	0.615	0.420
LOF	0.375	0.375	0.438	0.563	0.438	0.563	0.458
密度比	0.000	1.000	0.000	0.500	0.000	0.000	0.250
Mislead	0.125	0.333	0.375	0.583	0.286	0.429	0.355

表 7: 外れ値検出手法を利用した負の転移が生じない単語の検出・再現率

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	0.667	0.833	1.000	0.889	0.429	0.889	0.784
LOF	1.000	1.000	1.000	1.000	1.000	1.000	1.000
密度比	0.000	0.333	0.000	0.556	0.000	0.000	0.148
Mislead	0.167	0.500	0.429	0.778	0.286	0.333	0.415

表 8: 外れ値検出手法を利用した負の転移が生じない単語の検出・F 値

	OC		PB		PN		avr
	PB	PN	OC	PN	OC	PB	
Erk	0.400	0.476	0.667	0.667	0.316	0.727	0.542
LOF	0.545	0.545	0.609	0.720	0.609	0.720	0.625
密度比	-	0.500	-	0.526	-	-	0.513
Mislead	0.143	0.400	0.400	0.667	0.286	0.375	0.378

表 8 を見ると，Misleading の正解データを用いても負の転移のない単語を検出する能力は高くない．それに比較すれば外れ値検出手法を利用した場合の検出する能力は高い．外れ値検出手法を利用して負の転移が生じない単語を判定できる可能性もあり，この点で精度改善が可能であると考える．

5　おわりに

本論文では WSD の領域適応における Misleading データの検出に外れ値検出手法を利用することを試みた．総当たり的に各データが Misleading データと見なせるかどうかを調べることで，Misleading データの存在を確認できた．また，それらを正解集合として外れ値検出を用いた検出能力も調べた．結論的には外れ値検出手法を利用しても，本論文で設定したような Misleading データの検出は困難であることがわかった．ただし Misleading データと関連の深い負の転移現象の有無を判定することには利用可能だと考えている．今後は外れ値ではない Misleading データの特徴を調査することで，新たな Misleading データの検出法を考えたい．

文献

Markus M. Breuning, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander (2000) “LOF: Identifying Density-Based Local Outliers,” in *ACM SIGMOD 2000*, pp. 93–104.

Katrin Erk (2006) “Unknown Word Sense Detection As Outlier Detection,” in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 128–135.

Jing Jiang and Chengxiang Zhai (2007) “Instance weighting for domain adaptation in NLP,” in *Proc. of ACL-2007*, pp. 264–271.

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

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono (2010) “SemEval-2010 Task: Japanese WSD,” in *Proc. of the 5th International Workshop on Semantic Evaluation*, pp. 69–74.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich (2005) “To transfer or not to transfer,” in *Proc. of the NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*.

Anders Sogaard (2013) *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*: Morgan & Claypool.

高村大也 (2010) 言語処理のための機械学習入門，コロナ社.

新納浩幸、佐々木稔 (2014) 「共変量シフトの問題としての語義曖昧性解消の領域適応」，自然言語処理，第 21 巻，第 1 号，(to appear).

神嶌敏弘 (2010) 「転移学習」，人工知能学会誌，第 25 巻，第 4 号，pp.572–580.