Safe Multiclass Transfer Learning

No Author Given

No Institute Given

Abstract. In transfer learning, domain adaptation tries to exploit the knowledge from a source domain with plentiful data to help learn a classifier for the target domain with a different distribution and little labeled training data. In this paper, we investigate this problem under the setting of *Hypothesis Transfer Learning* (HTL) where we can only access the source model instead of the data. We aim at two important issues: effectiveness of the transfer and compatibility of the target model with different types of source models in the HTL scenario and proposed our method, SMTLe. We illustrate that our strategy that uses the class probabilities as the auxiliary bias in SMTLe can greatly increase the compatibility of the target model to fit different types of source models. To better exploit the source model, we use the bi-level optimization (BO) method to estimate the transfer parameter which measures the similarity of the source and target domains. We demonstrate that our BO problem is a strongly convex optimization problem and we can effectively obtain the optimal transfer parameter with the sub-gradient descent method. Empirical results show that SMTLe can effectively exploit the knowledge from different types of source models and outperform other HTL baselines as well.

1 Introduction

Domain adaptation for image recognition tries to exploit the knowledge from a source domain with plentiful data to help learn a classifier for the target domain with a different distribution and little labeled training data. In domain adaptation, the source and target domains share the same label but their data are drawn from different distributions.

In domain adaptation, the knowledge of the source domain can be transferred by 3 different approaches: *instance transfer*, *model transfer* and *feature representation transfer* [14]. In this paper, we focus on the method that transfers knowledge from the source model. Some recent works show that exploiting the knowledge from the source model can boost the performance of the target model effectively[18][10]. Moreover, in some real applications, we can only obtain the source models and it is difficult to access their training data for different reasons such as the data credential. Recently, a framework called Hypothesis Transfer Learning (HTL) [9] has been proposed to handle this situation. HTL assumes only source models (called the *hypotheses*) trained on the source domain can be utilized and there is no access to source data, nor any knowledge about the relatedness of the source and target distributions.

Previous research [2] [3] shows that without carefully measuring the distribution similarity between the source and target data, the source knowledge could not be exploited effectively or even hurt the learning process (called *negative transfer*)[14]. However, as we are not able to access the source data in an HTL setting, how to effectively and safely exploit the knowledge from the source model could be an important issue in HTL (Safety issue). Moreover, different source models can be trained with different kinds of classifiers. For example most models trained from ImageNet are deep convolutional neural networks while some models of the VOC recognition task could be SVMs or ensemble models. Therefore, a practical HTL algorithm should be compatible with different types of source classifiers (Compatibility issue). To the best of our knowledge, none of the previous work in HTL is able to solve these two issues at the same time.

In this paper, we propose our method, called Safe Multiclass Transfer Learning (SMTLe), that can solve these two issues simultaneously. Previous work [7] suggests that feature augmentation can greatly increase the compatibility of the target model in the HTL scenario. To solve the compatibility issue, we propose our strategy that uses the class probabilities as the auxiliary bias. For each example, we use class probabilities from the source model as an auxiliary bias to adjust the output of the target model. Moreover, we apply different weights (called transfer parameters) for different auxiliary biases to control the amount of the knowledge transferred from each source model. As a result, the value of the transfer parameter reflects the similarity between the source and target domain. By carefully estimating the transfer parameters, we can obtain the optimal target model that can exploit the knowledge from the source model effectively.

To better estimate the transfer parameters, we treat them as the hyperparameters of a convex optimization problem. Then to estimate the optimal values, we introduce bi-level hyperparameter optimization[15], which has been widely used for many different hyperparameter optimization problems. Specifically, on the low-level optimization problems, we use a least-square SVM to obtain the hyperplane and on the high level, we use the novel multi-class hinge loss with ℓ_2 penalty. Different from many other bi-level optimization problems which are non-convex optimization problems, we show that our transfer parameter estimation problem is a strongly convex optimization problem and demonstrate that our method SMTLe can find the $O(\log(t)/t)$ optimal solution with t iteration.

In our experiment, we use the popular benchmarks Office and Caltech256 as our dataset. We show that SMTLe can successfully transfer the knowledge with different types of source models. Moreover, we show that our novel high level objective function with ℓ_2 penalty can improve the performance of the target model effectively compared with SMTLe without ℓ_2 penalty and other baselines in HTL.

The rest of this paper is organized as follows: In Section 2 we introduce the issues in transfer learning and some related work regarding these issues. In Section 3, we introduce our strategy using the class probabilities as the auxiliary bias to adapt different types of source models. Then, we propose a novel objective function using ℓ_2 penalty term for transfer parameter estimation, called SM-

TLe in Section 4. We use Bi-level hyperparameter optimization to estimate the transfer parameter. We demonstrate that the ℓ_2 penalty term of our objective function can help SMTLe estimate the transfer parameter better especially when the size fo the target training set is small. In Section 5, we show the performance comparison between SMTLe and other baselines on a variety of experiments on Office and Caltech datasets.

2 Related Work

The motivation of transferring knowledge between different domains is to apply the previous information from the source domain to the target one, assuming that there exists a certain relationship, explicit or implicit, between the feature space of these two domains [14]. Technically, previous work can be categorized into solving the following three issues: *what*, *how* and *when* to transfer [18].

What to transfer. Previous work tried to answer this question from three different aspects: (1) selecting transferable instances, (2) learning transferable feature representations and (3) transferable model parameters. Instance-based transfer learning assumes that part of the instances in the source domain could be re-used to benefit the learning for the target domain. Lim et al. [11] proposed a method of augmenting the training data by borrowing data from other classes for object detection. Learning transferable features means to learn common features that can alleviate the bias of data distribution in the target domain. Recently, Long et al. [12] proposed a method that can learn transferable features using deep neural network and showed some impressive results on the benchmarks. A model transfer approach assumes that the parameters of the model for the source task can be transferred to the target task. Yang et al. [20] proposed Adaptive SVMs transferring parameters by incorporating the auxiliary classifier trained from the source domain. In addition to Yang's work, Ayatar et al. [1] proposed PMT-SVM that can determine the transfer regularizer automatically according to the target data. Tommasi et al. [18] proposed Multi-KT that can utilize the parameters from multiple source models for the target classes . Kuzborskij et al. [10] proposed a similar method to learn new categories by leveraging the known source models.

When and how to transfer. The question *when to transfer* arises when we want to know if the information acquired from the previous task is relevant to the new one (i.e. in what situations knowledge should not be transferred). *How to transfer* the prior knowledge effectively should be carefully designed to prevent inefficient and negative transfer. Previous work [6] [19] [21] has used the generative probabilistic method. Bayesian learning methods can predict the target domain by combining the prior source distribution to generate a posterior distribution. Alternatively, max margin methods [10] [17] show that it is possible to learn from a few examples by minimizing the Leave-One-Out (LOO) error for the training model. Cawley et al. [4] show that there is a closed-form implementation of LOO cross-validation that can generate an unbiased model estimation for LS-SVM.

Our work corresponds to the context above. In this paper, we propose SMTLe based on the model transfer approach with LS-SVM. Our work addresses how to prevent negative transfer while only the source model is accessible for domain adaptation. Compared to other works, we propose a new perspective which provides insight on negative transfer. Based on this, we propose our novel objective function and show that SMTLe can better leverage knowledge from different source models. As a result, SMTLe can achieve a better performance and alleviate negative transfer.

3 Using the Source Knowlege as the Auxiliary Bias

In this section, we introduce our strategy in SMTLe that can exploit the knowledge from different types of classifiers. In general, for each example in the target domain, we use its output class probabilities from the source models as the auxiliary bias term to adjust the final output.

To make our method compatible with different types of source model, we have to find a solution to align the knowledge of different types of classifier. It is clear that most of the existing classifiers can output the class probability for the input examples. Therefore, in SMTLe, we use the class probability of the target examples from the source model to align the knowledge of the different types of source model.

To use the aligned source knowledge, we use a straight-forward way, by using the class probability as the auxiliary information to adjust the output of the target model (see Figure 3). However, as we know that due to the domain shift, the performances of the different source models vary on the target domain. Therefore, to compensate this domain shift, we apply different weights to different source model outputs. Here, the weight of each model reflects the relatedness between the source model and the target domain. The more related they are, the larger weight we should apply. Specifically, in this paper, we call it *transfer parameter*. Therefore, for any target learning $D_T = \{x, y\}$ and the given source model f', our goal is to find the target model f:

$$f = \arg\min_{f \in \mathcal{F}} \ell\left(f + \beta f', D_T\right) \tag{1}$$

where β is the transfer parameter and $\ell(\cdot, \cdot)$ is the loss function to learn the target model.

There are several advantages of our feature augmentation with class probability: (1) It is an effective and easy way to align the knowledge from different types of source model. (2) Features are naturally normalized in the same dimension as the class probability is always in the interval $[0, 1]$. (3) The bonus advantage: increase the selection of the source domain. As SMTLe can select more types of source model, we have more options to select our source domain. As long as the model has the knowledge of the class we want (even though we may only need knowledge of one class in a multi-class classifier), we can still exploit the knowledge.

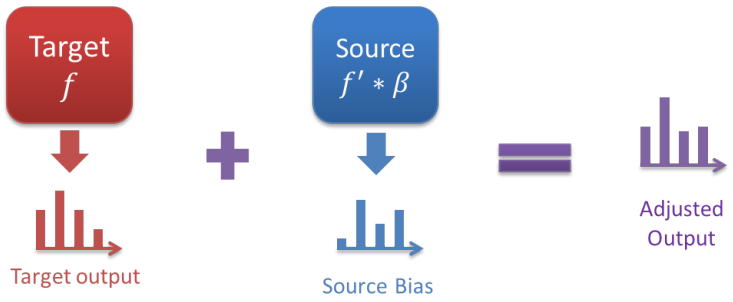

Fig. 1. Demonstration of using the source class probability as the auxiliary bias to adjust the output of the target model.

From Eq. (1) we can see that, once we can determine the value of the transfer parameter β, we are able to find the target model f_T. In the next part, we will show how we can effective estimate the unbiased transfer parameter effectively.

4 Bi-level Optimization for Transfer Parameter Estimation

In Eq. (1), we have to find the optimal target function f that minimize the training error on the target domain in addition with the source model f' and the transfer parameter β. As we discussed before, the transfer parameter in Eq. (1) is a hyperparameter that is decided by the relatedness between the source model and target domain. A simple way to measure the relatedness is to evaluate the performance of the source model on the target training data. However, this estimate could lead to a relatively large variance when the target training data is small. In this paper, we use the leave-one-out cross-validation (LOOCV) strategy to reduce this variance. Previous research [9] suggests that LOOCV can increase the robustness of the estimated hyperparameter especially on small dataset. For the i-th round, the target set for training $D_T = \{x_i, y_i | i \in 1, ..., l\}$ is split into training set $D_T^{\setminus i} = \{x_j, y_j | j \neq i\}$ and validation set $D_T^i = \{x_i, y_i\}$. The transfer parameter can be optimized with the following bi-level optimization (BO) function:

$$\begin{aligned} &\arg\min_{\beta} \sum_{i}^{l} \mathcal{L}(f^i(\beta), D_T^i) \\ &f^i(\beta) = \arg\min_{f \in \mathcal{F}} \ell\left(f + \beta f', D_T^{\setminus i}\right) \end{aligned} \tag{2}$$

Here, we can use any convex objective function (e.g. SVM objective function) as $\ell(\cdot, \cdot)$ for the low-level optimization problem and $\mathcal{L}(\cdot, \cdot)$ is our the high-level cross-validation objective function. In many previous works[13][15], BO optimization is a non-convex problem and can only obtain the approximate solution. However,

in our paper, we will show that problem (2) is strongly convex and we are able to obtain its optimal solution.

4.1 Low-level optimization problem using mean square loss

To better illustrate our learning scenario, define our learning process as follows. Suppose we have N visual categories and can obtain N source binary classifiers $f' = \{f'_1, ..., f'_N\}$ from the source domain. We want to train a target function f consists of N binary classifier $f = \{f_1, ..., f_N\}$ using the target training set D_T and the source models f'. Specifically, in our BO problem Eq. (2), for the low level optimization problem, we consider the scenario where each of the binary target model is a linear classifier $f_i = w_i x + b_i$. Fand we use Least-square loss as the objective function to optimize each target model f_n for any given transfer parameter β:

$$\begin{aligned} \text{Low-level:} \quad & f^{\setminus i}(\beta) : \min \sum_{n}^{N} \frac{1}{2}||w_n||^2 + \frac{C}{2} \sum_{j}^{l} \left(Y_{jn} - f_n(x_j) - \beta_n f'_n(x_j)\right)^2 \\ & \text{s.t.} \qquad x_j \in D_T^{\setminus i} \end{aligned} \tag{3}$$

Here, Y is an encoded matrix of y using one-hot strategy where $Y_{in} = 1$ if $y_i = n$ and 0 otherwise.

The reason why we use the objective function (3) is that, it can provide an unbiased closed form Leave-one-out error estimation of each binary model f_n[4]. Let $K(X, X)$ be the kernel matrix and

$$\psi = \left[K(X, X) + \frac{1}{C}\mathrm{I}\right] \tag{4}$$

Let ψ^{-1} is the inverse of matrix ψ and ψ_{ii}^{-1} is the ith diagonal element of ψ^{-1}. $\hat{Y}_{in}$, the LOO estimation of binary model f_n for sample x_i, can be written as:

$$\hat{Y}_{in} = Y_{in} - \frac{\alpha_{in}}{\psi_{ii}^{-1}} \quad \text{for} \quad n = 1, ..., N \tag{5}$$

where the matrix $\boldsymbol{\alpha} = \{\alpha_{in} | i = 1, ...l; n = 1, ..., N\}$ can be calculated as:

$$\boldsymbol{\alpha} = \psi^{-1} Y - \psi^{-1} f'(X) \boldsymbol{\beta}^T \tag{6}$$

4.2 High-level optimization problem using multi-class hinge loss with ℓ_2 pentalty

For high level optimization problem, we use multi-class hinge loss[5] with ℓ_2 penalty as our objective function.

$$\begin{aligned} \text{High-level:} \quad & \beta : \min \frac{\lambda}{2} \sum_{n=1}^{N} \|\beta_n\|^2 + \sum_{i,n} \left[1 - \varepsilon_{ny_i} + \hat{Y}_{in} - \hat{Y}_{iy_i} - \xi_i\right] \\ & \text{s.t.} \qquad 1 - \varepsilon_{ny_i} + \hat{Y}_{in} - \hat{Y}_{iy_i} \leq \xi_i \end{aligned} \tag{7}$$

Here, $\varepsilon_{ny_i} = 1$ if $n = y_i$ otherwise 0. Compared to the previous work [18][10] which uses the multi-class hinge loss without the ℓ_2 penalty, there are two main advantages for our objective function: (1) When the training set is small, our LOOCV estimation could have a large variance. Similar to the penalty term in low-level problem (3), the ℓ_2 penalty here can reduce this variance and improve the generalization ability of the estimated transfer parameter. (2) With the ℓ_2 penalty, optimization problem (7) become a strongly convex optimization problem w.r.t. the transfer parameter β. Therefore, we can obtain an $O(\log(t)/t)$ optimal solution with t iterations using Algorithm 1 (see proof in Theorem 1 in Appendix).

Algorithm 1 SMTLe

Input: $\lambda, \psi, Y, f', T,$
Output: $\beta = \{\beta^1, ..., \beta^n\}$
$\beta^0 \leftarrow 1$
$\alpha' = \psi^{-1}Y, \alpha'' = \psi^{-1}f'$
for $t = 1$ to T **do**
  $\hat{Y} \leftarrow Y - \left(\psi^{-1} \circ I\right)^{-1}(\alpha' - \alpha''\beta)$,$\Delta_\beta = 0$
  for $i = 1$ to l **do**
    $\Delta_\beta \leftarrow \Delta_\beta + \lambda\beta$
    $l_{ir} = \max(1 - \varepsilon_{y_i r} + \hat{Y}_{ir} - \hat{Y}_{iy_i})$
    if $l_{ir} > 0$ **then**
      $\Delta_\beta^{y_i} \leftarrow \Delta_\beta^{y_i} - \frac{\alpha''_{iy_i}}{\psi_{ii}^{-1}}$, $\Delta_\beta^{r} \leftarrow \Delta_\beta^{r} + \frac{\alpha''_{ir}}{\psi_{ii}^{-1}}$
    end if
  end for
  $\beta^t \leftarrow \beta^{(t-1)} - \frac{\Delta_\beta}{\lambda \times t}$
end for

5 Experiment

In this section, we show empirical results of our algorithm for different transferring situations on two image benchmark datasets: Office and Caltech.

5.1 Dataset & Baseline methods

Office contains 31 classes from 3 subsets (Amazon,Dslr and Webcam) and Caltech contains 256 classes. We select 13 shared classes from two datasets[1]. The input features of all examples are extracted using AlextNet[8]. Because the two subsets Dslr and Webcam are relatively small and don't have data for testing, we don't use them as our target domain.

[1] 13 classes include: backpack, bike, helmet, bottle, calculator, headphone, keyboard, laptop, monitor, mouse, mug, phone and projector

We compare our algorithm SMTLe with two kinds of baselines. The first one is the methods without leveraging any prior knowledge (no transfer baselines). **No transfer:** SVMs trained only on target data. Any transfer algorithm that performs worse than it suffers from negative transfer. **Batch:** We combined the source and target data, assuming that we have full access to all data, to train the SVMs. The result of the Batch method might be considered as the best performance achieved during the transfer learning. The second kind of baseline consists of two previous transfer methods in HTL, **MKTL[7]** and **Multi-KT[18]**. Similar to SMTLe, both of them use the LOOCV method to estimate the relatedness of the source model and target domain, but they use their own convex objective function without the ℓ_2 penalty terms. We use linear kernel for all methods in all our experiments.

5.2 Extensive experiments on benchmarks

In this subsection, we perform 6 groups of experiments under the setting of HTL. In each group of experiment, the source model is trained using linear SVMs on the whole source data. We use 5 different sizes of training data for each class in the target domain. Experiment results are reported by averaging over 10 rounds and shown in Figure 2.

Observation & discussion: SMTLe can significantly outperform other baselines especially when the training size is small. Moreover, in some groups of experiments, they even suffer from negative transfer on the small training set. As we discussed above, without the ℓ_2 penalty in the objective functions when the training set is small, these two HTL baselines are not able to estimate the relatedness between the source model and target domain well. However, as the training size increases, the variance of the estimation of LOOCV decreases. The affect of the ℓ_2 penalty term become less significant. Meanwhile, the target data contains more useful information to learn a better target model. Therefore, SMTLe and the other two HTL baselines show similar performance. In some experiments, it is interesting to see that SMTLe can even outperform the Batch method which might be considered as the the best performance under the setting of HTL.

5.3 Transfer from Multi-Model & Multi-Source

As we mentioned, SMTLe can exploit knowledge from different types of models which could greatly extend our selection of the source domain. In this subsection we show that SMTLe can successfully transfer the knowledge from two source models using two different types of classifiers.

6 Conclusion

In this paper, we present a novel method called SMTLe that is able to transfer knowledge of the source model in domain adaptation. Inspired by previous work,

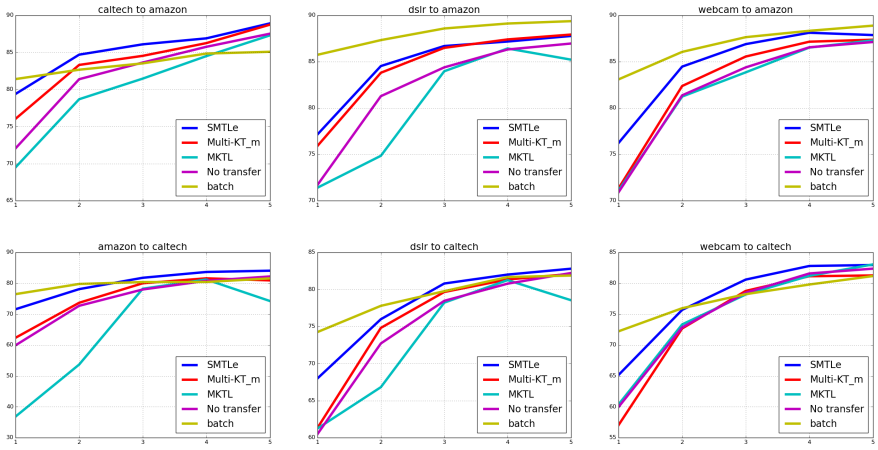

Fig. 2. HTL domain adaptation experiment results. 5 different sizes of target training sets are used in each group of experiments.

Table 1. A2C

	1	2	3	4	5
SMTLe	71.60	78.10	81.80	83.60	84.10
Multi-KT	62.40	73.70	80.00	81.60	80.90
MKTL	36.80	53.70	78.20	81.20	74.20
No transfer	59.90	72.70	78.00	80.80	82.20
batch	76.50	79.80	80.40	80.40	81.60

we propose a novel perspective on the work of HTL and show the reasons why positive and negative transfer would happen in the different scenarios. Based on our analysis, we propose our method SMTLe that can safely leverage the knowledge from the source models to achieve the improved target model performance by limiting the VC dimension of the transfer problem and reduce the empirical risk as well. Experiment results show that SMTLe can leverage related source knowledge and alleviate negative transfer in different scenarios and outperform other baseline methods.

In our perspective on the domain adaptation problem, the feature augmentation approach can fit a wider range of source classifiers. We can leverage the knowledge from any source model that can output the decision score/confidence, such as the Neural Networks and the inference model. Meanwhile, there are still many open issues to solve before we could maximize the utility of different kinds of source classifiers. For example, how to better exploit the knowledge from a deep neural network with our feature augmentation framework and achieve good positive transfer performance and avoid negative transfer simultaneously. These challenges could lead to our future interest.

Appendix

Theorem 1. *Let $L(\beta)$ be a λ-strongly convex function and β^* be its optimal solution.Let $\beta_1, ..., \beta_{T+1}$ be a sequence such that $\beta_1 \in B$ and for $t > 1$, we have $\beta_{t+1} = \beta_t - \eta_t \Delta_t$, where Δ_t is the sub-gradient of $L(\beta_t)$ and $\eta_t = 1/(\lambda t)$. Assume we have $||\Delta_t|| \leq G$ for all t. Then we have:*

$$L(\beta_{T+1}) \leq L(\beta^*) + \frac{G^2(1 + \ln(T))}{2\lambda T} \tag{8}$$

Proof. As $L(\beta)$ is strongly convex and Δ_t is in its sub-gradient set at β_t, according to the definition of λ-strong convexity [16], the following inequality holds:

$$\langle \beta_t - \beta^*, \Delta_t \rangle \geq L(\beta_t) - L(\beta^*) + \frac{\lambda}{2}||\beta_t - \beta^*||^2 \tag{9}$$

For the term $\langle \beta_t - \beta^*, \Delta_y \rangle$, it can be written as:

$$\langle \beta_t - \beta^*, \Delta_t \rangle = \left\langle \beta_t - \frac{1}{2}\eta_t \Delta_t + \frac{1}{2}\eta_t \Delta_t - \beta^*, \Delta_t \right\rangle = \frac{1}{2} \langle \beta_{t+1} + \beta_t - 2\beta^*, \Delta_t \rangle + \frac{1}{2}\eta_t \Delta_t^2 \tag{10}$$

Then we have:

$$||\beta_t - \beta^*||^2 - ||\beta_{t+1} - \beta^*||^2 = (\beta_t - \beta_{t+1})(\beta_t + \beta_{t+1} - 2\beta^*) = \langle \beta_{t+1} + \beta_t - 2\beta^*, \eta_t \Delta_t \rangle \tag{11}$$

Using the assumption $||\Delta_t|| \leq G$, we can rearrange (9) and plug (10) and (11) into it, we have:

$$Diff_t = L(\beta_t) - L(\beta^*) \leq \frac{\lambda(t-1)}{2}||\beta_t - \beta^*||^2 - \frac{\lambda t}{2}||\beta_{t+1} - \beta^*||^2 + \frac{1}{2}\eta_t G^2 \tag{12}$$

Due to the strong convexity, for each pair of $L(\beta_t)$ and $L(\beta_{t+1})$ for $t = 1, ..., T$, according to (9), we have:

$$L(\beta_{t+1}) - L(\beta_t) \leq \langle \beta_{t+1} - \beta_t, \Delta_t \rangle - \frac{\lambda}{2}||\beta_{t+1} - \beta_t||^2 = -\eta_t \Delta_t^2 (1 - \frac{1}{2t}) \leq 0 \tag{13}$$

Therefore, we have the following sequence $L(\beta^*) \leq L(\beta_T) \leq L(\beta_{T-1}) \leq ... \leq L(\beta_1)$. For the sequence $Diff_t$ for $t = 1, ..., T$, we have:

$$\sum_{t=1}^{T} Diff_t = \sum_{t=1}^{T} L(\beta_t) - TL(\beta^*) \geq T\left[L(\beta_T) - L(\beta^*)\right] \tag{14}$$

Next, we show that

$$\sum_{t=1}^{T} Diff_t = \sum_{t=1}^{T} \left\{ \frac{\lambda(t-1)}{2}||\beta_t - \beta^*||^2 - \frac{\lambda t}{2}||\beta_{t+1} - \beta^*||^2 + \frac{1}{2}\eta_t G^2 \right\}$$
$$= -\frac{\lambda T}{2}||\beta_{T+1} - \beta^*||^2 + \frac{G^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{G^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{G^2}{2\lambda}(1 + \ln(T)) \tag{15}$$

Combining (14) and rearranging the result, we have:

$$L(\beta_{T+1}) \leq L(\beta^*) + \frac{G^2(1 + \ln(T))}{2\lambda T}$$

□

References

1. Aytar, Y., Zisserman, A.: Tabula rasa: Model transfer for object category detection. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2252–2259. IEEE (2011)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning 79(1-2), 151–175 (2010)
3. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al.: Analysis of representations for domain adaptation. Advances in neural information processing systems 19, 137 (2007)
4. Cawley, G.C.: Leave-one-out cross-validation based model selection criteria for weighted ls-svms. In: Neural Networks, 2006. IJCNN'06. International Joint Conference on. pp. 1661–1668. IEEE (2006)
5. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. The Journal of Machine Learning Research 2, 265–292 (2002)
6. Davis, J., Domingos, P.: Deep transfer via second-order markov logic. In: Proceedings of the 26th annual international conference on machine learning. pp. 217–224. ACM (2009)
7. Jie, L., Tommasi, T., Caputo, B.: Multiclass transfer learning from unconstrained priors. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 1863–1870. IEEE (2011)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
9. Kuzborskij, I., Orabona, F.: Stability and hypothesis transfer learning. In: Proceedings of the 30th International Conference on Machine Learning. pp. 942–950 (2013)
10. Kuzborskij, I., Orabona, F., Caputo, B.: From n to n+ 1: Multiclass transfer incremental learning. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 3358–3365. IEEE (2013)

11. Lim, J.J.: Transfer learning by borrowing examples for multiclass object detection. Ph.D. thesis, Massachusetts Institute of Technology (2012)
12. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France. pp. 97–105 (2015)
13. Maclaurin, D., Duvenaud, D., Adams, R.P.: Gradient-based hyperparameter optimization through reversible learning. In: Proceedings of the 32nd International Conference on Machine Learning (2015)
14. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
15. Pedregosa, F.: Hyperparameter optimization with approximate gradient. In: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. pp. 737–746 (2016)
16. Rockafellar, R.T.: Convex analysis. Princeton university press (2015)
17. Tommasi, T., Orabona, F., Caputo, B.: Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3081–3088. IEEE (2010)
18. Tommasi, T., Orabona, F., Caputo, B.: Learning categories from few examples with multi model knowledge transfer. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(5), 928–941 (2014)
19. Wang, X., Huang, T.K., Schneider, J.: Active transfer learning under model shift. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 1305–1313 (2014)
20. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: Proceedings of the 15th international conference on Multimedia. pp. 188–197. ACM (2007)
21. Zhou, T., Tao, D.: Multi-task copula by sparse graph regression. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 771–780. ACM (2014)