様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　２　日現在

機関番号：　１２１０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２４４

研究課題名（和文）外れ値検出手法からの重み設定による共変量シフト下における語義曖昧性解消の領域適応

研究課題名（英文）Learning under covariate shift for domain adaptation for word sense disambiguation through weight setting using a outlier detection method

研究代表者

新納　浩幸（Shinnou, Hiroyuki）

茨城大学・工学部・教授

研究者番号：１０２５０９８７

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では語義曖昧性解消の領域適応の問題を共変量シフト下の学習によって解決する。その際に問題となるのは事例への重みである確率密度比の算出方法と、重み付き学習の手法である。確率密度比の算出方法としては拘束無し最小二乗重要度適合法を試みた。そこで用いるカーネル関数は通常、ガウスカーネルであるが、線形モデルの方が本タスクには適していることを示した。また重みは連続値ではなく、大中小の3つの離散値を用いる手法を提案した。また重み付き学習の手法は通常最大エントロピー法を利用するが、SVMも利用できることを示した。カーネル関数、重み付き学習、重みの大別処理の最善の組み合わせを求めた。

研究成果の概要（英文）：In this research, I solved the domain adaptation for word sense disambiguation by using the learning method under the assumption of covariate shift. The key point of this approach is how to estimation of the probability density ratio, and how to conduct the weighted learning. For the first problem, I adopt unconstrained least squares importance fitting (uLSIF). In this research, I showed that a linear kernel is better than a Gaussian kernel used as the basis function generally. Furthermore, I proposed to use 3 kinds of discrete values as a weight. For the second problem, I showed that SVM also is available but the maximum entropy method. Furthermore, I combined the kernel function, the weighted learning and discrete weights.

研究分野：　自然言語処理

キーワード：　語義曖昧性解消　領域適応　共変量シフト　外れ値検出

１．研究開始当初の背景

自然言語処理の多くのタスクで教師付き学習が利用されているが，そこでは訓練データとテストデータの領域が異なるという領域適応の問題が生じている．例えば「ゴルフ」という単語には sport と car の意味があり，その語義曖昧性解消を考えた場合，学習元のコーパスがスポーツ記事であれば主に sport と判定される規則が学習されるが，その規則を車の記事に適用すると誤る場合が多いという問題である．このような領域適応の問題に対する手法には，大きく分けて事例ベースの手法と素性ベースの手法が存在する．素性ベースの手法では，概略，ソース領域の素性空間をターゲット領域に合うように変換する．また事例ベースの手法では，概略，訓練事例に重みをつけて重み付き学習を行う．

自然言語処理の領域適応では素性ベースの手法が数多く試されてきたが，自然言語処理のタスクでは領域が変わっても通常，文の意味は変化しないという特徴がある．これは統計的には共変量シフトの仮定が成立していることに対応し，共変量シフト下の学習は事例ベースの手法と位置づけられる．

そこで本研究は自然言語処理の領域適応のタスクに対して，共変量シフト下の学習を試みる．

２．研究の目的

本研究では自然言語処理のタスクとして語義曖昧性解消を取りあげ，その領域適応の問題に対して効果的な共変量シフト下の学習方法を提案することである．

共変量シフトとは $P_s(c|x) = P_t(c|x)$ であるが $P_s(x) ¥= P_t(x)$ という仮定である．一般に，共変量シフト下の学習では，x の確率密度比 $w(x) = P_t(x)/P_s(x)$ 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ることで，$P_t(c|x)$ を構築するアプローチが取られる．つまり共変量シフト下の学習のポイントは確率密度比の推定である．直接的に推定する方法としては，$P_t(x)$ と $P_s(x)$ をそれぞれモデル化して求め，その比を取る方法があり，過去にもこの類の手法がいくつか提案されている．しかし $P_t(x)$ と $P_s(x)$ のモデル化が困難なことも多い．そこで確率密度比 $w(x)$ 自体をモデル化して推定するという手法が機械学習の分野では研究されている．ここではその中の比較的新しい手法である拘束無し最小二乗重要度適合法という手法をここでのタスクに試みる．また外れ値検出の手法を利用して重み自体を推定する手法も試みる．

また確率密度比 $w(x)$ が推定された後にそれを x の重みとして，学習する必要がある．通常は最大エントロピー法により $P_t(c|x)$ を学習するが，ここでは分類器自体を学習することを試みる．具体的には SVM の重み付き学習を試みる．

３．研究の方法

まず提案手法を評価するためのデータセットを構築する必要がある．ここでは現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese, BCCWJ）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 の日本語語義曖昧性解消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，語義曖昧性解消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 * 6 = 96 通りの語義曖昧性解消の領域適応の問題に対して提案手法の評価を行う．

利用する拘束無し最小二乗重要度適合法はカーネル関数の線形モデルである．カーネル関数は一般にガウスカーネルが利用されるが，ここでは自然言語処理の分野で一般に利用される線形カーネルを利用する．またここで求まった重みは連続値あるが，大中小の 3 種の離散値に直した方が経験的に精度が向上する．また重み付き学習は通常の最大エントロピー法の他に不均衡データに対する SVM の手法を応用し，
SVM の重み付き学習を行う．

重みの設定方法と重み付き学習を組み合わせて，評価用データセットの平均正解率で評価する．

また外れ値検出の手法を利用した重みの推定には近年深層学習を利用したものが提案された．そこで提案されている手法をインプリメントして試す．

４．研究成果

対象単語 w についてソース領域 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．ソース領域 S の訓練データのみを用いて，手法 A により分類器を学習し w に対する正解率を求める．16 種類の各対象単語に対する正解率の平均をソース領域 S からターゲット領域 T に対する手法 A の正解率とする．結果，手法 A について 6 種類の各領域適応に対しての正解率が得られる．それらの平均を手法 A の平均正解率とする．本論文で扱う手法 A は 2 つの要素から構成される．確率密度比の算出法と重み付き学習の種別である．以下の表の組み合わせが得られる．Base-M は重み付けなしで学習法に最大エントロピー法を用いたもの．Base-S は重み付けなしで学習法に SVM を用いたもの．Mtd-G-M はガウスカーネルを利用した拘束無し最小二乗重要度適合法（uLSIF）により確率密度比を推定し，重み付き学習には最大エントロピー法を用いたもの，Mtd-G-S は学習法に SVM を用いたもの．Mtd-L-M は線形カーネルを利用した uLSIF により確率密度

比を推定し，重み付き学習には最大エントロピー法を用いたもの，Mtd-L-S は学習法に SVM を用いたもの．Ours-G-M はガウスカーネルを利用した uLSIF により確率密度比を推定し，それを提案手法により 0.1，1.1 および 2.1 に変換し，重み付き学習には最大エントロピー法を用いたもの，Ours-G-S は学習法に SVM を用いたもの．Ours-G-M は線形カーネルを利用した uLSIF により確率密度比を推定し，それを提案手法により 0.1，1.1 および 2.1 に変換し，重み付き学習には最大エントロピー法を用いたもの，Ours-L-S は学習法に SVM を用いたものである．本研究の手法は Ours-L-S である．

手法	重み付け	学習法
Base-M	1	ME
Base-S	1	SVM
Mtd-G-M	Gauss	ME
Mtd-G-S	Gauss	SVM
Mtd-L-M	Linear	Me
Mtd-L-S	Linear	SVM
Ours-G-M	G→3 種	Me
Ours-G-S	G→3 種	SVM
Ours-L-M	L→3 種	ME
Ours-L-S	L→3 種	SVM

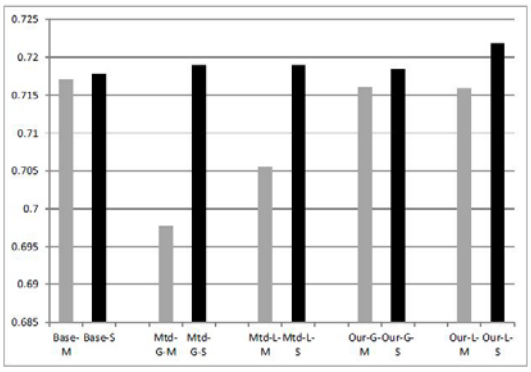

図 1： SVM と ME の比較

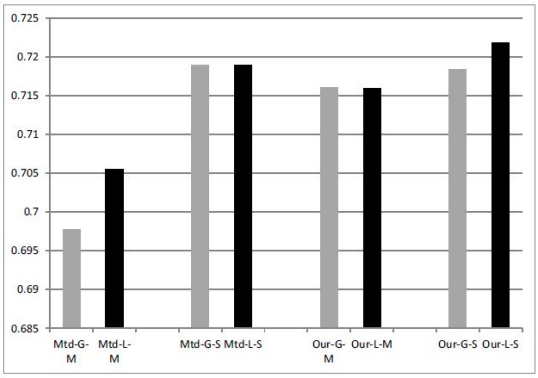

図 2：ガウスと線形の比較

実験結果は Base-M < Base-S，Mtd-G-M < Mtd-G-S，Mtd-L-M < Mtd-L-S，Our-G-M < Our-G-S，Our-L-M < Our-L-S が成立したので，最大エントロピー法を用いるよりも SVM を用いる方が効果があることがわかる．また Mtd-G-M < Mtd-L-M，Mtd-G-S = Mtd-L-S，Ours-G-S < Mtd-L-S が成立し，Ours-G-M と Ours-L-M はそれぞれ 0.7160 と 0.7159 でありほぼ等しい．このため uLSIF ではガウスカーネルを利用するよりも線形カーネルを利用する方が効果的であることがわかる．さらに本研究成果による手法 Ours-L-S は最も高い平均正解率である．また各領域適応においても，PN → PB を除いて最も高い正解率を示した(図 1，図 2 参照)．

本研究で得られた知見は共変量シフト化の学習を語義曖昧性解消の領域適応の問題に利用する場合には，拘束無し最小二乗重要度適合法を離散値に変換し，重み付き学習には SVM を利用することが精度向上のポイントであることである．この知見を利用して，語義曖昧性解消以外の様々な自然言語処理の領域適応の問題に利用する．また外れ値検出を利用した手法では良い結果が得られなかった．深層学習を利用する場合には素性ベースの手法の方が相性がよいということが分かった．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① 新納 浩幸 他，クラスタリングを利用した語義曖昧性解消の誤り原因のタイプ分け，自然言語処理，査読有，Vol.22, No.5, 2015, pp.319-362.
② 新納 浩幸 他，共変量シフト下の学習による語義曖昧性解消の教師なし領域適応，自然言語処理，査読有，Vol.21, No.5, 2014, pp.1101-1035.

〔学会発表〕（計 40 件）

① Hiroyuki Shinnou, etc., "Learning under Covariate Shift for Domain Adaptation for Word Sense Disambiguation", PACLIC-29, 査読有, 2015, pp.215-223.
② Hiroyuki Shinnou, etc., "Hybrid Method of Semi-supervised Learning and Feature Weighted Learning for Domain Adaptation of Document Classification",PACLIC-29, 査読有, 2015, pp.500-507.
③ Hiroyuki Shinnou ,etc., "Active Learning to Remove Source Instances for Domain Adaptation for Word Sense Disambiguation", PACLING-2015, 査読有, 2015, pp.156-162.
④ 新納浩幸, 他, "語義曖昧性解消におけるシソーラス利用の問題分析", 言語処理学会第 21 回年次大会, 査読無, 2015, P1-15.
⑤ 新納浩幸, 他, "uLSIF を用いた事例へ

の重み付けによる語彙曖昧性解消の領域適応",情報処理学会自然言語処理研究会, 査読無, 2014, NL-218-2.

〔図書〕(計　1　件)

① 新納 浩幸, オーム社, Chainer による実践深層学習, 2016, 182.

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者
新納 浩幸（SHINNOU, Hiroyuki）
茨城大学・工学部・教授
研究者番号：10250987

(2)研究分担者
無し

(3)連携研究者
無し

(4)研究協力者
無し