4K1-1in2

特徴変換と重み付けを併用したドメイン適応

Joint Optimization of Feature Transform and Instance Weighting for Domain Adaptation

石井 雅人　　佐藤 敦
Masato ISHII　　Atsushi SATO

NEC 情報・メディアプロセッシング研究所
NEC Information and Media Processing Laboratories

サンプルごとの重み付けと特徴変換を同時に行うドメイン適応手法を提案する。両者を同時に最適化することで、ターゲットデータに関連の無いソースデータを効率的に排除しつつ、ドメイン適応に適した特徴変換を得る。さらに、重みを均一にする正則化を導入することで、なるべく多くのソースデータを適応し、後段の識別学習への悪影響を低減する。実際の映像監視システムで得られたデータで実験を行い、本手法の有効性を示す。

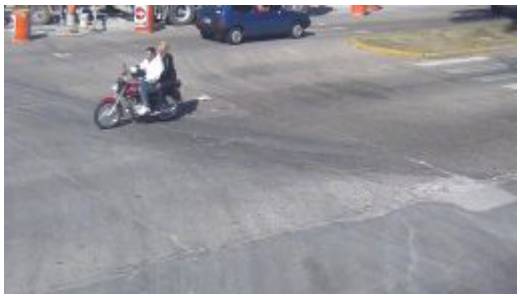
1. はじめに

多くの機械学習手法では、学習データの特性が、実際に認識処理が適用される評価データの特性に十分近いことを仮定している。しかし、実用においては、学習データを収集した環境と評価データが得られる環境が異なることが多く、この仮定が必ずしも成り立たない。監視映像の例を図 1 に示す。図 1(a) と図 1(b) は異なる 2 地点の監視カメラから撮影された画像であるが、監視カメラの位置・姿勢の違いや地理的な違いによって、物体の画像の特性が若干異なることが分かる。図 1(a) の映像で識別学習を行い、図 1(b) の映像で評価すると、この特性の違いによって識別精度は大きく低下する。このような問題を回避するため、ドメイン適応と呼ばれる技術が提案されている。

ドメイン適応は、適応元のソースデータの分布が適応先のターゲットデータの分布に合うようにデータを変換する処理である。学習データをソースデータ、評価データをターゲットデータとしてドメイン適応を行うことで、学習データと評価データの特性の違いを低減し、学習データで学習した識別モデルの評価データに対する識別精度を向上させることができる。既存のドメイン適応は大きく 2 つのアプローチに分類できる。1 つはサンプル重み付けによる方法 [Shimodaira 00, Huang 06, Kanamori 09]、もう 1 つは特徴変換による方法 [Saenko 10, Gong 12, Fernando 13] である。サンプル重み付けによる方法では、ソースデータに対してサンプルごとの重み付けを行うことによってデータの分布を合わせる。この手法は、ターゲットデータに適応できないソースデータのサンプルを効率的に排除できる利点があるが、ソースデータの実効的なサンプル数を低減するため、適応後の識別学習に悪影響を与えてしまう場合がある。一方、特徴変換による手法では、特徴変換を行うことでデータの分布を合わせる。重み付けによる手法とは異なり、全てのソースデータを適応できる利点があるが、ターゲットデータに適応できないソースデータが無理に適応され、ドメイン適応の精度が下がるという問題がある。このように、ドメイン適応では、ターゲットデータに適応できないソースデータの排除と、適応後のソースデータの実効的なデータ数の維持との間にトレードオフが存在する。しかし、このトレードオフは、従来の研究において明示的には取り扱われてこなかった。

本稿では、特徴変換と重み付けを併用したドメイン適応手法を提案する。両者を併用することによって、ターゲットデータに適応できないソースデータを効率的に排除しつつ、ドメイン適応に適した特徴変換を求める。さらに、サンプル重みを均一にする正則化を導入することにより、適応後のソースデータの実効的なデータ数ができるだけ維持され、適応後の識別学習において多くの学習データを用いることができる。実際の映像監視システムで得られたデータを用いた実験で、本手法が有効に働くことを示す。

(a) A地点の監視映像の例



(b) B地点の監視映像の例

図 1: 監視映像の例

連絡先: NEC 情報・メディアプロセッシング研究所　石井雅人
m-ishii@nd.jp.nec.com

2. 提案手法

2.1 提案手法の定式化

ソースデータを $\{\mathbf{x}_i^{(S)}\}(i=1,...,N_s)$、ターゲットデータを $\{\mathbf{x}_j^{(T)}\}(j=1,...,N_t)$ とし、特徴変換 $f_\theta(\mathbf{x})$ のパラメータを θ、ソースデータの各サンプルの重みを w_i とする。提案手法では、ソースデータへの重みをできるだけ均一に保ちながら、

変換後の重み付きソースデータ $\{w_i f_\theta(\mathbf{x}_i^{(S)})\}$ と変換後のターゲットデータ $\{f_\theta(\mathbf{x}_j^{(T)})\}$ が同じ分布となるように θ と $\{w_i\}$ を学習する。具体的には以下の最適化問題を解くことによって θ と $\{w_i\}$ を得る。

$$\min_{\{w_i\},\theta}\left[\sum_i^{N_s} w_i L(\mathbf{x}_i^{(S)}|\theta) + \sum_j^{N_t} L(\mathbf{x}_j^{(T)}|\theta) + \lambda\sum_i^{N_s} w_i^2\right],$$
$$\text{subject to } \sum_i^{N_s} w_i = N_s, \forall i\ w_i \geq 0. \quad (1)$$

ここで、$L(\mathbf{x}|\theta)$ は特徴変換 f_θ を用いた時のサンプル $\mathbf{x}$ に関する損失である。損失の定義は、従来の feature learning で採用されているものを用いることができ、一般的には再構成誤差が良く用いられる [Vincent 10, Hinton 02]。特徴変換として線形変換を使った場合の具体例は 2.4 節で述べる。式 (1) の第 1 項はソースデータに対する損失の重み付き和、第 2 項はターゲットデータに対する損失の総和である。これらを最小化することで、重み付きソースデータとターゲットデータの双方の情報を良く記述する特徴変換を学習できる。変換後の特徴次元数が十分に小さければ、双方の分布は特徴変換上で大きく重なることが期待され、ドメイン適応を実現できる。式 (1) の第 3 項は重みに関する正則化項であり、λ は正則化の強さを決めるハイパーパラメータである。拘束条件により、重みは平均 1 となる非負の値であるため、λ を大きくすると重みが均一に近づく。これにより、一部の学習データのみに大きく重みが割り当てられて実効的なソースデータ数が大きく減少することを防ぐ。したがって、式 (1) の最小化により、実効的なソースデータの数をできるだけ保ったまま、ドメイン適応に適した特徴変換を求めることができる。式 (1) の最小化は、$\{w_i\}$ に関する最小化と θ に関する最小化を交互に繰り返すことで行う。以下では、それぞれの最適化方法について述べる。

2.2 サンプルの重み $\{w_i\}$ に関する最適化

式 (1) の目的関数はそれぞれの w_i に対して 2 次であり、拘束条件は w_i に対して線形である。したがって、式 (1) の $\{w_i\}$ に関する最小化は、ラグランジュの未定乗数法を用いて解析的に解くことができる。未定乗数 α, $\{\beta_i\}(i = 1, ..., N_s)$ を導入すると、式 (1) の双対問題は以下のように得られる。

$$\max_{\alpha,\{\beta_i\}} \phi(\alpha, \{\beta_i\}), \text{ subject to } \forall i\ \beta_i \geq 0 \quad (2)$$
$$\phi(\alpha, \{\beta_i\}) = \min_{\{w_i\}} \psi(\{w_i\}, \alpha, \{\beta_i\}) \quad (3)$$
$$\psi(\{w_i\}, \alpha, \{\beta_i\}) = \sum_i^{N_s} w_i L(\mathbf{x}_i^{(S)}|\theta) + \lambda\sum_i^{N_s} w_i^2 + \alpha\left(\sum_i^{N_s} w_i - N_s\right) - \sum_i^{N_s}\beta_i w_i. \quad (4)$$

上記の双対問題を解くことにより、最適な重み $\{w_i^*\}$ は以下のように得られる。

$$w_i^* = \begin{cases} 0 & \text{if } \beta_i > 0 \\ 1 - \frac{1}{2\lambda}\left(\overline{\beta} + s_i\right) & \text{if } \beta_i = 0 \end{cases} \quad (5)$$
$$\beta_i = \max\left(0, \overline{\beta} + s_i - 2\lambda\right) \quad (6)$$
$$\overline{\beta} = \frac{1}{N_S}\sum_i \beta_i \quad (7)$$
$$s_i = L(\mathbf{x}_i^{(S)}|\theta) - \frac{1}{N_S}\sum_i L(\mathbf{x}_i^{(S)}|\theta). \quad (8)$$

Algorithm 1 線形変換を用いた場合の提案手法

Require: $\{\mathbf{x}_i^{(S)}\}$, $\{\mathbf{x}_j^{(T)}\}$, D, and λ.
  Set all w_i to 1.
  repeat
    Calculate weighted covariance matrix of $\{\mathbf{x}_i^{(S)}\}$ and $\{\mathbf{x}_j^{(T)}\}$ based on the fixed $\{w_i\}$.
    Obtain D eigenvectors $\mathbf{v}_d (d = 1, ..., D)$ corresponding to D largest eigenvalues of the above covariance matrix.
    Set $P = \{\mathbf{v}_1, ..., \mathbf{v}_D\}^T$.
    Calculate $\{w_i\}$ based on the fixed P (Eq. (5)-(8) and (10))
  until convergence
  return P and $\{w_i\}$

式 (6) の β_i と式 (7) の $\overline{\beta}$ は互いに依存しているため、反復計算が必要となる。収束に必要な反復回数は、実験的には多くの場合数回、多くても数十回である。

2.3 特徴変換のパラメータ θ に関する最適化

式 (1) の正則化項と拘束条件は θ に依存しないため、式 (1) の θ に関する最適化は以下のように書ける。

$$\min_\theta \left[\sum_i^{N_s} w_i L(\mathbf{x}_i^{(S)}|\theta) + \sum_j^{N_t} L(\mathbf{x}_j^{(T)}|\theta)\right]. \quad (9)$$

上式の最適化はサンプルごとの損失の全ての総和の最小化である。これは、ソースデータに関する損失に重み付けがされていることを除けば、従来の feature learning において多く採用されている目的関数の最適化である。したがって、それらと同様の最適化手法によって最適な θ を得ることができる。

2.4 線形変換を用いた場合

提案手法では様々な特徴変換を用いることができるが、本稿では、最も単純な特徴変換として線形変換を用いる。この時、特徴変換のパラメータ θ は射影行列 P である。また、損失 $L(\mathbf{x}|P)$ として以下の再構成誤差を用いる。

$$L(\mathbf{x}|P) = \|\mathbf{x} - P^T P\mathbf{x}\|^2. \quad (10)$$

ただし、このままでは、一般に $R^T R = I$ を満たす R について $L(\mathbf{x}|P) = L(\mathbf{x}|RP)$ となり、最適解が無限に存在してしまう。そこで、以下の拘束条件を追加する。

$$PP^T = I. \quad (11)$$

ここで、式 (10) を式 (9) に代入すると、P に関する最適化は以下のようになる。

$$\min_P\left[\sum_i^{N_s} w_i\|\mathbf{x}_i^{(S)} - P^T P\mathbf{x}_i^{(S)}\|^2 + \sum_j^{N_t}\|\mathbf{x}_j^{(T)} - P^T P\mathbf{x}_j^{(T)}\|^2\right]$$
$$\text{subject to } PP^T = I. \quad (12)$$

上式の最適化問題は、ソースデータへの重み付けを除けば、主成分分析で解く最適化問題と同じである。したがって、この最適化問題は、射影先の次元数を D とすると、重み付き共分散行列の固有値分解を行い、固有値が大きい順に D 個の固有ベクトルを並べた行列を求めることで解くことができる。特徴変換として線形変換を用いた場合の提案手法の処理の流れを Algorithm 1 に示す。

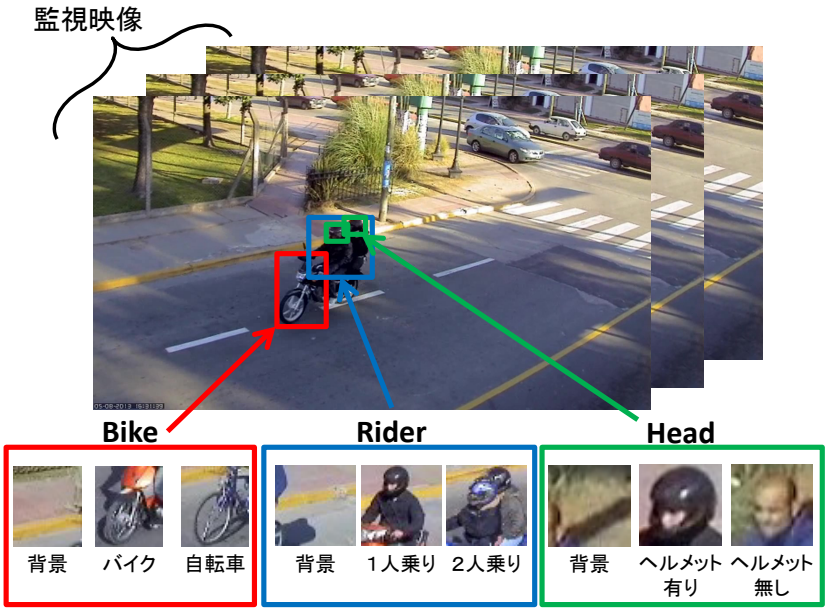


図 2: 実験に用いた監視映像と識別タスク

(a) Bike タスクの画像例（上段：学習データ、下段：評価データ）

(b) Head タスクの画像例（上段：学習データ、下段：評価データ）

(c) Rider タスクの画像例（上段：学習データ、下段：評価データ）

図 3: 実験に用いた画像の例

3. 実験

3.1 実験の概要

提案手法の有効性を示すため、図 1 に例を示した実際の監視映像を用いた実験を行った。2 つの地点（A 地点と B 地点）で撮影された監視映像を用意し、A 地点で撮影された映像を学習データ、B 地点で撮影された映像を評価データとして用いた。識別タスクとして、3 種類の 3 クラス識別タスクを行った。

Bike (バイク検知) バイク／自転車／背景の識別
Rider (2 人乗り検知) 1 人乗り／2 人乗り／背景の識別
Head (ノーヘル検知) ヘルメット有りの頭部／ヘルメット無しの頭部／背景の識別

表 1: 実験に用いたデータセット

データ名	特徴次元数	学習データ数	評価データ数
Bike set1	270	1779	1443
Bike set2	270	1661	1404
Head set1	225	1859	1486
Head set2	225	1683	1417
Rider set1	324	1758	1440
Rider set2	324	1665	1403

監視映像から作成されるデータと識別タスクの関係を図 2 に示す。実験では、ドメイン適応によって、A 地点の映像で学習した識別器が B 地点の映像における識別精度を向上させる。図 3 に各タスクのパターンの例を示す。特徴量には輝度勾配特徴を用い、識別器には改良版 GLVQ [Sato 13] を用いた。実験では、ドメイン適応なし、GFK [Gong 12]、subspace alignment [Fernando 13]、提案手法の 4 手法を比較した。ドメイン適応および識別器のハイパーパラメータは、学習データを用いた 5 分割交差検証によって決定した。実験に用いるデータセットは各タスクについて set1 と set2 の 2 つ作成し、その平均識別誤り率で性能を比較した。各データセットの詳細を表 1 に示す。

3.2 実験結果

表 2 に各ドメイン適応手法を用いた時の平均識別誤り率、図 4 にドメイン適応後で特徴変換を行った後の平均次元数を示す。表 2 より、いずれのタスクにおいても提案手法が他の手法よりも低い誤り率を達成しており、最も効果の高い Bike タスクでは、ドメイン適応を行わない場合と比較して誤り率が 1/4 程度低減していることが分かる。さらに、図 4 より、提案手法は既存手法よりも低次元の特徴変換によって高い性能を達成していることが分かる。これは、提案手法ではサンプル重み付けによって適応の難しいデータを効率的に排除しつつ、学習データと評価データの双方を良く記述する特徴変換が得られているた

表 2: 各手法の識別誤り率

Method	Bike	Head	Rider
w/o domain adaptation	7.45%	13.82%	19.68%
GFK	6.85%	15.00%	19.57%
Subspace alignment	7.54%	13.73%	19.67%
Proposed method	**5.71**%	**13.68**%	**18.27**%

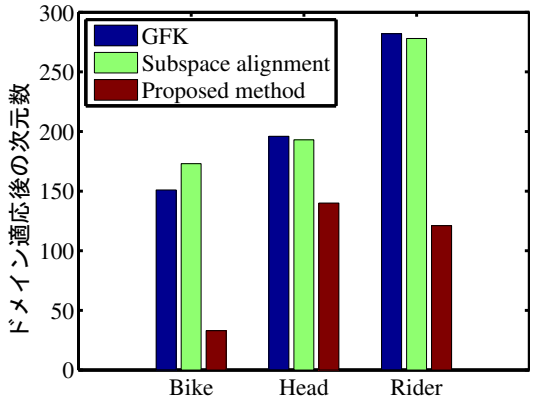


図 4: ドメイン適応後の次元数

めである。また、いずれの手法においても、Bike タスクと比較して Head タスクや Rider タスクではドメイン適応の効果が相対的に低い傾向が確認できる。これは、学習データと評価データの特性の違いの複雑さに起因すると考えられる。図 3 を見ると、A地点の監視映像とB地点の監視映像の違いは主にカメラの角度であり、A地点の映像では斜め上方向から撮影しているのに対し、B地点の映像ではより水平に近い方向であることが分かる。バイクは平面に近い物体であるため、カメラの角度による違いはアフィン変換で近似できるような比較的単純な違いである。一方、頭部や人体は立体的な対象であるため、カメラの角度の違いによってオクルージョンなどの複雑な変化が発生する。これにより、Head タスクや Rider タスクでは適切なドメイン適応が難しくなり、Bike タスクと比較すると、相対的にその効果が低くなったと考えられる。

4. 結論

本稿では、サンプル重み付けと特徴変換を併用したドメイン適応手法を提案した。本手法では、重み付けしたソースデータとターゲットデータの双方が良く分布するように、サンプル重みと特徴変換を同時に求める。重みを均一に近づける正則化を導入することによって、適応されるソースデータの実効的なサンプル数をできるだけ保ち、適応後の識別学習において多くの学習データで学習を行うことができる。実際の映像監視システムで得られたデータで実験を行い、提案手法によって、既存手法よりも効果的にドメイン適応を行うことができ、適応後の識別学習の精度が向上することを示した。今後は、RBM や auto-encoder などの性能の高い非線形な特徴変換を用いた場合の提案手法の評価などを行う。

参考文献

[Fernando 13] Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T.: Unsupervised Visual Domain Adaptation Using Subspace Alignment, in *IEEE International Conference on Computer Vision*, pp. 2960–2967 (2013)

[Gong 12] Gong, B., Shi, Y., Sha, F., and Grauman, K.: Geodesic Flow Kernel for Unsupervised Domain Adaptation, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073 (2012)

[Hinton 02] Hinton, G.: Training products of experts by minimizing contrastive divergence, *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800 (2002)

[Huang 06] Huang, J., Gretton, A., Borgwardt, K. M., Scholkopf, B., and Smola, A. J.: Correcting Sample Selection Bias by Unlabeled Data, in *Advances in Neural Information Processing Systems*, pp. 601–608 (2006)

[Hull 94] Hull, J. J.: A Database for Handwritten Text Recognition Research, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 5, pp. 550–554 (1994)

[Kanamori 09] Kanamori, T., Hido, S., and Sugiyama, M.: A Least-squares Approach to Direct Importance Estimation, *the Journal of Machine Learning Research*, Vol. 10, pp. 1391–1445 (2009)

[LeCun 98] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based Learning Applied to Document Recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998)

[Saenko 10] Saenko, K., Kulis, B., Flitz, M., and Darrell, T.: Adapting Visual Category Models to New Domains, in *European Conference on Computer Vision*, pp. 213–226 (2010)

[Sato 13] Sato, A. and Ishii, M.: Inverse of Lorentzian Mixture for Simultaneous Training of Prototypes and Weights., in *International Conference on Pattern Recognition Applications and Methods*, pp. 151–158 (2013)

[Shimodaira 00] Shimodaira, H.: Improving Predictive Inference under Covariate Shift by Weighting the Log-likelihood Function, *Journal of Statistical Planning and Inference*, Vol. 90, No. 2, pp. 227–244 (2000)

[Vincent 10] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, *ACM Journal of Machine Learning Research*, Vol. 11, pp. 3371–3408 (2010)