生成モデルと識別モデルを共用する転移学習法に関する研究

1X14C112-5　降矢 佳代子
指導教員　後藤 正幸

1　研究背景・目的

自動文書分類とは，所属するカテゴリが既知な学習用文書（以下，ラベルありデータ）集合を用いて識別関数を学習し，その関数に従って分類対象であるカテゴリが未知な新規文書（以下，ラベルなしデータ）を分類する問題である．その際，実問題では，学習に用いるラベルありデータ集合と分類対象のラベルなしデータ集合の単語出現頻度分布（以下，単語分布）が異なることがある．カテゴリが既知の過去の新聞記事を学習した分類器を用いて，カテゴリが未知の新しい記事を分類するような問題である．例えば，2010 年の記事で分類器を構築し，2015 年の記事を分類する場合，2011 年の東日本大震災を背景に，2010 年には低頻度であった津波や原子力発電所に関連する単語が，2015 年には多く出現するなど，それぞれの記事の単語分布は異なる．一方，文書分類における一般的な学習法は，ラベルありデータ集合とラベルなしデータ集合で単語分布が類似した場合を想定している．そのため，一般的な学習法を単語分布が異なる実問題に適用した場合，学習を上手く行えず，分類精度低下の恐れがある．

上記の課題に対して，異なる単語分布を持つラベルあり・なしデータ両方を利用する転移学習 [1] という枠組みがあり，その一手法として，Maximum Hybrid Log-likelihood Expectation（以下，MHLE）が提案されている [2]．しかし，MHLE ではモデルを学習する際に，ラベルありとラベルなしのデータ数が等しい状況を暗に仮定している．実問題では，ラベルありデータの作成には人手でのラベル付けが必要であり，大量の付与は困難なため，ラベルありデータがラベルなしデータよりも少ない場合も考えられる．特に，新聞記事において，オンライン記事の普及によって，新規のラベルなし記事の数は増加している．そのため，このような場合，MHLE ではラベルなしデータと比べラベルありデータの学習が十分に行われず，分類精度が低くなる恐れがある．

そこで本研究では，ラベルありデータがラベルなしデータより相対的に少ない問題に対し，学習に対する重みを導入することで，双方のデータをバランス良く学習に寄与させる手法を提案し，分類精度の向上を目指す．また，新聞記事データを用いて検証を行い，本手法の有効性を示す．

2　準備

2.1　文書分類

各文書の単語の出現頻度ベクトルを $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_V)^{\mathrm{T}}$，文書のカテゴリラベルを $y \in \{1, \ldots, k, \ldots, K\}$ とする．ただし，V は文書集合全体に含まれる語彙の総数とする．また，予めカテゴリラベルが与えらている N 件からなるラベルありデータ集合を $D_l = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ とする．ここで，y_n はラベルありデータ $\boldsymbol{x}_n$ が所属するカテゴリである．一方，分類対象であるラベルなしデータ集合を $D_u = \{\boldsymbol{x}_m\}_{m=1}^{M}$ とする．このラベルなしデータ $\boldsymbol{x}_m$ を，カテゴリ y のいずれかに所属するものとし，カテゴリ y への事後確率 $p(y|\boldsymbol{x}_m)$ が最大となるカテゴリ $\hat{y}$ へ分類する．

文書分類のモデルには，生成モデルと識別モデルの 2 つのアプローチがある．生成モデルとはデータ $\boldsymbol{x}$ とラベル y の同時確率分布 $p(\boldsymbol{x}, y)$ を予め仮定し，その分布を推定するモデルであり，一般にラベルなしデータの分布の分析に有効である．一方，識別モデルとはデータ $\boldsymbol{x}$ から事後確率 $P(y|\boldsymbol{x})$ を推定するモデルであり，一般にラベルありデータの分類に有効である．

2.2　転移学習

転移学習とは，分類したい対象とは本来は異なるデータによって学習された知識も再利用して，分類器を構成する方法である．本研究では，学習用のラベルありデータの所属領域（以下，元ドメイン），分類したいラベルなしデータの所属領域（以下，目標ドメイン）に関して，次の 4 点を仮定する．i) 目標ドメインと元ドメインの特徴空間は同一とする．ii) 目標ドメインと元ドメインのデータ分布は異なるものとする．iii) 目標ドメインと元ドメインのカテゴリラベル集合は同一とする．iv) データがカテゴリに属する条件付き確率分布は，目標ドメインと元ドメイン間で異なるが，高い類似性はあるとする．

2.3　MHLE

MHLE は，異なるドメインに属するラベルあり・なしデータの両方を用いて，生成モデルと識別モデルを共用し学習する手法である．これにより，ラベルなしデータの分類に利用できる情報や知識を，異なるドメインのラベルありデータから転移させる．

生成モデルの学習には，最大事後確率推定を適用する．ラベルなしデータはカテゴリラベルが未知であるため，ラベルなしデータ $\boldsymbol{x}_m$ のカテゴリラベルが k である確率 $P(k|\boldsymbol{x}_m)$ を導入する．そして，生成モデルのパラメータを $\boldsymbol{\Theta}$，生成モデルを $p_g(\boldsymbol{x}, y; \boldsymbol{\Theta})$ としたとき，$P(k|\boldsymbol{x}_m)$ による対数尤度の重み付き和に基づく目的関数 $J_g(\boldsymbol{\Theta})$ の最大化によって生成モデルを学習する．一方，識別モデルの学習には，生成モデルの学習のために導入した $P(k|\boldsymbol{x}_m)$ を用いて識別モデルのパラメータ $\boldsymbol{W}$ を推定する．そして，識別モデルを $P_d(y|\boldsymbol{x}; \boldsymbol{W})$ としたとき，ラベルありデータの条件付き対数尤度の最大化と，$P(y|\boldsymbol{x}_m)$ と $P_d(y|\boldsymbol{x}_m; \boldsymbol{W})$ の Kullback-Leibler 情報量最小化に基づく目的関数 $J_d(\boldsymbol{W})$ を最大化させる $\boldsymbol{W}$ を求める．ここで，$J_g(\boldsymbol{\Theta})$, $J_d(\boldsymbol{W})$ はそれぞれ式 (1), (2) で算出される．ただし，$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_K]^{\mathrm{T}}$, $\boldsymbol{W} = [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k, \ldots, \boldsymbol{W}_K]^{\mathrm{T}}$, $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{ki}, \ldots, \theta_{kV})^{\mathrm{T}}$ である．また，θ_{ki} はカテゴリ k での i 番目の単語の生起確率を表し，$\|\boldsymbol{\theta}_k\|_1 = 1$ とする．

$$J_g(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \log p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n}) + \sum_{m=1}^{M} \sum_{k=1}^{K} P(k|\boldsymbol{x}_m) \log p_g(x_m, k; \boldsymbol{\theta}_k) + \log p(\boldsymbol{\Theta}) \quad (1)$$

$$J_d(\boldsymbol{W}) = \sum_{n=1}^{N} \log P_d(y_n|\boldsymbol{x}_n; \boldsymbol{W}) - \sum_{m=1}^{M} \sum_{k=1}^{K} P(k|\boldsymbol{x}_m) \log \frac{P(k|\boldsymbol{x}_m)}{P_d(k|\boldsymbol{x}_m; W)} + \log p(\boldsymbol{W}) \quad (2)$$

生成モデルにはナイーブベイズモデル，$\boldsymbol{\Theta}$ の事前確率分布にはディレクレ事前確率分布を適用できる．一方，識別モデルには多項ロジスティック回帰モデル，$\boldsymbol{W}$ の事前確率分布 $p(\boldsymbol{W})$ にはガウス事前確率分布を適用できる．

ここで，$J_g(\boldsymbol{\Theta})$ を最大化させる $\boldsymbol{\Theta}$ と $J_d(\boldsymbol{W})$ を最大化させる $\boldsymbol{W}$ は，$J_g(\boldsymbol{\Theta})$ と $J_d(\boldsymbol{W})$ の和を最大化させる $\boldsymbol{\Theta}$ と $\boldsymbol{W}$ と同値である．そこで，$J_g(\boldsymbol{\Theta})$ の統合重みを β とし，$J_g(\boldsymbol{\Theta})$ と $J_d(\boldsymbol{W})$ の重み付き和 $J(\boldsymbol{W}, \boldsymbol{\Theta})$ を式 (3) で定義する．$J(\boldsymbol{W}, \boldsymbol{\Theta})$ を最大化させる $\boldsymbol{W}$ と $\boldsymbol{\Theta}$ を，EM アルゴリズムにより推定する．

$$J(\boldsymbol{W}, \boldsymbol{\Theta}) = J_d(\boldsymbol{W}) + \beta J_g(\boldsymbol{\Theta}) \quad (3)$$

3 提案手法

3.1 概要

MHLE では，ラベルあり・なしデータの両方を生成・識別モデルの双方の学習に用いることで，他の転移学習手法と比較し，分類精度が向上することが実験的に示されている．また，全データで双方のモデルを学習するため，ラベルあり・なしデータの数が目的関数に大きく影響する．現状の手法では，元ドメインと目標ドメインのデータ数 N, M が同程度の場合に有効な手法であると考えられる．一方，ラベルありデータの数がラベルなしデータの数と比べて少ない場合には，ラベルなしデータが学習に強く寄与し，相対的にラベルありデータのカテゴリ情報が十分に学習されない可能性がある．また，データ数を等しくするために，データ数が多いラベルなしデータをサンプリングにより減らした場合，単語情報が少なくなり，分類精度低下の恐れがある．そこで，本研究ではラベルありデータとラベルなしデータの目的関数への寄与度合いを等しくすることで，この問題の解決を図る．具体的には，ラベルありデータ数とラベルなしデータ数をそれぞれ N, M とするとき，ラベルありデータの目的関数の重みを M，ラベルなしデータの目的関数の重みを N とする．これにより，各データのデータ数による学習への寄与度を等しくすることで，分類精度の向上を図る．

3.2 モデル式

式 (1), (2) におけるラベルありデータの目的関数の和に重み M，ラベルなしデータの目的関数の和に重み N をそれぞれ乗じる．このとき，提案する生成モデルの目的関数を $J'_g(\boldsymbol{\Theta})$，識別モデルの目的関数を $J'_d(\boldsymbol{W})$ とおくと，式 (4), (5) が与えられる．

$$J'_g(\boldsymbol{\Theta}) = M\sum_{n=1}^{N}\log p_g(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}_{y_n}) + N\sum_{m=1}^{M}\sum_{k=1}^{K} P(k|\boldsymbol{x}_m)\log p_g(x_m, k; \boldsymbol{\theta}_k) + \log p(\boldsymbol{\Theta}) \quad (4)$$

$$J'_d(\boldsymbol{W}) = M\sum_{n=1}^{N}\log P_d(y_n|\boldsymbol{x}_n; \boldsymbol{W}) - N\sum_{m=1}^{M}\sum_{k=1}^{K} P(k|\boldsymbol{x}_m)\log\frac{P(k|\boldsymbol{x}_m)}{P_d(k|\boldsymbol{x}_m; \boldsymbol{W})} + \log p(\boldsymbol{W}) \quad (5)$$

そして，$J'_d(\boldsymbol{W})$ と $J'_g(\boldsymbol{\Theta})$ の重み付き和 $J'(\boldsymbol{W}, \boldsymbol{\Theta})$ を最大化させるパラメータ $\Psi = \{\boldsymbol{W}, \boldsymbol{\Theta}\}$ を，以下に示す EM アルゴリズムの繰り返し計算により推定する．

E-step)

$$P(y|\boldsymbol{x}_m; \boldsymbol{W}, \boldsymbol{\Theta}, \beta) = \frac{P_d(y|\boldsymbol{x}_m; \boldsymbol{W})p_g(\boldsymbol{x}_m, y; \boldsymbol{\theta}_y)^\beta}{\sum_{k=1}^{K} P_d(k|\boldsymbol{x}_m; \boldsymbol{W})p_g(\boldsymbol{x}_m, k; \boldsymbol{\theta}_k)^\beta} \quad (6)$$

M-step)

$$\boldsymbol{\Theta}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} q_g(\boldsymbol{\Theta}, \Psi^{(t)}) \quad (7)$$

$$\boldsymbol{W}^{(t+1)} = \arg\max_{\boldsymbol{W}} q_d(\boldsymbol{W}, \Psi^{(t)}) \quad (8)$$

ただし，式 (7) 中の $q_g(\boldsymbol{\Theta}, \Psi^{(t)})$ は $J'_g(\boldsymbol{\Theta})$ に繰り返し計算の t 回目での Ψ の推定値 $\Psi^{(t)}$ を用いた関数，式 (8) 中の $q_d(\boldsymbol{W}, \Psi^{(t)})$ は $J'_d(\boldsymbol{W})$ に $\Psi^{(t)}$ を用いた関数である．また，以上の提案手法は，$N = M$ のとき，従来の MHLE と等価である．

4 実験

4.1 実験条件

2010 年と 2015 年の読売新聞記事データセットを用いて実験を行う．カテゴリは，政治，経済，社会，スポーツ，文化，生活，犯罪・事件，科学の計 8 カテゴリである．語彙集合には全体で出現頻度が 100 以上の単語を用い，総数 $V = 3152$ である．元ドメインを 2010 年，目標ドメインを 2015 年の記事とする．各実験で両ドメイン 12,000 件ずつから表 1 に示した記事数を無作為に抽出し，双方の記事で学習とテストを行う．ただし，元ドメインの記事は全実験で同一とし，それに対して目標ドメインの記事数が異なる 5 通りの状況を考え，各実験を 10 回ずつ行う．汎化性能を従来手法（MHLE）と比較するため，目標ドメインだけでなく，元ドメインの記事に対しても，分類評価を行う．さらに，事前実験により，重みパラメータ $\beta = 0.5$ とする．評価指標として式 (9) に示す分類精度を用いる．

$$\text{分類精度} = \frac{\text{正しく分類された記事数}}{\text{全記事数}} \quad (9)$$

表 1. 各実験の各ドメインの記事数 [件]

No.	元ドメイン	目標ドメイン
1	4000	4000
2	4000	6000
3	4000	8000
4	4000	10000
5	4000	12000

4.2 実験結果と考察

表 2 に，各実験における 10 回の平均分類精度を示す．

表 2. 各実験における平均分類精度

	元ドメイン		目標ドメイン	
No.	従来	提案	従来	提案
1	0.7728	0.7728	0.7040	0.7040
2	0.7065	0.7913	0.6468	0.7169
3	0.6991	0.8006	0.6321	0.7195
4	0.6868	0.8225	0.6256	0.7340
5	0.6216	0.8332	0.5686	0.7361

表 2 より，ラベルありデータがラベルなしデータより少ない場合（No.2〜5），両ドメインとも提案の精度が従来を上回っていることがわかる．すなわち，提案手法では，元ドメインに対する汎化性能を維持しつつ，目標ドメインでも高い汎化性能を持つモデルが構築できたといえる．また，提案手法においては，目標ドメインの記事数が多いほど，分類精度が高くなることがわかる．これらは，提案手法では従来手法と比べ，ラベルありデータによってカテゴリ情報を十分に学習できたためと考えられる．

5 まとめと今後の課題

本研究では，MHLE に対して，ラベルありデータがラベルなしデータよりも少ない場合に，データ数の偏りを考慮する手法を提案した．また，新聞記事データを用いた実験により提案手法が従来手法に比べ，分類精度の観点で優れていることを示した．今後の課題として，最適な生成モデルと識別モデルに関する研究などが挙げられる．

参考文献

[1] 神嶌敏弘,“転移学習,” 人工知能学会誌, Vol.25, No.4, pp.572–580, 2010.

[2] 藤野昭典, 上田修功, 永田昌明,“ラベルありデータの選択バイアスに頑健な半教師あり学習,” 情報処理学会論文誌 数理モデル化と応用, Vol.4, No.2, pp.31–42, 2011.