Report for CSE 5339 2018 — (OTMLSA)
Optimal Transport in Machine Learning and Shape Analysis

Joint distribution optimal transportation for domain adaptation

Changhuang Wan

04/24/2018

1 Background

In this report, the work in paper [1] is going to be presented and their contributions will be discussed. The paper is published in 2017 and written by Nicolas Courty, Rémi Flamary, Amaury Habrard and Alain Rakotomamonjy. In paper [1], solving the domain adaptation problem based on optimal transport is investigated and studied, and the unsupervised domain adaptation problem where data labels are only available in the source domain is addressed.

1.1 Domain Adaptation

Domain Adaptation (DA) is a technique related to machine learning and transfer learning[2]. At the beginning, in most theoretical and empirical studies, models are assumed to be trained and tested using data gathered from some fixed distributions.[5] When training and test data are drawn from the same distribution, many discriminative learning methods for classification have good performance.[6] However, in many practical applications, source domains and target domains are different. In other words, the data used for training the model may not follow the same distribution with target domain. Then, two questions appear:

1) under what conditions can a classifier trained from source data be expected to perform well on target data?

2) how can we train a model from some source domains and then apply it to different target domains with the lowest target error at test time?

Therefore, we are going to read paper [1] to find out the answers are addressed by the authors for these two questions.

If the domain adaptation is done correctly, models built on a specific data representation become more robust when confronted to data depicting the same classes, but described by another observation system. Among the many strategies proposed, finding domain-invariant representations has shown excellent properties, in particular since it allows to train a unique classifier effective in all domains.

1.2 Optimal Transport

Since the Optimal Transport (OT) can be applied in computing distances between probability distributions, many researchers in different fields have raised interest in OT problems. Wasserstein, Monge-Kantorovich or Earth Mover distances are three typical distances which are widely used. They share some central properties: i) They can be evaluated directly on empirical estimates of the distributions without having to smoothen them using nonparametric or semi-parametric approaches; ii) By exploiting the geometry of the underlying metric space, they provide meaningful distances even when the supports of the distributions do not overlap. In the work of paper [1], the 1-Wasserstein distance is used.

2 Introduction of Work in Paper[1]

2.1 Notations

Let $\Omega \in \mathbb{R}^d$ be an input measurable space of dimension d and $\mathcal{C}$ the set of possible labels. $\mathcal{P}(\Omega)$ denotes the set of all the probability measures over Ω. The standard learning paradigm assumes the existence of a set of training data $X_S = \{x_i^s\}_{i=1}^{N_s}$ associated with a set of class labels $Y_S = \{y_i^s\}_{i=1}^{N_s}$ where $y_i^s \in \mathcal{C}$, and a testing data set $X_T = \{x_i^t\}_{i=1}^{N_t}$ with unknown labels. N_s and N_t are the number of components of X_S and X_T respectively. $\mathcal{P}(\Omega \times \mathcal{C})$ refers to all probability over $\Omega \times \mathcal{C}$. $P_s(\mathbf{x}, y), P_t(\mathbf{x}, y) \in \mathcal{P}(\Omega \times \mathcal{C})$ are respectivel joint probability distributions in source and target domain. μ_s and μ_t denotes marginal distributions over X_S and X_T.

2.2 Assumptions

In domain adaptation problems, most domain adaptation methods would have at least one of the two following assumptions:

Class imbalance: Label distributions are different in the two domains ($P_s(y) \neq P_t(y)$), but the conditional distributions of the samples with respect to the labels are the same ($P_s(\mathbf{x}^s|y) = P_t(\mathbf{x}^t|y)$);

Covariate shift: Conditional distributions of the labels with respect to the data are equal ($P_s(y|\mathbf{x}^s) = P_t(y|\mathbf{x}^t)$, or equivalently $f_s = f_t = f$). However, data distributions in the two domains are supposed to be different ($P_s(\mathbf{x}^s) \neq P_t(\mathbf{x}^t)$).

Note that this difference should be small if the adaptation techniques are supposed to be effective. The first assumption is considering the difference of label distributions, whereas the second one talks about the difference of data distributions. But the conditional distributions are supposed to be the same. Although there is no clear reason which make these assumptions hold, they are still widely considered since it is too difficult to adapt both marginal feature and conditional distributions by minimzing a global divergence between D_S and D_T. In practical applications, the drift occurring between the source and the target domains generally implies a change in both marginal and conditional distributions.[3]

Thus, in paper[1], both marginal and conditional distributions are considered. C. Nicolas et al.[1] assumed that there exists a nonlinear transformation between the label space distributions of the source and target domains that can be estimated with optimal transport $\mathbf{T} : \Omega_s \to \Omega_t$. Additionnally, it is assumed that the transformation preserves the conditional distribution.

$$P_s(y|\mathbf{x}^s) = P_t(y|\mathbf{T}(\mathbf{x}^s)) \tag{1}$$

The (1) indicates that the label information is preserved by the transformation, and the Bayes decision functions are tied through the equation

$$f_t(\mathbf{T}(\mathbf{x})) = f_s(\mathbf{x}) \tag{2}$$

2.3 DA problem with OT

2.3.1 Problem formulation

In DA problem, to find out the set of labels Y_T associated with X_T, the empirical estimate of the joint probability distribution $P(X, Y) \in P(\Omega \times \mathcal{C})$ from (X_S, Y_S) will be usually relied on under the assumption that X_S and X_T are drawn from the same distribution $\mu \in P(\Omega)$. In paper[1], the authors also assume that there exist two distinct joint probability distributions $P_s(X, Y)$ and $P_t(X, Y)$ which correspond respectively to two different *source* and *target* domains. we study two different (but related) distributions D_S and D_T on $X \times Y$. The DA task consists of the transfer of knowledge from D_S to D_T. optimal transport formulation in domain adaptation:

$$\gamma_0 = \underset{\gamma \in \Pi(\mu_s, \mu_t)}{\arg\min} \int_{\Omega \times \Omega} d(\mathbf{x}_1, \mathbf{x}_2) \mathrm{d}\gamma(\mathbf{x}_1, \mathbf{x}_2) \tag{3}$$

where $\Pi(\mu_s, \mu_t) = \{\gamma \in \mathcal{P}(\Omega \times \Omega) | p^+ \# \gamma = \mu_s, p^- \# \gamma = \mu_t\}$, p^+, p^- denotes the two marginal projections of $\Omega \times \Omega$ to Ω, and $p\#\gamma$ is the image measure of γ by p, $d(\mathbf{x}_1, \mathbf{x}_2)$ is the distance function between $\mathbf{x}_1$ and $\mathbf{x}_2$.

To handle a change in both marginal and conditional distributions, a joint distribution optimal transport loss is investigated

$$\gamma_0 = \underset{\gamma \in \Pi(P_s, P_t)}{\arg\min} \int_{(\Omega \times \mathcal{C})^2} \mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) \mathrm{d}\gamma(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) \tag{4}$$

where $\mathcal{D} = \alpha d(\mathbf{x}_1, \mathbf{x}_2) + \mathcal{L}(y_1, y_2)$ is a joint cost measure combining both distance $d(\mathbf{x}_1, \mathbf{x}_2)$ and a loss function $\mathcal{L}(y_1, y_2)$ which measures the discrepancy between y_1 and y_2. Moreover, the loss function $\mathcal{L}$ is assumed to be bounded, symmetric, $k-$Lipschitz and satisfying the triangle inequality, that is, $\forall y_1, y_2, y_3 \in \mathcal{C}$

$$\begin{align} \text{Symmetric:} \quad & \mathcal{L}(y_1, y_2) = \mathcal{L}(y_2, y_1), \tag{5} \\ k\text{-Lipschitz:} \quad & \exists k > 0, \text{such that} |\mathcal{L}(y_1, y_2) - \mathcal{L}(y_1, y_3)| \leqslant k|y_2 - y_3|, \tag{6} \\ \text{Triangle inequality:} \quad & \mathcal{L}(y_1, y_3) \leqslant \mathcal{L}(y_1, y_2) + \mathcal{L}(y_2, y_3), \tag{7} \end{align}$$

These properties will be used in proof of bounded error.

In the unsupervised DA problem, we do not have access to labels in the target domain, and thus it is very difficult to find the coupled optimal solution. Let $f : \Omega \to \mathcal{C}$ be an hypothesis label function from a given class of hypothesis $\mathcal{H}$ in the target domain. Define the following joint distribution based on f as a proxy for y in D_T:

$$P_t^f = (\mathbf{x}, f(\mathbf{x}))_{\mathbf{x} \sim \mu_t} \tag{8}$$

And the empirical versions of P_s, P_t are considered:

$$\hat{P}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{\mathbf{x}_i^s, y_i^s}, \hat{P}_t^f = \frac{1}{N_t} \sum_{i=1}^{N_t} \delta_{\mathbf{x}_i^t, f(\mathbf{x}_i^t)}. \tag{9}$$

In the discrete case, γ will be a matrix which belongs to Δ, that is, the transportation polytope of nonnegative matrices between uniform distributions. The goal is to estimate a prediction f on target domain, and then the OT problem can be formulated as

$$\min_{f, \gamma \in \Delta} \sum_{i,j} \mathcal{D}(\mathbf{x}_i^s, y_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t)) \gamma_{ij} = \min_f W_1(\hat{P}_s, \hat{P}_t^f) \tag{10}$$

where W_1 is the 1-Wasserstein distance for $\mathcal{D}$, formulated as:

$$W_1(P_s, P_t^f) = \inf_{\Pi \in \Pi(P_s, P_t^f)} \int_{(\Omega \times \mathcal{C})^2} \alpha d(\mathbf{x}_s, \mathbf{x}_t) + \mathcal{L}(y_s, y_t^f) \mathrm{d}\Pi((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)). \tag{11}$$

2.3.2 Proof of Bound on the Target Error

The authors[1] addressed the first question by proving that the proposed classifier's target error in terms of its source error and the divergence between the two domains is bounded. Define the expected loss in the target domain $err_T(f)$ and source domain $err_S(f)$ by

$$err_T(f) \triangleq \mathbf{E}_{(\mathbf{x},y)\sim P_t}\mathcal{L}(y, f(\mathbf{x})), err_S(f) \triangleq \mathbf{E}_{(\mathbf{x},y)\sim P_s}\mathcal{L}(y, f(\mathbf{x})) \tag{12}$$

Here, $\mathbf{E}$ denotes the expectation function. Then the expected inter function loss $err_T(f, g) = \mathbf{E}_{(\mathbf{x},y)\sim P_t}\mathcal{L}(g(\mathbf{x}), f(\mathbf{x}))$. And in order to get the bound of error, Probabilistic Transfer Lipschitzness (PTL) is defined by

Definition 1 *Let μ_s and μ_t be respectively the source and target distributions. Let $\phi : \mathbb{R} \rightarrow [0, 1]$. A labeling function $f : \Omega \rightarrow \mathbb{R}$ and a joint distribution $\Pi(\mu_s, \mu_t)$ over μ_s and μ_t are ϕ-Lipschitz transferable if for all $\lambda > 0$:*

$$\mathbf{Pr}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\Pi(\mu_s,\mu_t)}[|f(\mathbf{x}_1) - f(\mathbf{x}_2)| > \lambda d(\mathbf{x}_1, \mathbf{x}_2)] \leqslant \phi(\lambda). \tag{13}$$

Here, $\mathbf{Pr}$ is the probability function, and $|f(\mathbf{x}_1) - f(\mathbf{x}_2)|$ refers to the difference of labels. This definition says that the probability of finding pairs of source-target instances labeled differently is bounded in a $(1/\lambda)$-ball when given deterministic labeling functions f and a coupling Π. In addition, definition 1 is useful for proving empirical concentration result for Wasserstein distance. Based on the definitions and assumptions above, a theorem for upper boundof $err_T(f)$ is given in paper [1]: there exists, c_0 and $N > 0$, such that for $N_s > N$ and $N_t > N$, for all $\lambda > 0$, with $\alpha = k\lambda$, we have with probability at least $1 - \phi$:

$$\begin{aligned} err_T(f) \quad \leqslant & \quad W_1(\hat{P}_s, \hat{P}_t^f) + \sqrt{\frac{2}{c'}\log(\frac{2}{\delta})}(\frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}}) \\ & + err_S(f^*) + err_T(f^*) + kM\phi(\lambda) \end{aligned} \tag{14}$$

In (14), $\sqrt{\frac{2}{c'}\log(\frac{2}{\delta})}(\frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}})$ corresponds to the objective function, $err_S(f^*) + err_T(f^*)$ relates to the joint error minimizer illustrating that domain adaptation can work only if we can predict well in both domains, and $kM\phi(\lambda)$ assesses the probability under which the PTL does not hold.

2.3.3 Learning with Joint Distribution OT

In paper [1], an optimization approach based on block coordinate Descent has been proposed to solve the Joint Distribution OT problem. Assume that the function space to which f belongs is either a RKHS or a function space parametrized by some parameters $\mathbf{w} \in \mathbb{R}^p$. Instead of solving (10), a regularized optimal transport formulation is considered:

$$\min_{f\in\mathcal{H},\gamma\in\Delta}\sum_{i,j}\gamma_{i,j}(\alpha d(\mathbf{x}_i^s, \mathbf{x}_i^t) + \mathcal{L}(y_i^s, f(x_j^t))) + \lambda\Omega(f) \tag{15}$$

where $\Omega(f)$ is the regularization term either a non-decreasing function of the squared-norm or a squared-norm on the vector parameter. Additionally, $\Omega(f)$ is continuously differentiable. The idea of the regularization form is that due to most elements of γ_0 should be zero with high probability, a smoother version of the transport could be found by lowering its sparsity, by increasing its entropy. As a result, the optimal transport γ for (15) will have a denser coupling between the distributions. Then, with fixed γ, the optimization problem (15) can reformulated as

$$\min_{f\in\mathcal{H}}\sum_{i,j}\mathcal{L}(y_i^s, f(\mathbf{x}_j^t)) + \lambda\Omega(f) \tag{16}$$

Then the Frame of Block coordinate descent algorithm for joint distribution OT problem can be concluded as below

```
Initialization function $f^0$ and set $k = 1$;
Set $\alpha$ and $\lambda$;
while not converged do
    $\gamma^k \leftarrow$ Solve OT problem (10) with fixed $f^{k-1}$;
    $f^k \leftarrow$ Solve learning problem (16) with fixed $\gamma^k$;
    $k \leftarrow k + 1$;
end
```

Algorithm 1: Optimization with BCD

With this method above, the authors gave the answer for second question in the first section.Also, the authors present some examples to show its effectiveness and efficincy.

3 Discussions

According to the introduction and analysis, as the authors declared the major contribution of this work is the new framework for unsupervised domain adaptation between joint distributions. However, there are some concerns from me: (a) The theorem holds based on many assumptions, then, will it limit the application in other practical problem? Or which kinds of problem, this approach can preform well. Since the difference between the source and target domain is supposed to be small, how could we determine it 'small enough'? (b) It seems that the BCD method relies on the initial guess of f. Does f^0 matter or not? If yes, how to get a good initial f^0. (c) When I try to repeat the proof of theorem in that paper, I don't go smoothly for formula (20) in Appendix section D in [1]. It looks like we can't get it from inequality (19).

References

[1] Courty, Nicolas, et al. "Joint distribution optimal transportation for domain adaptation." Advances in Neural Information Processing Systems. 2017. https://arxiv.org/pdf/1705.08848.pdf.

[2] https://en.wikipedia.org/wiki/Domain_adaptation#cite_note-1.

[3] Courty, Nicolas, et al. "Optimal transport for domain adaptation." IEEE transactions on pattern analysis and machine intelligence 39.9 (2017): 1853-1865.. https://ieeexplore.ieee.org/document/7586038/.

[4] R. Flamary et al. "Optimal Transport for Domain Adaptation." http://remi.flamary.com/biblio/courty2016optimal.pdf. IEEE-TPAMI 2016.

[5] Ben-David, Shai, et al. "A theory of learning from different domains." Machine learning 79.1-2 (2010): 151-175." https://link.springer.com/content/pdf/10.1007%2Fs10994-009-5152-4.pdf.

[6] Juang, B-H., and Shigeru Katagiri. "Discriminative learning for minimum error classification (pattern recognition)." IEEE Transactions on signal processing 40.12 (1992): 3043-3054. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=175747.

[7] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transportation," in Neural Information Processing Systems (NIPS), 2013, pp. 2292–2300. https://arxiv.org/pdf/1306.0895.pdf