修士研究論文

ターゲット領域のキーワード含有率を事例の重みとした感情分析の領域適応

平成31年2月5日

茨城大学　工学部情報工学科
17NM719X　白　静

指導教員：新納浩幸教授

平成 30 年度茨城大学工学部情報工学科修士研究

ターゲット領域のキーワード含有率を事例の重みとした感情分析の領域適応

氏名：17NM719X 白 静
指導教員：新納 浩幸 教授

論文要旨

本論文では感情分析の領域適応に対して、新たな事例ベースの手法を提案する。

感情分析とはレビュー文書（例えば映画のレビュー）が肯定的なものか、否定的なものかを判定するタスクである。これは文書分類の一種であり、教師あり学習を用いて解決できる。しかし判定先の文書が学習データの領域とは異なる領域の文書（例えば書籍のレビュー）であった場合に、教師あり学習で得られた分類器の精度が下がってしまう。これが領域適応の問題である。

領域適応の問題に対する手法は、素性ベースのものと事例ベースのものに大別できる[7]。概略、素性ベースの手法は学習データの素性に重みを付けた学習であり、事例ベースの手法は学習データの事例に重みを付けた学習である。

実験では Amazon dataset [1] を用いた、確率密度比を求める 2 つ手法（ uLSIF [11] と論文 [13]の提案手法）と提案手法を比較することで、提案手法の有効性を示す。

Master's Thesis in Scholastic 2019, Major in Computer and In formation Sciences, Granduate School of Science and Engineering, Ibaraki University

Domain Adaptation for Sentiment Analysis using Keywords in the Target Domain as the Learning Weight

Author : 17NM719X Bai Jing
Adviser : Prof. Hiroyuki Shinnou

Abstract

This paper proposes a new method of instance-based domain adaptation for sentimentanalysis. First, our method de

nes the likelihood of keywords, through the value of inversedocument frequency (IDF), for each word in documents in the target domain. Next,the keyword content rate of a document is calculated using the likelihood of keywords and the domain adaptation is performed by giving the keyword content rate to eachdocument in the source domain as the weight. The experiment used an Amazon dataset to demonstrate the effectiveness of our proposed method. Although the instance-based method has not shown great efficiency, the advantages combining instance-based method and feature-based method are shown in this paper.

目 次

第1章　序論

1.1　概要

本論文では感情分析の領域適応に対して、新たな事例ベースの手法を提案する。

感情分析とはレビュー文書（例えば映画のレビュー）が肯定的なものか、否定的なものかを判定するタスクである。これは文書分類の一種であり、教師あり学習を用いて解決できる。しかし判定先の文書が学習データの領域とは異なる領域の文書（例えば書籍のレビュー）であった場合に、教師あり学習で得られた分類器の精度が下がってしまう。これが領域適応の問題である。

領域適応の問題に対する手法は、素性ベースのものと事例ベースのものに大別できる[7]。概略、素性ベースの手法は学習データの素性に重みを付けた学習であり、事例ベースの手法は学習データの事例に重みを付けた学習である。

ここでは新たな事例ベースの手法を提案する。従来、事例ベースの手法としては共変量シフトを仮定し、確率密度比を事例の重みとする手法が一般的である。ただし確率密度比を算出する計算コストは大きい。ここで提案する手法は単純なものであるが、確率密度比以上の効果がある。

提案手法ではターゲット領域の文書集合から、TF-IDF を用いてターゲット領域における単語wのキーワードの度合いd_wを設定する。このd_wを用いて、ソース領域における事例xがどの程度ターゲット領域のキーワードを含むかというキーワード含有率w_xを設定する。このw_xをxの重みとして重み付け学習を行うことで、領域適応の問題に対処する、

実験では Amazon dataset [1] を用いて、確率密度比を求める 2 つ手法（ uLSIF [11] と論文 [13]の提案手法）と提案手法を比較することで、提案手法の有効性を示す。

1.2 構成

本論文では感情分析の領域適応に対して、新たな事例ベースの手法を提案する。3章では感情分析の領域適応の説明を行う．4章では仮定されている共変量シフトを述べる．5章では確率密度比の算出手法を述べる．7章では本研究における提案を述べる．8章では，実験に利用したデータについてを述べる．9章では本研究においての問題点，改善点また今後の課題について述べる．

第2章　関連研究

領域適応の手法は、まず、ターゲット領域のラベル付きデータを用いる教師ありの手法と、それを用いない教師なしの手法に大別できる。教師ありの手法の場合、Daumé の手法 [3] が簡易でしかも能力が高いため、標準手法となっている。また様々な半教師あり学習を応用した手法も数多く提案されている。

本論文で扱うのは教師なしの手法である。教師なしの手法の場合、素性ベースのものと事例ベースのものに大別できる[7]。概略、素性ベースの手法は学習データの素性に重みを付けた学習であり、事例ベースの手法は学習データの事例に重みを付けた学習である。素性ベースの手法としては古典的には SCL [2] が有名である。近年は CORAL [8] が簡易でしかも能力が高いため注目されている。さらに素性ベースの手法は深層学習とも相性がよく [5]、CORAL を拡張した手法 [9]や敵対性ネットワークを利用した手法 [4][10] が state of the art と思われる。

第3章　領域適応

感情分析とはレビュー文書（例えば映画のレビュー）が肯定的なものか、否定的なものかを判定するタスクである。これは文書分類の一種であり、教師あり学習を用いて解決できる。しかし判定先の文書が学習データの領域とは異なる領域の文書（例えば書籍のレビュー）であった場合に、教師あり学習で得られた分類器の精度が下がってしまう。これが領域シフトの問題である。

領域シフトの問題に対する手法は、領域適応と呼ばれる。

領域適応は，転移学習（Transfer Learning）と呼ばれる学習手法の一つです．十分な教師ラベルを持つ領域（Source Domain，ソース領域）から得られた知識を，十分な情報がない目標の領域（Target Domain, ターゲット領域）に適用することで，目標領域において高い精度で働く識別器などを学習します。図3.1を示します。https://en.wikipedia.org/wiki/Domain_adaptation#/media/File:Transfer_learning_and_domain_adaptation.png

機械学習と転送学習に関連する領域である、手法は、素性ベースのものと事例ベースのものに大別できる[7]。概略、素性ベースの手法は学習データの素性に重みを付けた学習であり、事例ベースの手法は学習データの事例に重みを付けた学習である。

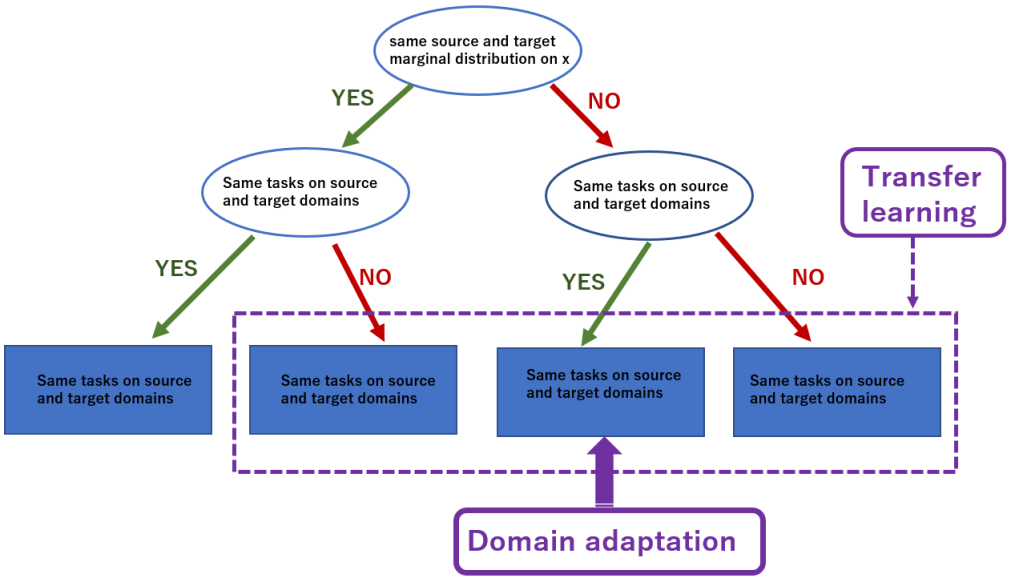
same source and target marginal distribution on x
YES
NO
Same tasks on source and target domains
Same tasks on source and target domains
Transfer learning
YES
NO
YES
NO
Same tasks on source and target domains
Same tasks on source and target domains
Same tasks on source and target domains
Same tasks on source and target domains
Domain adaptation

図 3.1: 領域シフト

第4章　共変量シフト下の学習

4.1　概要

事例ベースの手法では共変量シフトを仮定する。共変量シフトとは$P_S(c|\boldsymbol{x}) = P_T(c|\boldsymbol{x})$かつ$P_S(\boldsymbol{x}) = P_T(\boldsymbol{x})$という仮定である。共変量シフト下では、ソース領域のデータ$\boldsymbol{x}$に対して確率密度比$r = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$を重みとした重み付き学習から、$P_T(c|\boldsymbol{x})$を得ることができる。

4.2　期待損失最小化に基づく共変量シフト下の学習

対象単語wの語義の集合をC，またwの用例$\boldsymbol{x}$内のwの語義をcと識別したときの損失関数を$l(\boldsymbol{x}, c, d)$で表す．dはwの語義を識別する分類器である．$P_T(\boldsymbol{x}, c)$ をターゲット領域上の分布とすれば，本タスクにおける期待損失L_0は以下で表せる．

$$L_0 = \sum_{\boldsymbol{x},c} l(\boldsymbol{x}, c, d) P_T(\boldsymbol{x}, c)$$

また$P_S(\boldsymbol{x}, c)$ をソース領域上の分布とすると以下が成立する．

$$L_0 = \sum_{\boldsymbol{x},c} l(\boldsymbol{x}, c, d) \frac{P_T(\boldsymbol{x}, c)}{P_S(\boldsymbol{x}, c)} P_S(\boldsymbol{x}, c)$$

ここで共変量シフトの仮定から

$$\frac{P_T(\boldsymbol{x}, c)}{P_S(\boldsymbol{x}, c)} = \frac{P_T(\boldsymbol{x}) P_T(c|\boldsymbol{x})}{P_S(\boldsymbol{x}) P_S(c|\boldsymbol{x})} = \frac{P_T(\boldsymbol{x})}{P_S(\boldsymbol{x})}$$

となり，$w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$とおくと以下が成立する．

$$L_0 = \sum_{\boldsymbol{x},c} w(\boldsymbol{x}) l(\boldsymbol{x}, c, d) P_S(\boldsymbol{x}, c)$$

訓練データを$D = \{(\boldsymbol{x}_i, c_i)\}_{i=1}^N$とし，$P_S(\boldsymbol{x}, c)$を経験分布で近似すれば，

$$L_0 \approx \frac{1}{N}\sum_{i=1}^{N} w(\boldsymbol{x}_i) l(\boldsymbol{x}_i, c_i, d)$$

となるので，期待損失最小化の観点から考えると，共変量シフトの問題は以下の式L_1を最小にするdを求めればよいことがわかる．

$$L_1 = \sum_{i=1}^{N} w(\boldsymbol{x}_i) l(\boldsymbol{x}_i, c_i, d) \tag{4.1}$$

分類器dとして以下の事後確率最大化推定に基づく識別を考える．

$$d(\boldsymbol{x}) = \arg\max_c P_T(c|\boldsymbol{x})$$

また損失関数として対数損失$-\log P_T(c|\boldsymbol{x})$を用いれば，式(4.1)は以下となる．

$$L_1 = -\sum_{i=1}^{N} w(\boldsymbol{x}_i) \log P_T(c_i|\boldsymbol{x}_i)$$

つまり，分類問題の解決に$P_T(c|\boldsymbol{x}, \boldsymbol{\lambda})$のモデルを導入するアプローチを取る場合，共変量シフト下での学習では，確率密度比を重みとした以下に示す重み付き対数尤度$L(\boldsymbol{\lambda})$を最大化するパラメータ$\boldsymbol{\lambda}$を求める形となる．

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{N} w(\boldsymbol{x}_i) \log P_T(c_i|\boldsymbol{x}_i, \boldsymbol{\lambda}) \tag{4.2}$$

ここではモデルとして以下の式で示される最大エントロピー法を用いる．

$$P_T(c|\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{\lambda})} \exp\left(\sum_{j=1}^{M} \lambda_j f_j(\boldsymbol{x}, c)\right) \tag{4.3}$$

$\boldsymbol{x} = (x_1, x_2, \cdots, x_M)$が入力，$c$がクラスである．関数$f_j(\boldsymbol{x}, c)$は素性関数であり，実質$\boldsymbol{x}$の真のクラスが$c$のときに$x_j$を返し，そうでないとき 0 を返す関数に設定される．$Z(\boldsymbol{x}, \boldsymbol{\lambda})$は正規化項であり，以下で表せる．

$$Z(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum_{c \in C} \exp\left(\sum_{j=1}^{M} \lambda_j f_j(\boldsymbol{x}, c)\right) \tag{4.4}$$

そして$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_M)$が素性に対応する重みパラメータとなる．

第5章　確率密度比

5.1　概要

確率密度比の算出方法としては様々な手法がある。単純には$P_S(\boldsymbol{x})$と$P_T(\boldsymbol{x})$を求めればよいが、それらのモデルが複雑な場合は問題をより複雑にするために、直接、確率密度比をモデル化する手法が研究されている。それらの中でも uLSIF [11] は比較的計算量が少なく、広く利用されている。ただし、自然言語処理に限れば、bag-of-words の$P(\boldsymbol{x})$をNaive Bayesでモデル化できるので、論文 [13] では領域Rでのデータ$\boldsymbol{x}$が素性リスト$\{f_1, f_2, \cdots, f_n\}$であるとき$P_R(\boldsymbol{x}) = \prod_{i=1}^{n} P_R(f_i)$としている。また$P_R(f_i)$は以下で求めている。

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

ここで$n(R, f)$ は領域R内での素性fの頻度、$n(R)$ は領域Rのデータ数である。以上より確率密度比は以下となる。

$$r = \frac{P_T(\boldsymbol{x})}{P_S(\boldsymbol{x})} = \frac{n(T, f) + 1}{N(T) + 2} \cdot \frac{N(S) + 2}{n(S, f) + 1} \tag{5.1}$$

5.2　Naive Bayesの手法

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[12].

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

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．

$$w(\boldsymbol{x}) = \frac{P_T(\boldsymbol{x})}{P_S(\boldsymbol{x})} = \prod_{i=1}^{n} \left(\frac{n(T,f_i)+1}{N(T)+2} \cdot \frac{N(S)+2}{n(S,f_i)+1} \right)$$

5.3 uLSIFの手法

ソース領域内のデータを$\{\boldsymbol{x}_i^s\}_{i=1}^{N_s}$，ターゲット領域内のデータを$\{\boldsymbol{x}_i^t\}_{i=1}^{N_t}$とするuLSIF では確率密度比$w(\boldsymbol{x})$を以下の式でモデル化する．

$$\begin{aligned} w(\boldsymbol{x}) &= \sum_{l=1}^{b} \alpha_l \psi_l(\boldsymbol{x}) \\ &= \boldsymbol{\alpha} \cdot \boldsymbol{\psi}(\boldsymbol{x}) \end{aligned}$$

ただしここで，$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_b)$，$\boldsymbol{\psi}(\boldsymbol{x}) = (\psi_1(\boldsymbol{x}), \psi_2(\boldsymbol{x}), \cdots, \psi_b(\boldsymbol{x}))$である．また$\alpha_l$は正の実数であり，$\psi_l(\boldsymbol{x})$は基底関数と呼ばれるソース領域のデータ$\boldsymbol{x}$から正の実数値への関数である．uLSIF では，概略，自然数bと基底関数$\boldsymbol{\psi}(\boldsymbol{x})$を定めた後に，パラメータ$\boldsymbol{\alpha}$を推定する手順をとる．

説明の都合上，b と $\boldsymbol{\psi}(\boldsymbol{x})$が定まった後の$\boldsymbol{\alpha}$の推定を先に説明する．$w(\boldsymbol{x})$のモデルを$\hat{w}(\boldsymbol{x})$とおくと，パラメータ$\alpha_l$を推定するには，$w(\boldsymbol{x})$と$\hat{w}(\boldsymbol{x})$の平均2乗誤差$J_0(\boldsymbol{\alpha})$を最小にするような$\boldsymbol{\alpha}$を求めれば良い．$w(\boldsymbol{x}) = P_T(\boldsymbol{x})/P_S(\boldsymbol{x})$に注意すると，$J_0(\boldsymbol{\alpha})$ は以下のように変形できる．

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &= \frac{1}{2}\int (\hat{w}(\boldsymbol{x}) - w(\boldsymbol{x}))^2 P_S(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{1}{2}\int \hat{w}(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} - \int \hat{w}(\boldsymbol{x}) w(\boldsymbol{x}) P_S(\boldsymbol{x}) d\boldsymbol{x} + \frac{1}{2}\int w(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{1}{2}\int \hat{w}(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} - \int \hat{w}(\boldsymbol{x}) P_T(\boldsymbol{x}) d\boldsymbol{x} + \frac{1}{2}\int w(\boldsymbol{x})^2 P_S(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

3項目の式は定数なので，$J_0(\boldsymbol{\alpha})$を最小にするには，以下の$J(\boldsymbol{\alpha})$を最小にすればよい．

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\int \hat{w}(\boldsymbol{x})^2 P_S(\boldsymbol{x})d\boldsymbol{x} - \int \hat{w}(\boldsymbol{x})P_T(\boldsymbol{x})d\boldsymbol{x}$$

$J(\boldsymbol{\alpha})$を経験分布で近似した$\widehat{J}(\boldsymbol{\alpha})$は以下となる．

$$\begin{aligned}
\widehat{J}(\boldsymbol{\alpha}) &= \frac{1}{2N_s}\sum_{i=1}^{N_s}\widehat{w}(\boldsymbol{x}_i^s)^2 - \frac{1}{N_t}\sum_{j=1}^{N_t}\widehat{w}(\boldsymbol{x}_j^t) \\
&= \frac{1}{2}\sum_{l,l'=1}^{b}\alpha_l\alpha_{l'}\left(\frac{1}{N_s}\sum_{i=1}^{N_s}\psi_l(\boldsymbol{x}_i^s)\psi_{l'}(\boldsymbol{x}_i^s)\right) - \sum_{l=1}^{b}\alpha_l\left(\frac{1}{N_t}\sum_{j=1}^{N_t}\psi_l(\boldsymbol{x}_j^t)\right) \\
&= \frac{1}{2}\boldsymbol{\alpha}^T\widehat{H}\boldsymbol{\alpha} - \widehat{h}^T\boldsymbol{\alpha} \qquad (5.2)
\end{aligned}$$

ここで$\widehat{H}$は$b \times b$の行列であり，そのl行l'列の要素$\widehat{H}_{l,l'}$は以下である．

$$\widehat{H}_{l,l'} = \frac{1}{N_s}\sum_{i=1}^{N_s}\psi_l(\boldsymbol{x}_i^s)\psi_{l'}(\boldsymbol{x}_i^s)$$

また$\widehat{h}$はb次元のベクトルであり，そのl次元目の要素$\widehat{h}_l$は以下である．

$$\widehat{h}_l = \frac{1}{N_t}\sum_{j=1}^{N_t}\psi_l(\boldsymbol{x}_j^t)$$

$\widehat{J}(\boldsymbol{\alpha})$の最小値を求める際に正則化を行う．このとき付加する正則化項を L2 ノルムに設定し，$\boldsymbol{\alpha} > 0$の条件を外して，以下の最小化問題を解く．ここでパラメータλが導入されることに注意する．λは基底関数を設定する際に決められる．

$$\min_{\boldsymbol{\alpha}}\left[\frac{1}{2}\boldsymbol{\alpha}^T\widehat{H}\boldsymbol{\alpha} - \widehat{h}^T\boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha}\right]$$

この最小化問題は制約のない凸2次計画問題であるために，唯一の大域解が得られる．その解は以下である．

$$\tilde{\boldsymbol{\alpha}} = (\widehat{H} + \lambda I_b)^{-1}\widehat{h}^T \qquad (5.3)$$

最後に$\boldsymbol{\alpha} > 0$の条件に合わせるように，以下の調整を行う．

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= ((\max(0,\tilde{\alpha_1}),\max(0,\tilde{\alpha_2}),\cdots,\max(0,\tilde{\alpha_b})) \\ &= \max(0_b,\tilde{\boldsymbol{\alpha}}) \end{aligned} \tag{5.4}$$

パラメータbと基底関数の設定であるが，まず，bについては以下で設定する[1]．

$$b = \min(100, N_t)$$

次にターゲット領域のデータから重複を許さずにb個の点をランダムに取り出す．それらの点を$\{\boldsymbol{x}_j^t\}_{j=1}^b$ とおく．そして基底関数$\psi_l(\boldsymbol{x})$を以下のガウシアンカーネルで定義する．

$$\psi_l(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_l^t) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}_l^t||^2}{\sigma^2}\right)$$

以上より，確率密度比を求めるために残されているパラメータは正則化項の係数λとガウシアンカーネルの幅σの2つである．これらのパラメータはグリッドサーチの交差検定で求める．まずソース領域のデータとターゲット領域のデータをそれぞれ交わりのないR個の部分集合に分割する．それらの部分集合の中でr番目の部分集合を除き，残りを結合した集合を作る．それらを新たなソース領域のデータとターゲット領域のデータと見なす．そしてλとσをある値に設定し，式(5.3)と式(5.4)より$\boldsymbol{\alpha}$を求め，式(5.2)より$\widehat{J}(\boldsymbol{\alpha})^{(r)}$の値を求める．$r$を 1 から$R$まで変化させることで，$R$個の$\widehat{J}(\boldsymbol{\alpha})^{(r)}$の値が求まり，それらを平均した値を$\lambda$と$\sigma$に対する$\widehat{J}(\boldsymbol{\alpha})$の値とする．次に$\lambda$と$\sigma$を変化させ，上記手順で得られる$\widehat{J}(\boldsymbol{\alpha})$の値が最小となる$\hat{\lambda}$と$\hat{\sigma}$を求め，これを$\lambda$と$\sigma$の推定値とする．

[1]本実験ではbの値は最大 100 となるが，この 100 という数値はオリジナルの論文[6]で使われた値であり，本論文でのなんらかの予備実験から得た値ではない．uLSIF の実験結果はこの値を調整することで多少の向上があったかもしれない．

第6章　SVM

6.1　概要

SVM（upport vector machine）は、教師あり学習を用いるパターン認識モデルの一つである。分類や回帰へ適用できる。SVMは線形二値分類器であり,クラス数が2つであるような問題に用いられてきた.二つのクラスは,それぞれ正クラス及び負クラスを呼ばれ,正クラスに属する事例は正例と呼ばれ,負クラスに属する事例は負例と呼ばれる.

6.2　構造化SVM

2005年に Ioannis Tsochantaridis らが構造化SVM (英: structured SVM) を発表した[?] 。任意のデータ構造を扱えるように拡張したものである。

通常の二値分類SVMは以下の値で分類する。

$$\hat{y}(x;w) = \mathrm{sign}\langle w, x\rangle$$

これは、このようにも書ける。

$$\hat{y}(x;w) = \underset{y\in\{-1,1\}}{\arg\max}\ \langle w, yx\rangle$$

その上で、これを二値から一般の値に拡張する。 Ψは入出力から特徴量を作り出す実数ベクトルを返す関数。問題ごとに定義する。

$$\hat{y}(x;w) = \underset{y\in\mathcal{Y}}{\arg\max}\ \langle w, \Psi(x,y)\rangle$$

そして、下記の損失関数を最小化するように、最適化問題を解く。ここではL2正則化を付けている。 C は正則化の強さを表す定数。 Δ は出

力の類似度を表す実数を返す関数。問題ごとに定義する。 $\Delta(y, y) = 0$であり、異なる値同士なら 0 よりも大きくなるように設計する。

$$E(w) = \|w\|^2 + C\sum_{i=1}^{n} \Delta(y_i, \hat{y}(x_i; w))E(w) = \|w\|^2 + C\sum_{i=1}^{n} \Delta(y_i, \hat{y}(x_i; w))$$

上記の最適化問題を解くには工夫が必要であり、その後も提案が続いているが、2005年に提案された方法は下記のように上界となる関数$L_i(\boldsymbol{w})$を作る。

$$\Delta(y_i, \hat{y}(x_i; w)) \leq L_i(w)$$

その上で、下記の最適化問題を解く。

$$E(w) = \|w\|^2 + C\sum_{i=1}^{n} L_i(w)E(w) = \|w\|^2 + C\sum_{i=1}^{n} L_i(w)$$

$L_i(\boldsymbol{w})$の作り方として2通りが提案された。

マージン リスケーリング

$$L_i(w) = \sup_{y \in \mathcal{Y}} \Delta(y_i, y) + \langle w, \Psi(x_i, y)\rangle - \langle w, \Psi(x_i, y_i)\rangle$$

スラック リスケーリング

$$L_i(w) = \sup_{y \in \mathcal{Y}} \Delta(y_i, y)\left(1 + \langle w, \Psi(x_i, y)\rangle - \langle w, \Psi(x_i, y_i)\rangle\right)$$

第7章　提案手法

7.1　ターゲット領域におけるキーワードの度合い

ここでは単語wがターゲット領域においてどの程度キーワードとみなせるかという度合いd_xをwのターゲット領域の文書集合における IDF 値に設定する。

$$d_x = \log\left(\frac{N}{d_i}\right) + 1$$

ここでNはターゲット領域の文書集合の文書数、d_iはターゲット領域の文書集合の中で単語wを含む文書数である。

ターゲットの領域におけるどの単語wのキーワードの度合いを設定している。ターゲットの領域の文書におけるどの単語wのidfを求める。Nはターゲット領域の文書数、、d_iはターゲット領域の文章における含む単語w_iの文書の数量。、d_iが小さくなると、idfが増加してなる、この単語のキーワードの度合いが高める。

h_jはターゲット領域の文書とソース領域の文書中で含む単語w_jの数量。q_jとh_jのかけ算はTF-IDF用いた結果です。q_j h_jが大きくなると、単語のキーワードの度合いが増加してなる。この重みをつけた学習を行うことで、svmの正解率は高める。

7.2　ソース事例中のキーワードの含有率

ソース領域の事例xの重みw_xを定める。まず事例（文書）xの単語の集合を$\{w_i\}_{i=1}^{K}$とする。また単語w_iのx内での頻度をf_iとする。これらを用いて、w_xを以下の式で定める。

$$w_x = \frac{1}{\sum_{i=1}^{k} f_i} \sum_{i=1}^{K} f_i \cdot d_{w_i}$$

第8章　データセット

ここでは実験に使用したデータについての解説を行います.

実験では Amazon dataset [1]を用いる。具体的には以下のサイトで公開されている`processed_acl.tar.gz`を展開したデータを用いる。

`https://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

このデータは books (B), dvd (D), electronics (E) および kitchen (K) の4つの領域をもち、それぞれの領域に含まれる文書数は表8.1の通りである。どの領域でも positive データと negative データはそれぞれ 1,000 個あり、これらを合わせた 2,000 データをその領域の訓練データとする。

表 8.1: 領域毎の文書数

	positive	negative	test data
books	1,000	1,000	4,465
dvd	1,000	1,000	3,586
electronics	1,000	1,000	5,681
kitchen	1,000	1,000	5,945

第9章　実験

この章では，本論文で提案した手法を用いた実験について記述する．実験は英語キャプションに対する複数形表現の統一と日本語キャプションに対する複数形表現の統一の２つの実験を行った．

9.1　実験１：英語キャプションに対する複数形表現の統一

学習アルゴリズムには scikit-learn の SVM を用いる。カーネルは linear、C パラメータの値は 0.1 で固定とする。scikit-learn の SVM には重み付き学習もサポートされているので[1]、ここでの重み付き学習も scikit-learn の SVM を用いる。

領域適応としては、B→D、B→E、B→K、D→B、D→E、D→K、E→B、E→D、E→K、K→B、K→D、K→Eの12通りが存在する。それぞれの領域適応に対して、確率密度比を求める２つ手法（ uLSIF [11] と Naive Bayes を用いた式(5.1) ）及び提案手法を用いた結果(テストデータに対する正解率)を表9.1に示す。表9.1中の NONE は領域適応の手法を用いず、単にソース領域の訓練データから構築した分類器をそのままターゲット領域のテストデータに適用した結果である。また IDEAL はターゲット領域の訓練データを用いて分類器を学習し、それをターゲット領域のテストデータに適用した結果である。

9.1.1　素性ベース手法＋事例ベース手法

事例への重み付き手法である uLSIF、NB および提案手法を比較すると、12個の領域適応の中で 6個について提案手法が最も高い正解率を出

[1] http://scikit-learn.org/stable/auto_examples/svm/plot_weighted_samples.html

表 9.1: 実験結果

	IDEAL	NONE	uLSIF	NB	提案手法
B → D	0.822	0.806	0.806	**0.811**	0.809
B → E	0.852	0.761	0.756	0.755	**0.765**
B → K	0.878	0.845	0.778	0.779	**0.785**
D → B	0.831	0.762	0.733	**0.745**	0.741
D → E	0.852	0.761	0.748	0.753	**0.758**
D → K	0.878	0.795	0.773	0.782	**0.789**
E → B	0.831	0.712	0.714	**0.723**	0.719
E → D	0.822	0.722	0.708	**0.723**	0.714
E → K	0.878	0.849	0.854	**0.857**	0.855
K → B	0.831	0.713	0.707	0.714	**0.715**
K → D	0.822	0.740	0.733	0.723	**0.736**
K → E	0.852	0.842	0.847	**0.852**	0.845
平均	0.846	0.776	0.763	0.768	**0.769**

し、残りの 6個は NB が最も高い正解率を出している。12個の平均を取ると提案手法がわずかに NB を上回っており、提案手法は事例への重み付き手法としては、優秀であると言える。

表 9.2: 素性ベース手法＋事例ベース手法

	IDEAL	NONE	CORAL	提案手法	SCL	SCL + 提案手法
B → E	0.852	0.761	0.763	0.760	0.757	0.756
D → B	0.831	0.762	0.783	0.756	0.732	0.733
E → K	0.878	0.849	0.836	0.849	0.852	0.853
K → D	0.822	0.740	0.739	0.743	0.732	0.733
平均	0.846	0.778	0.780	0.777	0.768	0.769

第10章　考察

表9.1の NONE と事例への重み付き手法（uLSIF、NB および提案手法）を比較すると、明らかに NONE の方が正解率が高い。本実験データに限れば、事例ベースの手法は領域適応には効果がないと言える。

ただし事例ベースの手法は素性ベースの手法と容易に組み合わせられるという長所がある。ここでは論文[8] で行っている 4つの領域適応B→E、D→B、E→K、K→D に対して、最初に SCL を用いて訓練データの素性ベクトルを変換し、次にその変換されたベクトルに提案手法の重みを付けて学習を行う実験を行った。その結果を表9.2に示す。表9.2の CORAL は[8]の表から取り出した。

表9.2を見ると SCL はあまり効果がなく、SCL に提案手法を組み合わせた手法の精度は良くない。しかし SCL に提案手法を組み合わせた場合、SCL 単独の精度を改善できており、素性ベースの手法に事例ベースの手法を組み合わせた効果が確認できる。素性ベースの手法には SCL 以外にも多くの手法があるので、それらの手法と提案手法を組み合わせて改良できると考えている。

また本論文では重み付き学習には重み付き SVM を利用したが、ニューラルネットでは損失関数の損失値に重みを乗じて、それを損失値とすることで容易に重み付き学習が実現できる。深層学習を利用した領域適応の手法は数多く提案されており、それらと事例ベースの手法を組み合わせることも容易である。

チューニングが十分でなく、良い値は得られてはいないが、素性ベースの手法と事例ベースの手法を組み合わせることが、ニューラルネットワークで容易に実現できることがわかる。今後はこの枠組みでの領域適応の手法を考案したい。

簡単な例として、ソース領域のデータとターゲット領域のデータから AutoEncoder により次元縮約を行い、次元縮約したデータによりニューラルネットワークで学習を行うことを試してみる。学習では前述したように損失関数の損失値に本論文の提案手法から得られた重みを乗じて、そ

れを損失値とした（図10.1参照）。B → Eのみの実験であるが、表10.1と図10.2の結果が得られた。なお、この実験でのニューラルネットワークの学習は 25 epoch で終了させており、正解率は 25 epoch 後の学習によって得られたモデルをテストデータで評価した結果である。また次元縮約では 400次元に縮約している。

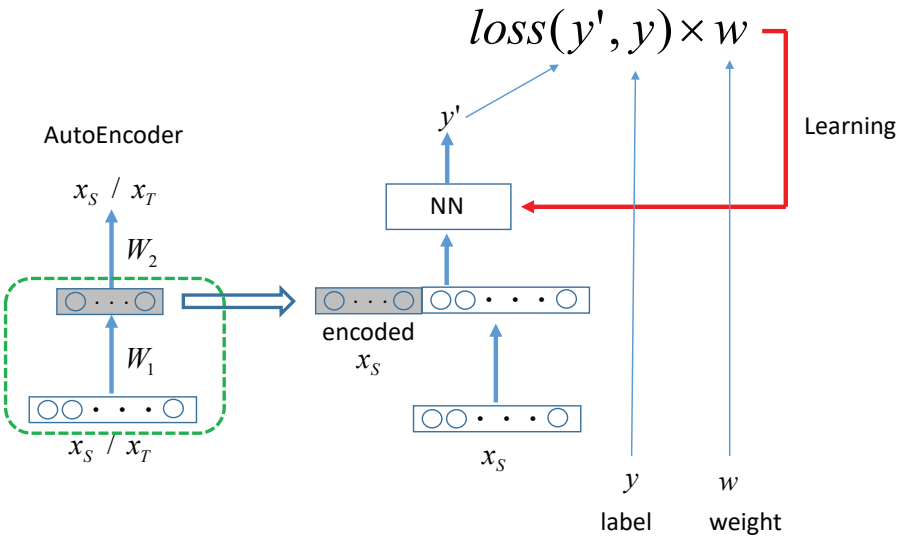

図 10.1: AE+NN+重み付き学習

表 10.1: ニューラルネットワークによる重み付き学習

IDEAL	NONE	NN	AE+NN	AE+NN+重み付け
0.852	0.761	0.7618	0.7667	0.7697

チューニングが十分でなく、良い値は得られてはいないが、素性ベースの手法と事例ベースの手法を組み合わせることが、ニューラルネットワークで容易に実現できることがわかる。今後はこの枠組みでの領域シフトに対する手法を考案したい。

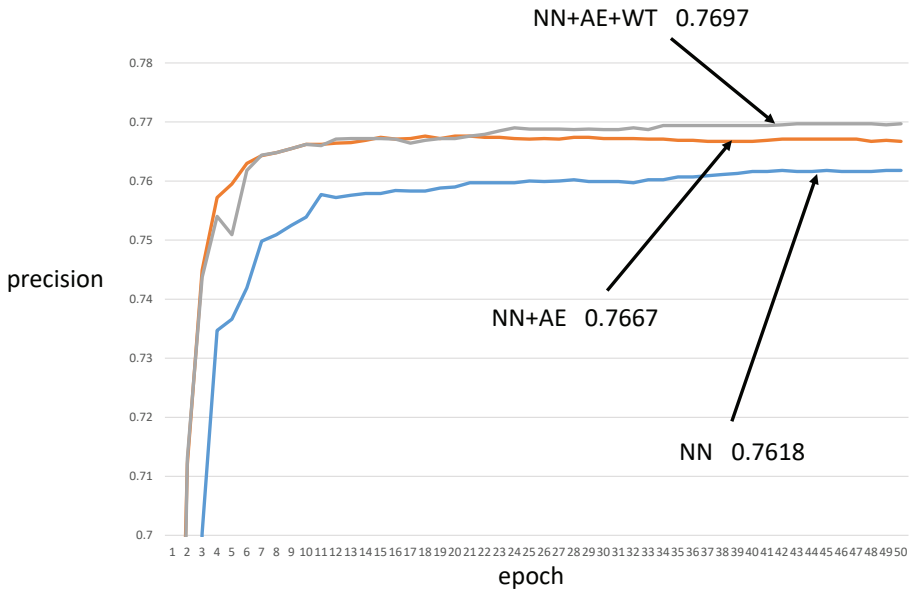

図 10.2: ニューラルネットワークによる重み付き学習

第11章　結論

本論文では感情分析の領域適応に対して、新たな事例ベースの手法を提案した。概略、ターゲット領域から TF-IDF を利用してキーワードの度合いを定め、ソース領域におけるデータがどの程度ターゲット領域のキーワードを含むかというキーワード含有率を求め、そのキーワード含有率をデータの重みとする。実験では確率密度比を求める 2 つ手法と比較することで、提案手法の優位性を示した。ただし事例ベースの手法単独では領域適応に対して効果は少なく、今後素性ベースの手法を組み合わせることが有効だと考える。その際にニューラルネットワークでの組み合わせが容易に実現できるため、今後はこの方向で研究を進めたい。

謝辞

修士研究を進めるにあたり，熱心にご指導いただいた情報工学科の新納教授に感謝いたします．また，多くのご指摘をいただきました自然言語処理研究室の皆様に感謝します．

参考文献

[1] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain adaptation for Sentiment Classification. In *ACL-2007*, pp. 440–447, 2007.

[2] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *EMNLP-2006*, pp. 120–128, 2006.

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

[4] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pp. 1180–1189, 2015.

[5] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *ICML-11*, pp. 513–520, 2011.

[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] 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.

[8] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of Frustratingly Easy Domain Adaptation. *AAAI*, 2016.

[9] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops*, pp. 443–450, 2016.

[10] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*, 2017.

[11] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, Vol. 25, No. 5, pp. 1370–1370, 2011.

[12] 高村大也. 言語処理のための機械学習入門. コロナ社, 2010.

[13] 新納浩幸, 佐々木稔. 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, Vol. 21, No. 1, pp. 61–79, 2014.