Safety Multiclass Transfer Learning

***Abstract*—In transfer learning, domain adaptation tries to exploit the knowledge from a source domain with a plentiful data to help learn a classifier for the target domain with a different distribution and little labeled training data. Negative transfer could happen when the source and target domain are not related, especially for the multi-class scenario. In this paper, following the framework of *Hypothesis Transfer Learning* (HTL), we propose a method that can safely transfer the knowledge from the source domain to the target and alleviate negative transfer using the LS-SVM in the multi-class scenario. Inspired by previous work, we first augment data in the target domain by adding the auxiliary features using the outputs of the models trained from the source domain. We show that the performance of the target model is greatly affected by the weights (called transfer parameters) of the auxiliary features. To better estimate the transfer parameter, we propose a novel objective function to estimate the transfer parameters for the auxiliary features and alleviate negative transfer. Experiment results show that our method can alleviate negative transfer and outperform other transfer methods in different scenario.**

I. Introduction

The success of transfer learning suggests that exploiting the knowledge of the existing models properly can greatly help us to learn new data. Transfer learning on image recognition is a very popular topic in recent years. Domain adaptation for image recognition tries to exploit the knowledge from a source domain with a 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 the different distribution.

In domain adaptation, the knowledge of the source domain can be represented in 3 different approaches: instance, model and feature representation [1]. In this paper, we propose a method that transfers the 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, especially when there are just a few examples in the target data [2] [3]. Moreover, in some real applications, we can only obtain the source models and it is difficult to access their training data because of various of reasons such as the data credential. Recently, some works have been proposed within a framework called Hypothesis Transfer Learning (HTL) to handle this situation [4]. HTL assumes only source models (called the *hypotheses*) trained on source task can be utilized and there is no access to source data, nor any knowledge about the relatedness of the source and target distributions. In HTL, a number of works have been attempted with Least Square Support Vector Machine (LS-SVM) [4]. Previous approaches show that the hypotheses can be evaluated effectively with LS-SVM via Leave-One-Out cross-validation [2].

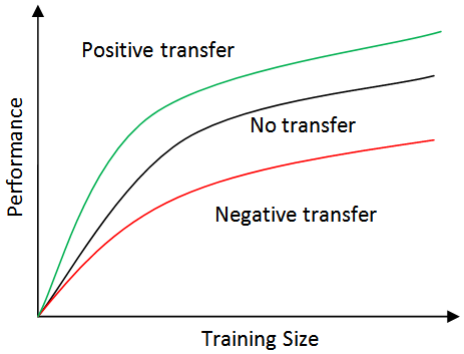

Fig. 1: Relying on the related source knowledge can improve the performance of the target model while forcing the target model to rely on the unrelated source could suffer from negative transfer.

In domain adaptation, different source domain can make the different contribution to the performance of the target model. The theoretical research shows that the utility of the source domain decreases as the distributions of the source and target data become less similar (or *less related*) [5] [6]. Moreover, when the source and target tasks are not related, negative transfer may happen. In transfer learning, *negative transfer* refers to the phenomenon where the source knowledge hurts the learning process and degrades the performance of the target model compare to a method without using any source knowledge [1]. Previous work of HTL assumes that the source and target domains are still very related. Most of them just consider the scenario where the target task is adding a new category to the source task (so called *from N classes to N+1 classes*) [2] [7] [8]. Moreover, their algorithms only focus on the performance on the newly added category, i.e. binary classification scenario, while paying less attention to the performance of the target model on all classes in the target data (the multi-class scenario). In some scenarios where the source and target tasks are less related, negative transfer could happen especially when we consider the performance of the target model on the whole target data (see Figure 2).

How to safely utilize the hypotheses to avoid negative transfer is still an open question in transfer learning [9]. To avoid negative transfer, we have to evaluate the utilities of the hypothesis to keep useful knowledge and reject bad information. This approach can be achieved by setting different weights (called transfer parameters) to each hypothesis. Previous works of HTL use Leave-One-Out error to estimate the transfer

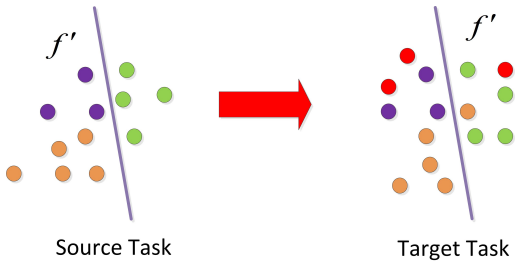

Fig. 2: Negative transfer happens when we transfer source hypothesis f' to target one. Points with different color represent different categories. The data distribution would change when a new category is added into the dataset. The newly added category (red points) can also greatly affect the data distribution in target task and negative transfer could happen when we consider the multi-class scenario.

parameters and avoid negative transfer [2] [7]. However, they try to solving a convex optimization problem which minimizes an upper bound of the leave-one-out error on the training set with a fix regularization term [1]. As a result, when source and target domains are not related, previous methods suffer from negative transfer if this regularization term is not set properly (see experiments in Section V). In this paper, we propose our method, called Safety Multiclass Transfer Learning (SMTLe), that can both alleviate negative transfer and leverage correct hypotheses to improve the performance of the target model. The main contributions of this paper include: (1) We propose a novel algorithm SMITLe within the HTL framework that can safely utilize the hypotheses to prevent negative transfer. We use a novel objective function with a L2 regularization term that can better estimate the transfer parameters and alleviate negative transfer. (2) We also show that by using sub-gradient descent, we can obtain the optimal solution at the rate of $O(\frac{\log(t)}{t})$ where t is training iteration.

The framework of HTL with LS-SVM has two major phrases: (I) Building binary One-Versus-All SVMs with transfer parameters and biased regularization. (II) Estimating the transfer parameters with Leave-one-out error. Following the two phrases of HTL, in Phrase I, inspired by the previous method, we reformulate the previous HTL problem as a data augmentation approach which reconstructs the target data by adding auxiliary features using the outputs of the source hypotheses. From the perspective of the data augmentation approach, we can turn our transfer learning problem into a traditional learning problem. We show that with proper values for the transfer parameter, we can always avoid negative transfer with our data augmentation approach. In Phrase II, based on the closed-form leave-one-out (LOO) error for model evaluation, we propose our novel objective function that can better estimate the transfer parameter and alleviate negative transfer with the L2 regularization term. We prove that transfer parameters learned from our novel objection function can alleviate negative transfer. Moreover, we show that we can always find a $\frac{\log(t)}{t}$ optimal solution with t iterations using sub-gradient descent while previous methods are not able to get any guaranteed convergence rate.

In our experiment, initially, the data of the source and target domains are drawn from the same distribution (dataset). By adding the different level of the noise to the source data, we can generate several sources with different relatedness of the target domain. Experiment results show that when the source and target domain are related (no noise or very little noise is added), all the transfer methods can get improved result and our method outperforms the other baselines. As the source and target domain become less related, the baseline methods suffer from negative transfer while our method can still exploit knowledge from the source domain and the target model can get improved performance.

The rest of this paper is organized as follow. In Section II we introduce the issues in transfer learning and some related work regarding these issues. In Section III, we introduce the biased regularization terms of our problem for Phrase I of HTL. Then, we propose a novel objective function for transfer parameter estimation, called SMTLe in Section IV. We show that the estimated transfer parameter can evaluate the utility of the source hypothesis and avoid negative transfer autonomously. In Section V, we show the performance comparison between SMTLe and other baselines on a variety of experiments on MNIST and USPS datasets.

II. Related Work

The motivation of transfer knowledge between different domains is to apply the previous information from the source domain to the target one, assuming that there exists certain relationship, explicit or implicit, between the feature space of these two domains [1]. Technically, previous work can be concluded into solving the following three issues: what, how and when to transfer [2].

What to transfer. Previous work tried to answer this question from three different aspects: selecting transferable instances, learning transferable feature representations and 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. proposed a method of augmenting the training data by borrowing data from other classes for object detection [10]. Learning transferable features means to learn common feature that can alleviate the bias of data distribution in the target domain. Recently, Long et al. proposed a method that can learn transferable features with deep neural network and showed some impressive results on the benchmarks [11]. Model transfer approach assumes that the parameters of the model for the source task can be transferred to the target task. Yang et al. proposed Adaptive SVMs by transferring parameters by incorporating the auxiliary classifier trained from source domain [12]. On top of Yang's work, Ayatar et al. proposed PMT-SVM that can determine the transfer regularizer according to the target data automatically [13]. Tommasi

[1]In their original papers, this value is fixed to be 1. In our experiments, we found that this setting leads to degraded performance.

et al. proposed Multi-KT that can utilize the parameters from multiple source models for the target classes [2]. Kuzborskij et al. proposed a similar method to learn new categories by leveraging over the known source [7].

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 situation, knowledge should not be transferred). *How to transfer* the prior knowledge effectively should be carefully designed to prevent inefficient and negative transfer. Some previous work consists in using generative probabilistic method [14] [15] [16]. Bayesian learning methods can predict the target domain by combining the prior source distribution to generate a posterior distribution. Alternatively, some previous max margin methods show that it is possible to learn from a few examples by minimizing the Leave-One-Out (LOO) error for the training model [7] [17]. Cawley et al. show that there is a closed-form implementation of LOO cross-validation that can generate unbiased model estimation for LS-SVM [18].

Our work corresponds to the context above. In this paper, we propose SMTLe based on model transfer approach with LS-SVM. We address our work on how to prevent negative transfer while just accessing the source model for domain adaptation. Compared to other works, propose a new perspective to the previous work of HTL, which brings more insight to negative transfer. Then we propose a novel strongly convex objective function for transfer parameters estimation. We show that SMTLe can converge at the rate of $O(\frac{\log(t)}{t})$. By optimizing this objective function, SMTLe can autonomously adjust the transfer parameters for different hypotheses. We theoretically show that, without any data distribution assumption, the superior bound of the training loss for SMTLe is the loss of a method learning directly (i.e. without using any prior knowledge). As a result, SMTLe can achieve a better performance and alleviate negative transfer.

III. Insight into negative transfer

In this section, we focus on the Phrase I of HTL and introduce our biased regularization for binary LS-SVM for our problem. Inspired by previous work, we bring a new perspective to the previous work and analysis the reasons why negative transfer could happen.

We define our transfer task in the following way: Suppose we have N visual categories. In our source task, N source binary classifiers $f'_n(x)$ for $n = 1, ..., N$, are trained from a distribution $\mathcal{D}_s$ to distinguish whether an object belongs to each of the N categories. In our target task, we have another small set of data (x, y) drawn from another distribution $\mathcal{D}_t$ with the same N categories as those in source task. We want to train N target binary classifiers $f_n(x)$ for $n = 1, ..., N$ on the data of the target domain so that they can perform well on the target domain.

A. Biased regularization in HTL

From previous works of HTL, the source and target classifiers f follow the hypothesis space of all linear model, i.e. $f_n = w\phi(x) + b$, where $\phi(x)$ can be any feature mapping that maps the example into another space. The transfer learning process of each target binary classifier f_n can be formalized as the following optimization problem:

$$\textbf{min} \quad R(w_n) + \frac{C}{2}\sum_{i}^{l}\mathcal{L}(f_n(x_i), Y_{in}) \tag{1}$$

Here $R(w)$ is the regularization term to guarantee good generalization performance and avoid overfitting. Y_{in} is the encoded label for binary classifier following $Y_{in} = 1$ if $y_i = n$ and -1 otherwise. $\mathcal{L}(\cdot)$ is the loss function. When we consider to use Lease Square SVM as the classifier, $\mathcal{L}(f, y) = (f - y)^2$.

In previous works, such as Multi-KT [2] , assume that the hyperplane w_n of the target classifier should be closed to the weighted combination of the hyperplane of the source models. The optimization problem (1) can be written as:

$$\begin{aligned} \textbf{min} \quad & \frac{1}{2}\left\| w_n - \sum_{k=1}^{N} w'_k\beta_k \right\|^2 + \frac{C}{2}\sum_{i}^{l} e_{in}^2 \\ \textbf{s.t.} \quad & e_{in} = Y_{in} - w_n\phi(x_i) - b_n \end{aligned} \tag{2}$$

The optimal solution to problem (2) is:

$$w_n = \sum_{k}^{N}\beta_k w'_k + \sum_{i}^{l}\alpha_{in}\phi(x_i) \tag{3}$$

It is obviously that once we can determine the values of the weights β, we can solve the optimization problem (2).

B. Data Augumentation in HTL

We can interpret Eq. (3) in the following way. Let $w''_n = \sum_{i}^{l}\alpha_{in}\phi(x_i)$, we have $w_n = w''_n + \sum \beta_k w'_k$. Therefore, for the target binary classifier $f_n(x)$, we have:

$$\begin{aligned} f_n(x) &= w''_n\phi(x) + \sum_{k}^{N}\beta_k w'_k\phi(x) + b_n \\ &= w''_n\phi(x) + (b_n - \sum_{k}^{N}\beta_k b_k) + \sum_{k}^{N}\beta_k f'_k(x) \end{aligned} \tag{4}$$

Here, we call the weight β_k the transfer parameter. From Eq. (5) we can see that the decision of each target binary classifier is made by combining the decision from target task $w''_n\phi(x)$ and the decision scores of the source model. The transfer parameters here is to control the amount of the knowledge transferred from the source models.

We can rewrite Eq. (4) as:

$$\begin{aligned} f_n(x) = &[w''_n, \beta_1, ..., \beta_N][\phi(x), f'_1(x), ..., f'_N(x)]^T \\ &+ (b_n - \sum_{k}^{N}\beta_k b_k) \end{aligned} \tag{5}$$

Here, we propose a novel insight to the transfer problem. From Eq. (5), we can see that solving the optimization problem (2) is equivalent to find the optimal hyperplane $\hat{w} = [w''_n, \beta_1, ..., \beta_N]$

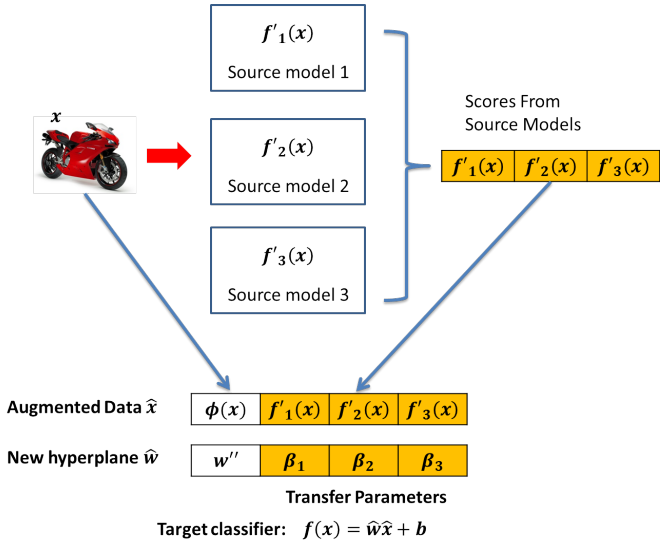

Fig. 3: The transfer learning process can be considered as the augmentation of the target data where the decision scores of the source models are appended as the auxiliary features. The transfer parameters can be considered as the a part of the corresponding hyperplane.

for the augmented data $\hat{x} = [\phi(x), f'_1(x), ..., f'_N(x)]$. Moreover, we can see that because the auxiliary features comes from the decision scores of the source models, we can greatly extend the choice of the source model. We can exploit the knowledge of any source model that can output the decision score of a example.

C. Reasons for negative transfer

From the perspective of the data augmentation, we can turn the problem of domain adaptation problem with HTL into a traditional learning problem, i.e. find the optimal values for the elements of the hyperplane hyperplane $\hat{w} = [w''_n, \beta_1, ..., \beta_N]$. According to the principle of Structural Risk Minimization (SRM) [19], the risk of a linear classifier $f(x) = wx + b$ on the unseen test data $R(f)$ (generalization risk) is bounded by:

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h(\ln(2l/h) + 1) + \ln(\delta/4)}{l}} \quad (6)$$

Here the first part on the right-hand side of the inequation $R_{emp}(f)$ is the empirical risk (training error) of the classifier and the second part is the confidence interval. h and l denote the VC dimension and number of training data of the classifier respectively and δ is the confidential parameter. According to [20], the VC dimension h is bounded by $h \leq \min([||w||^2 R^2], l) + 1$ where R is the radium of the smallest ball containing data x and $||w||$ is the *2-norm* of the hyperplane.

As we discussed above, we use the outputs of the source models as the auxiliary features to augment the target data. Let R and $\hat{R}$ denote the radiums of the data before and after augmentation. We should have $R^2 \leq \hat{R}^2$ and $||w||^2 \leq ||\hat{w}||^2$. This indicate that the VC dimension of the target model trained on the augmented data (augmented model) tends to increase compared to the model trained from the original data, i.e. method without transferring any source knowledge (no transfer model). As a result, data augmentation eventually increases the confidence interval of the risk of the augmented model. When the augmented model failed to decrease the empirical risk, its performance would degrade, i.e. suffer from negative transfer. For example, when the auxiliary features can't provide any extra useful information for classification, i.e. the source domain and target domain are unrelated, negative transfer could happen. In contrast, if we can significantly decrease the empirical risk of the augmented model, we can decrease its generalization risk and get improved performance, i.e. positive transfer.

However, for the original N categories, we already have their corresponding source category hypotheses and thus, their regularization term can be written as:

$$R(w_n, w'_n) = \frac{1}{2}\|w_n - \gamma_n {w'}_n\|^2 \quad (7)$$

As we can see that the regularization term (7) is a special case of (**??**) where only one β_k is none-zero.

Combining these two together, our multi-class incremental transfer problem can be solved by optimizing the following objective function:

Let $K(X, X)$ be the kernel matrix and

$$\psi = \begin{bmatrix} K(X,X) + \frac{1}{C}\mathrm{I} & \mathbf{1} \\ \mathbf{1^T} & 0 \end{bmatrix} \quad (8)$$

$$\psi \begin{bmatrix} \boldsymbol{\alpha}' \\ \boldsymbol{b}' \end{bmatrix} = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \quad \psi \begin{bmatrix} \boldsymbol{\alpha}'' \\ \boldsymbol{b}'' \end{bmatrix} = \begin{bmatrix} X(W')^T \\ 0 \end{bmatrix} \quad (9)$$

We have:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}' - \begin{bmatrix} \boldsymbol{\alpha}'' \boldsymbol{d_\gamma} & \boldsymbol{\alpha}'' \boldsymbol{\beta^T} \end{bmatrix} \quad (10)$$

Here $\boldsymbol{d_\gamma}$ is a diagonal matrix with $\{\gamma_i\}_{i=1,...,N}$ in its main diagonal. From Eq. (10) we can see that, the solution of Eq. (2) is completed once $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are set.

IV. SMTLE

V. Experiment

VI. Conclusion

References

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] T. Tommasi, F. Orabona, and B. Caputo, "Learning categories from few examples with multi model knowledge transfer," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 5, pp. 928–941, 2014.

[3] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 594–611, 2006.

[4] I. Kuzborskij and F. Orabona, "Stability and hypothesis transfer learning," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 942–950.

[5] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[6] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, p. 137, 2007.
[7] I. Kuzborskij, F. Orabona, and B. Caputo, "From n to n+ 1: Multiclass transfer incremental learning," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3358–3365.
[8] L. Jie, T. Tommasi, and B. Caputo, "Multiclass transfer learning from unconstrained priors," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1863–1870.
[9] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowledge-Based Systems*, vol. 80, pp. 14 – 23, 2015, 25th anniversary of Knowledge-Based Systems.
[10] J. J. Lim, "Transfer learning by borrowing examples for multiclass object detection," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.
[11] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France*, 2015, pp. 97–105.
[12] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 188–197.
[13] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2252–2259.
[14] J. Davis and P. Domingos, "Deep transfer via second-order markov logic," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 217–224.
[15] X. Wang, T.-K. Huang, and J. Schneider, "Active transfer learning under model shift," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1305–1313.
[16] T. Zhou and D. Tao, "Multi-task copula by sparse graph regression," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 771–780.
[17] T. Tommasi, F. Orabona, and B. Caputo, "Safety in numbers: Learning categories from few examples with multi model knowledge transfer," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3081–3088.
[18] G. C. Cawley, "Leave-one-out cross-validation based model selection criteria for weighted ls-svms," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 1661–1668.
[19] V. N. Vapnik, "An overview of statistical learning theory," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 988–999, 1999.
[20] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

ACKNOWLEDGMENT

The authors would like to thank...