Transfer Learning with an Ensemble of Background Tasks

Zvika Marx, Michael T. Rosenstein, Leslie Pack Kaelbling
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{zvim,mtr,lpk}@csail.mit.edu

Thomas G. Dietterich
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR 97331
tgd@cs.orst.edu

Abstract

We demonstrate the transfer of learning from an ensemble of background tasks, which becomes helpful in cases where a single background task does not transfer well. This approach is accomplished through a simple maximum a posteriori elaboration on the logistic regression approach and tested on real world data.

1 Transfer Learning Via Learned Prior Distributions

Transferring knowledge from a familiar domain or task (call it *task A*) to an unfamiliar or newly-encountered one (*task B*) is a fundamental and fascinating aspect of human learning. Although the motivating notion is intuitive, the simple approach of treating the two tasks as identical and pooling their training data does not usually work well. This is presumably because the decision boundaries for A and B are not in exactly the same places in the feature space, even when the feature spaces and input distributions are themselves identical. Hence, more sophisticated methods are required.

One interesting approach is to treat task A as defining a form of Bayesian prior distribution for task B. In this paper, we study this approach in a setting where we have many task As and only one task B, and we study whether we can learn a useful prior from those multiple task As that gives effective guidance when learning task B.

Consider the well-known logistic regression model,

$$P(y=1|\mathbf{x}) = \frac{1}{1+\exp[w_0 + \sum_{j=1}^{n} w_j x_j]}$$

where y is the class label, $\mathbf{x}$ is a vector of n features, the w_j are real-valued weights, and $P(y=0|\mathbf{x}) = 1 - P(y=1|\mathbf{x})$. A standard way of fitting this model is to assume an independent Gaussian prior on the weight values. That is, each weight is drawn from a Gaussian distribution: $w_j \sim N(\mu_j, \sigma_j)$. The standard way of fitting this model is to maximize the penalized log likelihood

$$\sum_{i=1}^{N} \log P(y_i | \mathbf{x}_i) - \sum_{j=0}^{n} \left(\frac{w_j - \mu_j}{2\sigma_j} \right)^2$$

Typically, the values $\mu_j = 0$ and $\sigma_j = \sigma$ are employed, with σ (a constant value > 0) set by holdout or cross-validation methods [3]. σ_0 is typically set to be relatively large to avoid penalizing the intercept weight w_0. The model can be fit via improved iterative scaling [1].

In the application described in this paper, we have available data from K different "tasks". We propose to select one of these to be task B and use the remaining $K - 1$ tasks to learn the values of μ_j and σ_j. Using these learned values, we then fit a logistic regression model to task B via penalized maximum likelihood. Specifically, we fit $K - 1$ independent logistic regressions (with prior mean 0 and $\sigma = 1$) and obtain fitted weights $\{w_j^k\}_{j=0}^n$, $k = 1, \ldots, K-1$ (the superscript k indicates the training task k). From these, we then estimate as follows:

$$\mu_j = \frac{1}{K-1} \sum_{k=1}^{K-1} w_j^k \quad \text{for } j = 0, \ldots, n$$

$$\sigma_j = \sqrt{\frac{1}{K-2} \sum_{k=1}^{K-1} (w_j^k - \mu_j^k)^2} \quad \text{for } j = 0, \ldots, n$$

This strategy is similar to previous work by Chelba and Acero [2]. However, they fit the μ_j to auxiliary data based on a single A task and set the σ_j to a constant σ tuned with holdout data. In our application (and in many transfer learning settings), there is not enough task B data to employ holdout methods.

2 Experimental Test

We tested this algorithm on a meeting invitation task that we call the "Busy People" task. Data were collected from 21 individuals for the task of deciding whether or not to accept an email invitation to a meeting. The data were generated by first collecting two months of calendar data and associated background knowledge from each individual. The background knowledge included definitions of the various projects that the person was working on, the other people working on those projects, and the relationships between the individual and these other people. Synthetic email invitations addressed to each individual were then generated for a two-week period (overlapping the two-month calendar), and the individual then classified each invitation as to whether he or she would accept or reject that invitation (independently of the other invitations but based on the actual state of the calendar).

Eight of the participants generated their data as part of a *military* simulation in which they were involved. For these people, the background knowledge was defined as part of the simulation, and the calendar information was collected during the simulation. The remaining 13 participants were all *researchers* in university or private research labs, and they provided the background knowledge and calendar information as described above.

We performed the following experiment 21 times and averaged the results: One individual was chosen to be Task B, and the remaining individuals constituted 20 Task As. 20 logistic regression classifiers were fit to these domain As. From the fitted weight values, the μ_j and σ_j values were estimated as described above. For Task B, data from 32 examples were chosen (randomly) for training and the rest were set aside as the test set. For training sets of size m= 2, 4, 8, 16, and 32, the Task B classifier was fit by penalized maximum likelihood and then the resulting model was evaluated on the holdout test data. The results are plotted in Figure 1.

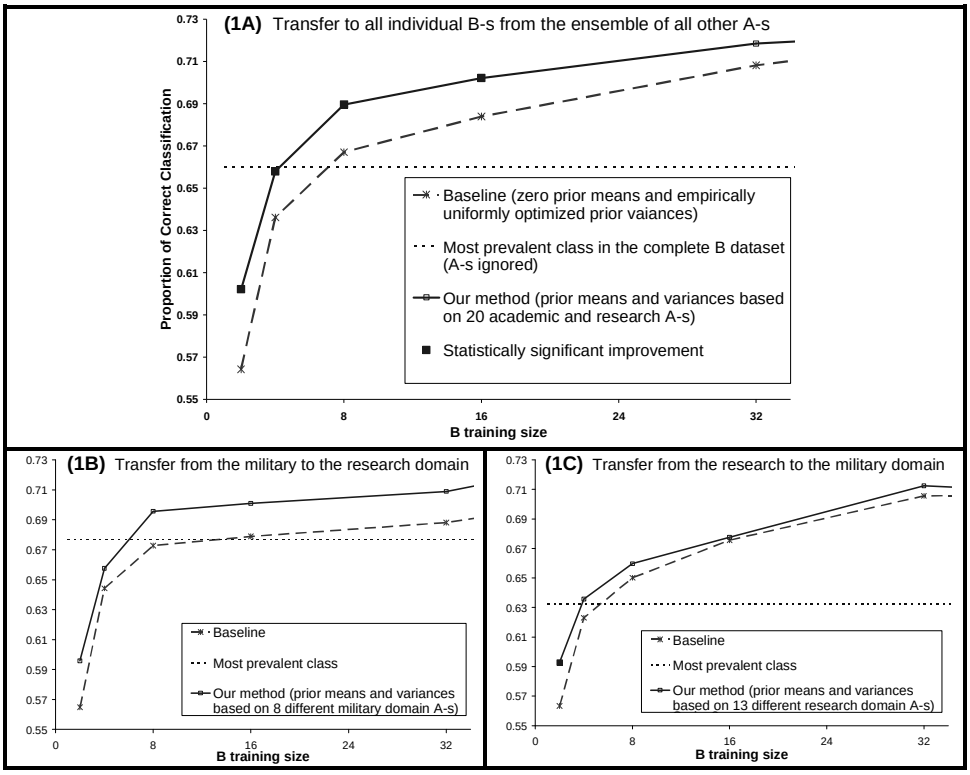

Figure 1: Our method improves on average over a transfer-unaware baseline. The numbers shown refer to averages over individual B tasks. The ensemble of A tasks consists of the remaining 20 individuals (1A), the eight military domain (1B), and the 13 research domain individuals (1C). In 1A, differences relative to the baseline are mostly statistically significant (one tailed paired t test, $p < .05$). In the cross-domain transfer cases (1B, 1C), most of these differences are not statistically significant.

Figure 1A shows the overall results compared to training only on the Task B data. The transfer algorithm shows improvement that is statistically significant compared to the baseline. Figure 1B shows the results if we transfer from the 8 military participants to the 13 research participants. We still observe a positive transfer effect, although a paired differences t test does not report statistical significance, perhaps because of the small number of replications. Figure 1C shows the results of transferring from the 13 researchers to the 8 military participants. There is a

positive effect (not statistically significant) only at very small sample sizes. Results from the within-domain transfer, to each participant from the rest of participants of the same type, were very close to the results of transfer from all participants, that is the improvement was mostly statistically significant (not shown in Figure 1).

3 Discussion

This work demonstrates that when multiple A tasks are available, we can obtain positive transfer learning by first fitting models to the individual A tasks, and then using those fitted models to estimate the parameters of an informative prior distribution for task B. This strategy can be viewed as a rough approximation of a full hierarchical Bayesian approach in which we adopt a hyperprior over the weights and assume that prior for each task A or B is drawn from this shared hyperprior. Such an approach can be expected to give better results than the simple strategy employed here.

In a related paper [4], we define a hierarchical naïve Bayes classifier and apply it to this same problem. However, in those experiments, we only studied transfer between pairs of study participants, rather than between the whole set of As and a single B. Those results (as well as applying the strategy described in [2] to transfer between two individuals), while positive, were not as strong as the results reported here for logistic regression. This suggests that transferring logistic regression weight values may be better suited to this meeting invitation task than transferring the naïve Bayes parameter values.

It is important to be aware of the assumptions necessary for any transfer method to work and to examine the degree to which they match reality. Our assumption here was that the tasks we dealt with are of the same kind or, in other words, that they were generated from a common source or taken from a common pool. Thus, although prominent differences prevented individual transfer, valuable transferable information could be extracted from the ensemble. This should not be expected to work under all circumstances, e.g. if there are two or more domains that are fundamentally different. With this observation in hand, an important further research direction seems to be to detect cross-task and cross-domain similarities and relevances (e.g. through unsupervised mechanisms such as data clustering).

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under Contract No. NBCHD030010. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

[1] Berger, A. (1997) The Improved iterative scaling algorithm: a gentle introduction. Technical report, Carnegie Mellon University.

[2] Chelba C. & Acero A. (2004) Adaptation of maximum entropy capitalizer: Little data can help a lot. In D. Lin and D. Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 285-292.

[3] Chen S. & Rosenfeld R. (2000) A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, **8**(1):37-50.

[4] Rosenstein M. T., Marx Z., Pack Kaelbling L. & Dietterich T. G. (2005) To Transfer or Not To Transfer, *Inductive Transfer: 10 Years Later NIPS 2005 Workshop*.