AccGenSVM: Selectively Transferring from Previous Hypotheses

Diana Benavides-Prado
Dept. of Computer Science
The University of Auckland
dben652@aucklanduni.ac.nz

Yun Sing Koh
Dept. of Computer Science
The University of Auckland
ykoh@cs.auckland.ac.nz

Patricia Riddle
Dept. of Computer Science
The University of Auckland
pat@cs.auckland.ac.nz

Abstract

In our research, we consider transfer learning scenarios where a target learner does not have access to the source data, but instead to hypotheses or models induced from it. This is called the Hypothesis Transfer Learning (HTL) problem. Previous approaches concentrated on transferring source hypotheses as a whole. We introduce a novel method for selectively transferring elements from previous hypotheses learned with Support Vector Machines. The representation of an SVM hypothesis as a set of support vectors allows us to treat this information as privileged to aid learning during a new task. Given a possibly large number of source hypotheses, our approach selects the source support vectors that more closely resemble the target data, and transfers their learned coefficients as constraints on the coefficients to be learned. This strategy increases the importance of relevant target data points based on their similarity to source support vectors, while learning from the target data. Our method shows important improvements on the convergence rate on three classification datasets of varying sizes, decreasing the number of iterations by up to 56% on average compared to learning with no transfer and up to 92% compared to regular HTL, while maintaining similar accuracy levels.

1 Introduction

Hypothesis transfer learning (HTL) aids the learning of a new classification task by exploiting source hypotheses or models learned on previous tasks. HTL attempts to remedy drawbacks of typical transfer learning [Pan and Yang, 2010], since in general it requires availability of instances, features or parameters, and of domain adaptation [Daumé III, 2009; Duan *et al.*, 2009; Gong *et al.*, 2012; Hoffman *et al.*, 2012], a technique that requires source data during transfer. The source for transfer in HTL are prior models. Therefore, it is possible to transfer even when source data is unavailable or difficult to access.

In HTL source hypotheses are usually transferred as a whole. The new function is driven towards a linear combination of these sources. Most common methods propose to use source hypotheses for predicting target data [Yang *et al.*, 2007; Kuzborskij *et al.*, 2015; Mozafari and Jamzad, 2016; Wang and Hebert, 2016], to learn contributions or weights of source hypotheses [Tommasi *et al.*, 2014; Kuzborskij *et al.*, 2015; Wang and Hebert, 2016] or to assign importances to these sources [Yang *et al.*, 2007]. Some others transfer from a single source hypothesis [Oneto *et al.*, 2015].

The problem of partially using source hypotheses to aid learning of a new task has been, to the best of our knowledge, utterly ignored. Here we propose AccGenSVM, a technique that selectively transfers from source hypotheses trained with an SVM, the state-of-the-art technique in HTL. The transferred information corresponds to previous coefficients. This information is used as upper-bounds of the coefficients to be learned on the target task. We exploit the fact that source support vectors represent a summary of source data, and transfer coefficients only when we find similarities between source support vectors and target data.

We propose to treat the elements of source hypotheses as privileged information at training time [Vapnik and Izmailov, 2016]. Recent works relying on this concept cover a wide variety of cases and application areas (for example [Zhou *et al.*, 2016]). A learning algorithm provides privileged or additional information to encourage faster or more accurate learning. This information is commonly represented as additional features for the target data. Here we propose to represent it as upper-bounds on the coefficients that need to be learned on a target task. This decision is based on the fact that, in SVM, a higher coefficient is an indication of the importance of a data point in the final function [Cristianini and Shawe-Taylor, 2000].

AccGenSVM uses source hypotheses and transfers learned coefficients as a means to emphasize target data points lying close to source support vectors. Learning is still based on the target data, though subject to modified constraints as indicated by the extra-information. This strategy resembles importance weighting on training data [Lapin *et al.*, 2014], where different data points get distinct weights depending on its relevance for the objective function. The challenge of SVM learning with weighted training data is precisely how to establish these weights.

The Kullback-Leibler divergence metric is used to filter source hypotheses. This allows AccGenSVM to select source hypotheses that might contain source support vectors from

Method	What to transfer	When to transfer	How to transfer
A-SVM [Yang *et al.*, 2007]	Source coefficient vectors as a whole.	When a source hypothesis is a good predictor of target data.	As an additional term that represents a linear combination of previous hypotheses.
PMT-SVM [Aytar and Zisserman, 2011], similarly in [Oneto *et al.*, 2015]			Directly, one source to one target.
MMKT [Tommasi *et al.*, 2014],		As indicated by an additional term that minimizes the leave-one-out error.	As an additional term along with a linear combination of previous hypotheses.
GreedyTL [Kuzborskij *et al.*, 2015], used by [Valerio *et al.*, 2016]		As indicated by a greedy search on the hypotheses set.	As new features, and an additional term for their total contribution.
MT-SVM [Wang and Hebert, 2016]		As far as required by the optimization procedure.	As an additional term that needs to be learned.
HMCA [Mozafari and Jamzad, 2016]	One-dimensional vector.	As indicated by a similarity measure between source and target data.	Directly in one dimension.
AccGenSVM (ours)	Selected elements of source hypotheses.	When there exist source support vectors similar to target data.	As upper-bounds of coefficients to be learned, following principles of [Vapnik and Izmailov, 2016] and [Lapin *et al.*, 2014].

Table 1: HTL in previous works and ours.

which to transfer. We then compare source support vectors with the target data using a Fast k-Nearest Neighbors (FNN) method [Beygelzimer *et al.*, 2013], and make a decision of whether to transfer afterwards.

Our approach is useful when source data is scarce or difficult to access, and related hypotheses are available. Furthermore, this method can be very useful when the target data distribution is slightly distinct from the source hypotheses. For example, when small changes over time cause new concepts to arise or old concepts to evolve. AccGenSVM can identify which source support vectors resemble parts of the target data. Domains like image classification or object recognition for data captured with different devices, patient or customer classification at different locations, or simply models with changing distributions are good candidates for this method.

Our main contributions are:

- A selective HTL method that follows principles of learning with privileged information. This additional information corresponds to source data points defining the decision boundary, support vectors, and their learned coefficients. These are used for two purposes: to decide when to transfer and to reinforce the importance of relevant target data points. Hence, our formulation looks similar to learning with importance weighting on training data, and our method provides a means to learn these weights using source hypotheses and their support vectors.
- A method for HTL that selects and uses source support vectors as required. It treats every source hypothesis independently, and performs a transfer only when necessary, still learning the new function from the target data.
- A method that only relies on availability of source hypotheses. It does not require to learn new terms, to predict target data using source hypotheses nor to have source data available. AccGenSVM deals effectively with a large number of source hypotheses and varying sizes of transfer level data.

2 Related Work

Transfer learning (TL) has been an increasingly active research area over the last few decades. With an aim to achieve faster or more accurate learning [Pan and Yang, 2010], typical settings of TL are based on transferring instances (instance-based), finding shared feature representations between sources and target data (feature representation-based), transferring parameters (parameter-transfer) or transferring common knowledge (knowledge-based transfer), in homogeneous or heterogeneous domains. More recently, domain adaptation has captured a lot of attention as an option for learning in the presence of changing distributions. Most common domain adaptation solutions [Dai *et al.*, 2007; Daumé III, 2009; Duan *et al.*, 2009; Gong *et al.*, 2012; Hoffman *et al.*, 2012] require availability of source data since their main purpose is to correct or learn common representations between sources and target.

HTL has arisen as an alternative when models are available but source data is not. A number of theoretical [Kuzborskij and Orabona, 2013], experimental [Yang *et al.*, 2007; Aytar and Zisserman, 2011; Tommasi *et al.*, 2014; Kuzborskij *et al.*, 2015; Oneto *et al.*, 2015; Mozafari and Jamzad, 2016; Wang and Hebert, 2016] and application-specific methods [Valerio *et al.*, 2016] have been proposed. These works provide similar solutions for the problems of what, when and how to transfer (see Table 1). All of them use source hypotheses as black-boxes. Moreover, most of them require the prediction of target data to evaluate how well a source hypothesis fits the target task, whereas others establish the importance of each source hypothesis by learning some additional term or through a user-defined parameter.

One of the first and most well-known attempts in HTL, A-SVM [Yang *et al.*, 2007], learns a new function with an additional λ term that controls the contribution of source hypotheses. This term represents the predictions of target instances using these sources. Accurate predictions on the training data imply that the new function should be biased towards previous hypotheses. The importance of each hypothesis for the target task can be also decided based on a user-defined parameter or on some similarity feature. A similar method is proposed in PMT-SVM [Aytar and Zisserman, 2011], an evolution of A-SVM. As in [Oneto *et al.*, 2015], transfer is performed between one source hypothesis and one target.

Another method, GreedyTL, uses a greedy search to find the best set of source hypotheses for transferring [Kuzborskij *et al.*, 2015]. Still source hypotheses must be used to predict target data. As in [Tommasi *et al.*, 2014], an additional β term

must be learned while optimizing the objective function. A distinguishing characteristic of GreedyTL is that it finds the best combination of weights for source hypotheses, without learning a new function from the target data.

Mozafari and Jamzad [2016] simplify the HTL problem to a one-dimensional setting between one source hypothesis and one target task, for homogeneous and heterogeneous transfer. They removed the need to predict target data by means of source hypotheses or by learning additional terms. Instead, a similarity metric must be calculated between source and target data, and consequently the source data must be available.

Wang and Hebert [2016] establish the contribution of each source hypothesis through an additional term that must be learned during the optimization procedure. Although this term controls the contribution of each source hypothesis, still these hypotheses are treated as a whole without distinguishing among their elements and the information that these might distinctly contribute to the target task.

The opportunity for learning by selectively transferring from source hypotheses is still unexplored. If fragments of each hypothesis can be transferred when appropriated for a target task, the learning process can have flexibility to learn from target data while being supported by source hypotheses only as necessary. It is possible to find a solution to a learning problem by explicitly exploiting the commonalities between source hypotheses and target data, while still considering the particularities of the target data. This could lead to learning faster or with higher accuracy on the new task.

3 AccGenSVM

For a classification task, generally an instance space X and a set of labels Y are available. Every example (x, y) is said to be i.i.d. drawn from an unknown distribution P on $X \times Y$. The task is to learn a function $f := X \rightarrow Y$ from a set of hypotheses H, that predicts Y (class) well for new instances, with the lowest expected loss $\ell(f) := \mathbf{E}\ell(Y \cdot f(x))$. An SVM solution can be found by solving a (soft margin) optimization problem, by the dual function [Schölkopf and Smola, 2002]:

$$\begin{aligned} \max_{\alpha} F(\alpha) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ s.t. \sum_{i=1}^{n} y_i \alpha_i &= 0, \forall i \ 0 \leq \alpha_i \leq C \end{aligned} \tag{1}$$

where n is the number of data points, K corresponds to the kernel function and C is the upper-bound on coefficients α to be learned. In recent years, the concept of learning with privileged information has arisen as an option for faster or more accurate learning when additional information is available at training time [Vapnik and Izmailov, 2016]. The dual function for learning with privileged information can be optimized to [Pechyony and Vapnik, 2011; Lapin *et al.*, 2014]:

$$\begin{aligned} \max_{\alpha, \overline{\alpha}} F(\alpha) - \frac{1}{\gamma} \overline{F}(\overline{\alpha}) \\ s.t. 1^\top \overline{\alpha} = 0, \sum_{i=1}^{n} y_i \alpha_i = 0, \forall i \ 0 \leq \alpha_i \leq C + \overline{\alpha_i} \end{aligned} \tag{2}$$

where $F(\alpha)$ is obtained by Eq. 1 and $\overline{F}(\overline{\alpha})$ by a dual function that depends both on the privileged information space and the coefficients α (see [Pechyony and Vapnik, 2011; Vapnik and Izmailov, 2016] for details). The coefficients learned on the space of the privileged information, $\overline{\alpha}$, serve as upper-bounds of the coefficients α on the space of the new task.

A connection between learning with privileged information and learning with weighted training data using SVM has been theoretically demonstrated [Lapin *et al.*, 2014]. In weighted learning, the influence of a data point (x, y) on the final decision function can be increased or decreased using weights. The dual function to optimize can be, then:

$$\begin{aligned} \max_{\alpha} F(\alpha) \\ s.t. \sum_{i=1}^{n} y_i \alpha_i = 0, \forall i \ 0 \leq \alpha_i \leq c_i \end{aligned} \tag{3}$$

where $F(\alpha)$ is found by Eq. 1, now subject to the modified constraint $\alpha_i \leq c_i$, with c_i corresponding to the weight of a particular data point x_i. Lapin, Hein and Schiele [2014] proved that by setting $c_i = C + \overline{\alpha_i}$, weighted learning becomes similar to learning with privileged information.

Finally, from the representer theorem [Schölkopf *et al.*, 2001], a solution to a learning problem using SVM is:

$$f(x) = \sum_{i=1}^{n} \overline{\alpha_i} y_i K(\overline{x_i}, x) \tag{4}$$

where $\overline{\alpha_i}$ is the learned coefficient for a support vector $\overline{x_i}$ with class y_i. x is the set of new points to be predicted using this function. We rely on this hypothesis representation for the transfer task: from every source hypothesis, their support vectors $\overline{x}$ will be used to decide when to transfer, given their similarity with the target data, while their coefficients $\overline{\alpha}$ will be transferred and used to obtain upper-bounds on the coefficients to be learned (as stated in Eqs. 5 and 6 below).

3.1 Problem Formulation for Selective HTL

Given a possibly large set of source hypotheses, $\overline{H}$, each hypothesis obtained with SVM by Eq. 1 and represented as in Eq. 4, the problem of HTL is that of using $\overline{H}$ to aid learning of a new function. This function is typically learned by regularizing the distance between the set of coefficients learned on the target data, α, and a linear combination of the coefficients learned for the source hypotheses, $\overline{\alpha}$. This solution is not selective among source hypotheses, and usually requires the prediction of target data using previous hypotheses, which is expensive when a large number of hypotheses are available.

AccGenSVM uses an alternative selective approach for transferring from source hypotheses. Each source support vector from the hypothesis set $\overline{H}$ can distinctly contribute to the new learning task, and therefore their coefficients $\overline{\alpha}$ can be transferred individually. This contribution is represented as upper-bounds on coefficients to be learned, and consequently the problem becomes similar to weighted learning with SVM and learning with privileged information [Lapin *et al.*, 2014]:

$$\max_{\alpha} F(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$s.t. \sum_{i=1}^{n} y_i \alpha_i = 0, \forall i \ 0 \leq \alpha_i \leq C + c_i, c_i = \sum_{k=1}^{sv} \overline{\alpha}_k \quad (5)$$

For a specific coefficient α_i, a modified constraint, c_i, is then learned from the privileged information, as represented by coefficients $\overline{\alpha}$ from the corresponding subset of similar source support vectors of size sv. This subset will be determined as explained in our solution. The new constraint can be learned from the source hypotheses, and allows some target data points to contribute more to the maximization problem, thus implying their importance for the target task. We introduce a term, $s_i/|\overline{H}|$, that balances the expected importance of a target data point:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$s.t.i \sum_{i=1}^{n} y_i \alpha_i = 0, \forall i \ 0 \leq \alpha_i \leq C + c_i, c_i = \frac{s_i}{|\overline{H}|} \sum_{k=1}^{sv} \overline{\alpha}_k \quad (6)$$

where s_i corresponds to the number of previous hypotheses contributing to c_i, as found by our method, from the set of available source hypotheses of size $|\overline{H}|$. As in weighted learning with SVM, the AccGenSVM solution is unique and satisfies the convexity requirements according to the Karush-Kuhn-Tucker conditions [Lapin *et al.*, 2014].

Figure 1 provides a graphical example of our selective HTL method. $\overline{h}_1$ and $\overline{h}_2$ are source hypotheses with different decision boundaries. For a new target task, our approach is able to identify common regions between target data and source hypotheses, and then to transfer coefficients as explained. Regions that are specific to the new data will be learned by the regular SVM, thus preserving the particularities of target data during the new learning task.

3.2 Selective HTL

A solution to Eq. 6 can be summarised in two phases. The first one is intended to select source hypotheses and elements for transfer, from a possibly large number of these sources; the second phase transfers coefficients and obtains new upper-bounds for relevant target data points. Our approach is detailed in Algorithm 1, and works as follows.

Phase 1 - Selection of elements for transfer.

1. Based on sequential minimal optimization (SMO) as an SVM solver [Bottou and Lin, 2007], AccGenSVM selects a pair of candidate target data points, x_i and x_j, one positive and one negative in binary classification. At every iteration, this method will gather privileged information from source hypotheses for these two data points.
2. The set $\overline{H}$ of source hypotheses is filtered using KL divergence. KL divergence is an information-theoretic metric for determining how divergent two probability distributions are. Every source hypothesis is compared to the target distribution, and then the subset of source hypotheses below a threshold is selected.

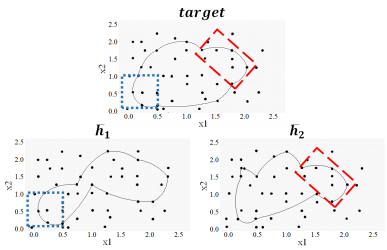

Figure 1: An example for selective HTL

3. FNN [Beygelzimer *et al.*, 2013] is used to find the source support vectors, $\overline{x}$, that more closely resemble the two target data points selected in Step 1, using the subset of source hypotheses selected in Step 2. Feature selection is performed as a previous step to speed up FNN.

Phase 2 - Transfer.

1. As a result from Phase 1, AccGenSVM obtains c_i and c_j, upper-bounds for α_i and α_j based on the coefficients $\overline{\alpha}$ from similar source support vectors.
2. AccGenSVM considers the number of contributing source hypotheses and balances new upper-bounds as in Eq. 6. The more source hypotheses contributing to an upper-bound, the higher this upper-bound will be for the corresponding target data point.

Algorithm 1: Pseudo-code for transfer with AccGenSVM

Data: Target data X, source hypotheses $\overline{H}$
Result: Upper-bounds of coefficients to be learned $\alpha_i up$, $\alpha_j up$
$x_i, x_j \leftarrow$ selected points by working set selection
$y_i, y_j \leftarrow$ corresponding classes for selected points
$c_i, c_j \leftarrow 0; s_i, s_j \leftarrow 0; \alpha_i up, \alpha_j up \leftarrow C$
// Phase 1
forall $\overline{h}$ *in* $\overline{H}$ **do**
  if $KL(\overline{h}, X) \leq threshold$ **then**
    forall $\overline{x}_h$ *in* $\overline{h}$ **do**
      if $\overline{x}_h$ *close to* x_i **and** $y_h = y_i$ **then**
        $c_i \leftarrow c_i + \overline{\alpha}_h$
      end
      else if $\overline{x}_h$ *close to* x_j **and** $y_h = y_j$ **then**
        $c_j \leftarrow c_j + \overline{\alpha}_h$
      end
    end
    $s_i \leftarrow s_i + 1; s_j \leftarrow s_j + 1$
  end
end
// Phase 2
$\alpha_i up \leftarrow C + \frac{s_i}{|\overline{H}|} c_i$
$\alpha_j up \leftarrow C + \frac{s_j}{|\overline{H}|} c_j$

The computational complexity of learning an SVM using SMO is $i * O(n)$, with i iterations and n target data points [Chang and Lin, 2011]. The FNN step increases this to a worst case $i * O(n * log(N))$. Here n depends on the num-

ber of points evaluated, which is smaller for faster convergence rates. Filtering hypotheses adds a constant $O(|\overline{H}|)$ that depends on the number of available source hypotheses. In terms of convexity, the algorithm behaves like learning an SVM with weighted training data [Lapin *et al.*, 2014].

4 Experiments

Here, we present experiments and discuss results for binary classification using AccGenSVM and other publicly available HTL methods for homogeneous transfer.

4.1 Datasets

Caltech256 is a benchmark dataset on image recognition, with 30,607 instances and 256 independent classes plus a clutter class. We exclude the clutter class, and work with 29,780 instances. ImageNet is a large benchmark dataset, with 14,197,122 instances and around 21,841 different classes. We work with 117 classes and 163,666 instances. Office is a small dataset that contains images from three different domains, with 93 classes and 4,652 instances. The number of instances per class varies for each dataset (see Table 2). For all datasets, we use available bag-of-words representations containing 1,000 features [Tommasi and Tuytelaars, 2014]. Features were scaled to [-1,1].

4.2 Experimental Set-Up

For each dataset, we extract binary random samples without replacement in two levels:

- Transfer level: we extract random samples of sizes 10%, 20% and 30% for training. We select 3 classes as positives and the rest as negatives, in a binary fashion, from each dataset. Positive classes are selected by their frequency (most frequent, least frequent and moderately frequent). We also select independent random binary samples of size 20%, 30%, and 50% as test sets. We perform 30 repetitions for all sampling configurations.
- Source hypotheses level: we extract 10%, 20% and 30% binary random samples for training hypotheses as sources. Positive classes of these samples are different, but related, to one of our three positives on the transfer level. Though for experiments we explicitly select related positives on this level, it is important to note that this step is not always necessary. AccGenSVM can filter hypotheses according to their relatedness with the transfer level data, as explained in Algorithm 1. KL divergence is used as a pre-step to select related positives from the original datasets. We also perform 30 repetitions for this sampling procedure. Models for these samples are trained using regular SVM.

For simplicity, we select 10 random samples for every positive class and every sample size of every dataset at the transfer level. We then train 10 models, one for each sample, using AccGenSVM, previous works and scenarios of learning without transfer. Experiments are performed using different percentages of hypotheses as sources (i.e. using 10%, 25%, 50% and 100% of the available hypotheses), selected randomly. These models are tested on corresponding test samples. We perform this procedure for every repetition (i.e. 30 times).

Method	Office (min. 7, max. 100)	Caltech256 (min. 80, max. 800)	ImageNet (min. 121, max. 2252)
TH	87.02 ± 1.64	89.88 ± 2.22	79.25 ± 3.63
SH	72.99 ± 7.19	73.08 ± 7.45	72.57 ± 7.48
S+TH	77.02 ± 6.64	79.88 ± 5.37	76.25 ± 7.89
A-SVM	87.02 ± 2.10	90.99 ± 2.67	80.68 ± 4.43
GreedyTL	78.31 ± 5.83	89.64 ± 6.83	$--$
AccGenSVM	85.66 ± 2.08	93.11 ± 1.50	83.94 ± 3.10

Table 2: Prediction accuracy of AccGenSVM and other methods, using all source hypotheses (100%). Maximum and minimum number of instances per class for the original datasets are also shown.

Based on a sensitivity analysis of the regularization parameter C and the γ parameter for the RBF kernel, we set $C = 1$ and $\gamma = 1/f$, with f number of features. We use a KL divergence threshold of 0.3 for all datasets. For FNN, we work with 3 nearest neighbours.

AccGenSVM[1] is built on top of LibSVM [Chang and Lin, 2011], using available KL divergence [Hausser and Strimmer, 2014] and FNN implementations [Beygelzimer *et al.*, 2013].

4.3 Results

Our results are compared against learning from models trained for samples on the transfer level using regular SVM (TH), models trained for samples on the source hypotheses level (SH) and models trained for samples on the transfer and corresponding source hypotheses level (S+TH). Results are also compared to two available previous works, A-SVM[2] [Yang *et al.*, 2007] and GreedyTL[3] [Kuzborskij *et al.*, 2015]. We provide further discussion for other methods.

Table 2 shows average results for prediction accuracy, for each dataset, using all of the available source hypotheses. For Caltech256, AccGenSVM outperforms TH by around 3%, SH by 20%, SH+T by 13%, A-SVM by 2% and GreedyTL by around 4%.

For ImageNet, AccGenSVM surpasses TH and A-SVM by similar percentages as Caltech256, while outperforms SH by around 12% and SH+T by around 7%. This might be due to tighter relations between classes on the original ImageNet dataset, which make SH and SH+T better predictors for transfer level data. As an example, in our experiments with ImageNet we had *coffee mug* as a positive class on the transfer level, and *mug* as a positive class on the source hypotheses level, two highly related classes. For Caltech256, we had *airplanes* class on the transfer level. This class is known to be related with various classes on the original dataset [Griffin *et al.*, 2007] but not so strongly. It is important to remark that our selective method can transfer in both cases.

For ImageNet, methods for weighting sources such as GreedyTL fail to find a solution for large sample sizes on

[1] Software available at: https://github.com/nanarosebp/PhDProject/tree/master/AccGenSVM

[2] We use the maximum supported by A-SVM software, 5 sources.

[3] Since this method does not train a new model, we do not measure its convergence rate.

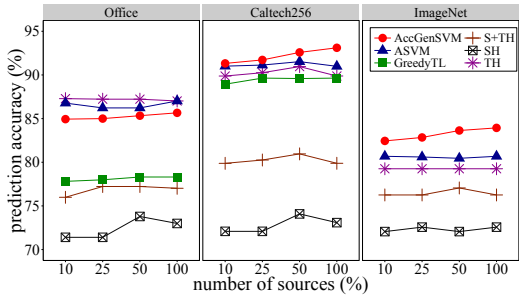

Figure 2: Prediction accuracy of our method vs. learning with no transfer (TH, SH, S+TH) and previous works. AccGenSVM is in the group with higher accuracy for small and medium-size datasets.

	Method	10% of sources	25% of sources	50% of sources	100% of sources
Office	TH	38.76 ± 7.40	39.07 ± 7.87	38.79 ± 7.66	40.04 ± 7.44
	A-SVM	118.07 ± 9.99	118.07 ± 9.99	118.07 ± 9.99	121.07 ± 9.69
	AccGenSVM	25.37 ± 9.00	25.07 ± 9.38	24.99 ± 8.87	24.37 ± 9.58
Caltech256	TH	55.05 ± 9.94	55.11 ± 10.74	55.11 ± 10.93	54.85 ± 9.94
	A-SVM	286.32 ± 138.57	284.45 ± 134.44	259.33 ± 131.30	265.36 ± 140.33
	AccGenSVM	24.22 ± 8.67	21.55 ± 9.09	21.41 ± 8.01	20.37 ± 8.41
ImageNet	TH	228.04 ± 46.20	230.42 ± 47.10	228.04 ± 48.46	229.47 ± 47.19
	A-SVM	1210.00 ± 258.46	1179.40 ± 242.38	1139.01 ± 233.19	1210.00 ± 233.19
	AccGenSVM	185.14 ± 30.53	178.78 ± 27.84	178.77 ± 26.26	171.09 ± 28.23

Table 3: Convergence rate measured as the number of iterations before a solution is found, for the three datasets, using different percentages of available source hypotheses.

the transfer level ("–" in Table 2). For Office, AccGenSVM results lie on the border of TH and A-SVM. Error rates of AccGenSVM are similar to TH and A-SVM, on the three datasets. Methods involving models based on source data (SH), source and transfer level data (S+TH) and source weighting (GreedyTL) have higher standard deviations. The more source hypotheses, the better our results (see Figure 2).

We also measure the convergence rate as the number of iterations before a solution is found. For the three datasets, AccGenSVM is able to find a solution with fewer iterations compared to its counterparts (TH and A-SVM). Table 3 shows an average decrease of up to 56% versus TH for at least one dataset, and up to 92% versus a regular HTL method. Even worst-case scenarios based on error rates show that AccGenSVM can learn at least 68% faster than regular HTL. A faster convergence rate can be achieved with AccGenSVM even with a small number of source hypotheses. In terms of execution time, our method performs similarly for small sample sizes. These results position our work as an effective implementation of the concept of privileged information, where it is expected that additional information facilitates faster convergence [Vapnik and Izmailov, 2016]. Similar accuracy levels with a faster convergence rate also indicate a higher slope in the learning curve [Tommasi *et al.*, 2014].

Finally, we test the stability of AccGenSVM in changing distributions. These tests are focused on Caltech256. We generate concept drift by randomly changing attributes values and then moving the distribution of the transfer level data (i.e. changing a given number of attributes values for a given number of instances, both of them randomly selected, and as indicated by the level of concept drift). From Figure 3, as the

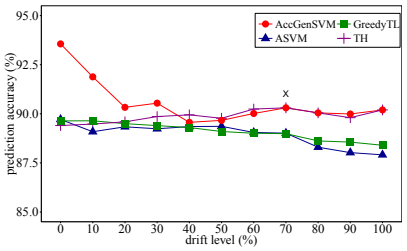

Figure 3: Concept drift stability of AccGenSVM at different drift levels (X is the drift level where transfer stops).

level of drift increases, our method achieves a prediction accuracy closer to learning with no transfer. Furthermore, when the distribution on the transfer level reaches a high percentage of concept drift (70%), AccGenSVM stops transferring and performs as learning with no transfer. As the transfer level data changes, it becomes more dissimilar to the source hypotheses, and hence a selective method like ours will choose not to transfer. Highly changing data distributions have a negative effect on regular HTL methods, since they use entire source hypotheses.

Discussion on other methods. Some of the existing TL methods address challenges like finding common feature representations, a problem that is not within the scope of our method [Pan and Yang, 2010]. One of the most referenced works, [Daumé III, 2009], is known to underperform when the source and the target data are very similar in terms of features [Pan and Yang, 2010], a strength in our case. GFK [Gong *et al.*, 2012], for instance, uses KL divergence for ranking source domains, though still requires to compare source data and target data. Others like DTMKL [Duan *et al.*, 2012] are based on similar concepts to regular HTL: learning linear combinations of sources (data, in this case) and learning additional terms. Other HTL approaches like [Mozafari and Jamzad, 2016] still rely on source data, while [Wang and Hebert, 2016] uses A-SVM to solve one of its minimization sub-problems, and therefore will depend on its results.

5 Conclusion and Future Work

We have proposed a selective approach for HTL, that distinguishes and selects parts of source hypotheses to transfer. Our method relies on the concept of privileged information and gives flexibility for a new task to learn a model with available training data. AccGenSVM maintains the prediction accuracy for TL scenarios with training data of varying sizes. Faster convergence rates can be achieved when learning a new model is supported by privileged information, represented in previous hypotheses and their support vectors.

A research avenue is to reduce the execution time of AccGenSVM for large datasets, as well as to analyse its generalization properties. A feasible extension is the consolidation of new and old concepts within a broader concept learning system, as supported by selective transfer methods. The problem of selectively transferring between heterogeneous domains is also a promising area. Finally, we aim to extend our approach to lifelong learning problems where source hypotheses are made available over time.

References

[Aytar and Zisserman, 2011] Yusuf Aytar and Andrew Zisserman. Tabula rasa: Model transfer for object category detection. In *2011 ICCV*, pages 2252–2259. IEEE, 2011.

[Beygelzimer *et al.*, 2013] A Beygelzimer, S Kakadet, J Langford, S Arya, D Mount, and S Li. FNN: Fast nearest neighbor search algorithms and applications. R package version 1.1, 2013.

[Bottou and Lin, 2007] Léon Bottou and Chih-Jen Lin. Support vector machine solvers. *Large scale kernel machines*, pages 301–320, 2007.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[Cristianini and Shawe-Taylor, 2000] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[Dai *et al.*, 2007] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *ICML*, pages 193–200. ACM, 2007.

[Daumé III, 2009] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[Duan *et al.*, 2009] Lixin Duan, Ivor W Tsang, Dong Xu, and Tat-Seng Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, pages 289–296. ACM, 2009.

[Duan *et al.*, 2012] Lixin Duan, Ivor W Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE TPAMI*, 34(3):465–479, 2012.

[Gong *et al.*, 2012] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE CVPR*, pages 2066–2073, 2012.

[Griffin *et al.*, 2007] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical report, 2007.

[Hausser and Strimmer, 2014] Jean Hausser and Korbinian Strimmer. *entropy: Estimation of Entropy, Mutual Information and Related Quantities*, 2014. R package v.1.2.1.

[Hoffman *et al.*, 2012] Judy Hoffman, Brian Kulis, Trevor Darrell, and Kate Saenko. Discovering latent domains for multisource domain adaptation. In *European Conference on Computer Vision*, pages 702–715. Springer, 2012.

[Kuzborskij and Orabona, 2013] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *ICML*, pages 942–950, 2013.

[Kuzborskij *et al.*, 2015] Ilja Kuzborskij, Francesco Orabona, and Barbara Caputo. Transfer learning through greedy subset selection. In *International Conference on Image Analysis and Processing*, pages 3–14. Springer, 2015.

[Lapin *et al.*, 2014] Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: SVM+ and weighted SVM. *Neural Networks*, 53:95–108, 2014.

[Mozafari and Jamzad, 2016] Azadeh Sadat Mozafari and Mansour Jamzad. A SVM-based model-transferring method for heterogeneous domain adaptation. *Pattern Recognition*, 56:142–158, 2016.

[Oneto *et al.*, 2015] Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Shrinkage learning to improve SVM with hints. In *IJCNN*, pages 1–9. IEEE, 2015.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[Pechyony and Vapnik, 2011] Dmitry Pechyony and Vladimir Vapnik. Fast optimization algorithms for solving SVM+. *Statistical Learning and Data Science*, 4:3–24, 2011.

[Schölkopf and Smola, 2002] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[Schölkopf *et al.*, 2001] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

[Tommasi and Tuytelaars, 2014] Tatiana Tommasi and Tinne Tuytelaars. A testbed for cross-dataset analysis. In *European Conference on Computer Vision*, pages 18–31. Springer, 2014.

[Tommasi *et al.*, 2014] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Learning categories from few examples with multi model knowledge transfer. *IEEE TPAMI*, 36(5):928–941, 2014.

[Valerio *et al.*, 2016] Lorenzo Valerio, Andrea Passarella, and Marco Conti. Hypothesis transfer learning for efficient data computing in smart cities environments. In *IEEE SMARTCOMP*, pages 1–8. IEEE, 2016.

[Vapnik and Izmailov, 2016] Vladimir Vapnik and Rauf Izmailov. Learning with intelligent teacher. In *Symposium on Conformal and Probabilistic Prediction with Applications*, pages 3–19. Springer, 2016.

[Wang and Hebert, 2016] Yu-Xiong Wang and Martial Hebert. Learning by transferring from unsupervised universal sources. In *AAAI*, pages 2187–2193, 2016.

[Yang *et al.*, 2007] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *ACM International Conference on Multimedia*, pages 188–197. ACM, 2007.

[Zhou *et al.*, 2016] Joey Tianyi Zhou, Xinxing Xu, Sinno Jialin Pan, Ivor W Tsang, Zheng Qin, and Rick Siow Mong Goh. Transfer hashing with privileged information. *arXiv preprint arXiv:1605.04034*, 2016.