Multiclass Continuous Correspondence Learning

Brian D. Bue
Electrical and Computer Engineering
Rice University
Houston, TX 77006
brian.d.bue@rice.edu

David R. Thompson
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
david.r.thompson@jpl.nasa.gov

Abstract

We extend the Structural Correspondence Learning (SCL) domain adaptation algorithm of Blitzer et al. [4] to the realm of continuous signals. Given a set of labeled examples belonging to a "source" domain, we select a set of unlabeled examples in a related "target" domain that play similar roles in both domains. We define a mapping into a common feature space using the pivot sample, which allows us to adapt a classifier trained on source examples to classify target examples. We show that when between-class distances are relatively preserved across domains, we can automatically select target pivots to bring the domains into correspondence, allowing us to adapt a classifier trained on source data, to classify target data.

1 Structural Correspondence Learning for Continuous Spaces

We extend the Structural Correspondence Learning (SCL) algorithm of Blitzer et al. [4] to the realm of continuous signals. SCL is a domain adaptation technique which creates a mapping between a "source" domain consisting of labeled examples, and an unlabeled "target" domain using a set of "pivot features" common to both domains. In text classification scenarios, these consist of terms (words) that serve similar roles in both domains, so that the role of other features can be inferred by correlation with these features. We extend this concept to continuous domains where the objects to be classified are continuous-valued functions, making it applicable to data such as time series or electromagnetic spectral signatures.

Recent work by Balcan et al. [1] provides an elegant method to define a correspondence mapping in continuous spaces. They illustrated that desiging a good feature space is similar to designing a good kernel function, and under certain conditions, a kernel which approximately preserves the margin of a max-margin separator can be constructed using a set of unlabeled samples. By projecting samples into a space defined by (distances to) the unlabeled samples, one can potentially harness the power of a high-dimensional kernel mapping in this lower-dimensional feature space. In this vein, we define our correspondence mapping with pivot *samples*, rather than pivot *features*. As with pivot features, pivot samples are samples which serve similar roles in both domains, and are diverse enough to characterize the nuances of the classification task.

Determining a mapping between domains is closely related to the topic of manifold alignment. Most manifold alignment algorithms assume some knowledge of the target domain in the form of paired (source to target) correspondences [**?**], [**?**] or a small number of labeled target examples [**?**], to define a transformation that reconciles the feature spaces, but recent work (e.g., [**?**]) determines the correspondence mapping automatically by matching local geometric properties across feature spaces.

In this work, we present Multiclass Continuous Correspondence Learning (MCCL): a domain adaptation technique for high-dimensional continuous data. We leverage structured relationships between a set of known source classes to automatically select a set of pivot samples to translate between source and target domains.

2 Domain Adaptation and Classification with MCCL

2.1 Approach Overview

We assume we have N labeled examples (X^S, Y^S) drawn from a "source" distribution $\mathcal{D}^S$ to train a predictor to classify M unlabeled examples X^T drawn from a "target" distribution $\mathcal{D}^T$ (assumed available at training time). We assume the two distributions share a set of classes with labels $Y = \{1, \ldots, K\}$, and examples $\mathbf{x}$ in both domains are F-dimensional vectors, where F is the number of features. We use the following transformation function to map a sample $\mathbf{x}$ to the feature space defined by pivots $\mathbf{p}_i \in P$ (we hereafter refer to this feature space as the "R-space").

$$R(\mathbf{x}, P) = \left(\frac{\mathrm{d}(\mathbf{x}, \mathbf{p}_1)}{\sum_{\ell=1}^{Q} \mathrm{d}(\mathbf{x}, \mathbf{p}_\ell)}, \ldots, \frac{\mathrm{d}(\mathbf{x}, \mathbf{p}_Q)}{\sum_{\ell=1}^{Q} \mathrm{d}(\mathbf{x}, \mathbf{p}_\ell)} \right) \tag{1}$$

Algorithm 1 describes the Multiclass Continuous Correspondence Learning Algorithm (MCCL). Given source pivots P^S, we select target pivots P^T which best preserve the (relative) relationships between source pivots (Step 1). Then, we train a multiclass predictor using the transformed source samples (Step 2) to classify the transformed target samples (Step 3).

Algorithm 1 Multiclass Continuous Correspondence Learning (MCCL)

```
Input: source training data $(X^S, Y^S)$, target data $X^T$, source pivots $P^S$.
Output: predicted target labels $Y^T$
1: Build target pivot set $P^T$ from $X^T$ by selecting best matching target pivot, $\mathbf{p}_i^T = \mathbf{x}_\ell^T$, for each
   source pivot $\mathbf{p}_i^S \in P^S$ according to $\ell = \operatorname*{argmin}_i \|R(\mathbf{p}_i^S, P^S) - R(\mathbf{x}_i^T, P^S)\|,\ i \in \{1, \ldots, M\}$
2: Train a multiclass predictor in the R-space $p : R(\mathbf{x}, P) \to Y$ using $R^S = (R(\mathbf{x}_i^S, P^S))_{i=1}^N$.
3: return Prediction vector $Y^T = (p(R(\mathbf{x}_i^T, P^T)))_{i=1}^M,\ \mathbf{x}_i^T \in X^T$.
```

We use a technique inspired by the H-divergence [3] to measure the quality of the pivots (P^S, P^T). The (empirical) H-divergence measures the difference between two distributions by finding a classifier which separates samples drawn from one distribution from samples drawn from the other. As low H-divergence scores indicate we cannot distinguish between domains, we seek a set of pivots with small average per-class H-divergence. We describe the Pivot Divergence (Pdiv) function below.

Algorithm 2 Pivot Divergence (Pdiv)

```
Input: pivot sets $(P^S, P^T)$, length $Q = \sum_{k=1}^K Q_k$
Output: divergence score $H$.
1: for $k = 1$ to $K$ do
2:     Define label vector $y = ((-1)_{i=1}^{Q_k}, (1)_{i=1}^{Q_k})$ for pivot samples belonging to class $k$.
3:     Train binary predictor $h : R(\mathbf{p}, P) \to \{-1, 1\}$.
4:     Calculate divergence between class $k$ source and target pivots
       $H_k = \frac{1}{2Q_k} \left( \sum_{i=1}^{Q_k} \mathcal{I}(h(\mathbf{p}_i, P^S) = y_i) + \sum_{i=Q_k+1}^{2Q_k} \mathcal{I}(h(\mathbf{p}_i, P^T) = y_i) \right)$
5: return $H = \frac{1}{K} \sum_{i=1}^K H_k$
```

2.2 Evaluation Methodology

We consider several classification contexts to evaluate the performance of the MCCL algorithm. First, we calculate the baseline "within-domain" source (S) and target (T) classification accuracies. The maximum of these provides an approximate upper bound on the best achievable accuracy with domain adaptation (ST), while the minimum of these gives a lower bound we expect to improve by reconciling the source and target spaces. Next, we calculate the classification accuracy in the R-space defined by Q_k pivots per class sampled from *labeled* source and target data (R-S, R-T, and R-ST, respectively), which measures the change in accuracy induced by the R-transform when labels are available in both domains. Last, we calculate the accuracy using Algorithm 1 (R*-ST) to select target pivots using only labeled source data. In the R-space cases, we select the set of Q_k samples nearest to each class mean as the source pivots P^S. We classify samples using the multiclass (one-vs-one) Support Vector Machine implemented in the LIBSVM package [5], with 5 fold cross-validation. We estimate the slack parameter C via grid search over values $\{10^{-2}, \ldots, 10^2\}$.

2.3 Synthetic Data Example

We first provide an illustrative example on a simple synthetic dataset, shown in Figure 1 (left two plots), along with classification accuracies for the cases defined in section 2.2. Each class consists of 500 samples drawn from one of four 2D Gaussians. The mean of each target Gaussian (bottom plot) is a randomly perturbed version of its corresponding source mean (top plot). Diamond markers indicate the $Q_k = 50$ selected source/target pivots. On the right we have the source (top, offset for clarity) and target (bottom) class means μ_i^S, μ_i^T in the R-space $R(\mu_i, P)$, where P is the set of pivots in the corresponding space (ordered by class membership). Visually, the means in the R-space appear better reconciled than in the original feature space, though not perfectly so due to the non-linear transformation between the two domains (particularly classes 2 (cyan) and 3 (yellow)). Despite this, we see a significant improvement in accuracy in the R-space cases (R-ST and R*-ST) over the baseline (ST) accuracy, as the transformed samples better capture the structured class relationships in each domain.

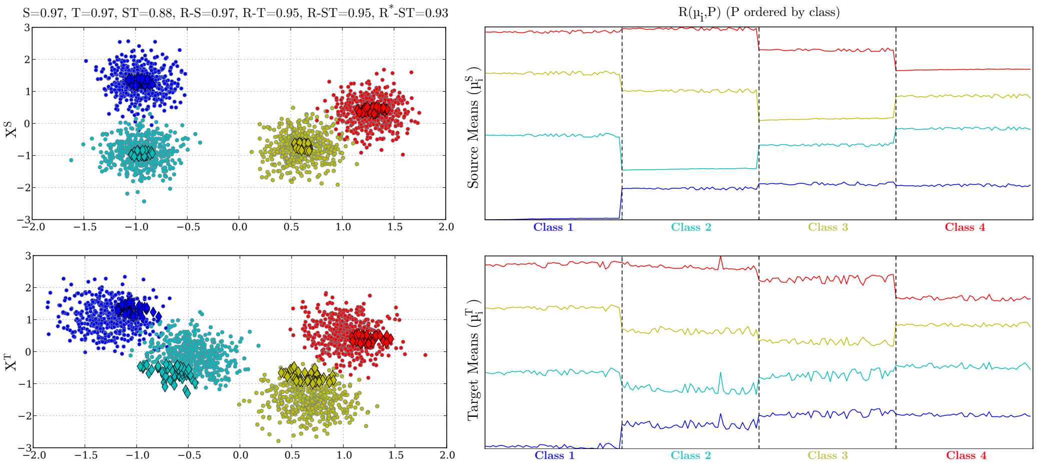

Figure 1: Left: 4 class synthetic source (top) and tartet (bottom) data. Right: means of the source pivots (top, offset for clarity) and the selected target pivots (bottom) in the R-space $R(\mu_i, P)$

2.4 Domain Adaptation for Hyperspectral Imagery

We next evaluate our algorithm on a domain adaptation problem in the hyperspectral imaging domain. Here, our goal is to classify a set of mineralogical samplestaken from one image, using training data from another image which was captured under different conditions. Our data consists of five mineralogical classes manually labeled by an expert geologist from two hyperspectral images of the Cuprite mining district in Cuprite, NV. The first image (hereafter referred to as "Av97") was captured in June 19, 1997 by the AVIRIS instrument, consists of 512×614 samples ("pixels"), and was studied in detail in [6]. The second image (hereafter referred to as "Hyp11") was captured on Feb. 06, 2011 by the Hyperion instrument onboard the EO-1 satellite, and contains 1798×779 pixels. Each pixel is a 29-dimensional vector of image radiance values measured at wavelengths in the range 2.1029-2.3249μm – an expert-defined range which captures relevant mineralogical characteristics for the selected classes. Preprocessing steps for each image include: atmospheric calibration (i.e., conversion from spectral radiance to surface reflectance) and illumination normalization (i.e., scaling each pixel by its L^2 norm). Also, because the smallest image consists of over 300,000 pixels, we segment each image using the technique described in [7]. The means of the resulting segments form the set of target pixels we use to build the target pivot set P^T.

Due to differences in sensor types, environmental conditions, capture dates, and atmospheric calibration techniques, identical classes are represented differently in each image. Scaling each sample by its L^2 norm accounts for scale differences to some degree, but can be further reconciled by applying whitening filters to each image. Figure 2 shows the whitened class means in each image. However, as we show in subsequent sections, these steps generally do not allow for robust class knowledge transfer between images.

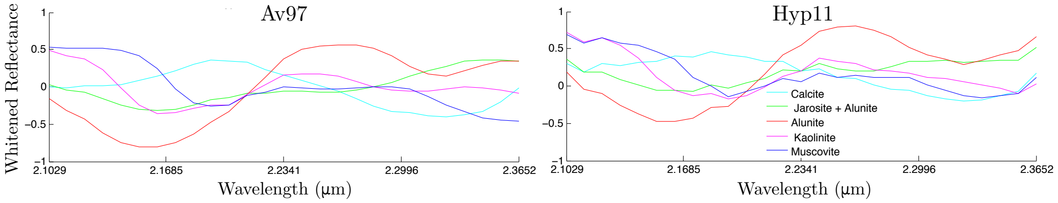

Figure 2: Whitened class means for Av97 (left) and Hyp11 (right) images.

We consider two domain adaptation scenarios: in the first, we train a classifier using the Av97 image as the source data and test the classifier using the Hyp11 image as the target data. We refer to this scenario as "Av97⇒Hyp11." In the second scenario we use the Hyp11 data as the source, with Av97 as the target data. We refer to this scenario as "Hyp11⇒Av97." Figure 3 gives classification accuracies and Pdiv scores with respect to the number of pivots per class Q_k. In both scenarios, we see significant improvements in accuracy in the domain adaptation cases (R-ST and R^*-ST) over the baseline (ST). We also observe that selecting pivots using Algorithm 1 (R^*-ST) yields comparable results to using labeled pivots (R-ST) for domain adaptation. However, in the Av97⇒Hyp11 scenaro, we see worse domain adaptation performance than in the Hyp11⇒Av97 scenario, along with a larger gap between the R-ST and R^*-ST results. Because the mapping between domains is defined by the source pivots, if the classes are more separated in the target domain, then in the source (which is the case in the Hyp11⇒Av97 scenario), the mapping performs well, even if the source classes aren't well-separated. However, if the target data is less separable than the source (which occurs in the Av97⇒Hyp11 scenario), then the source pivots may not capture enough info to resolve the ambiguous target classes.

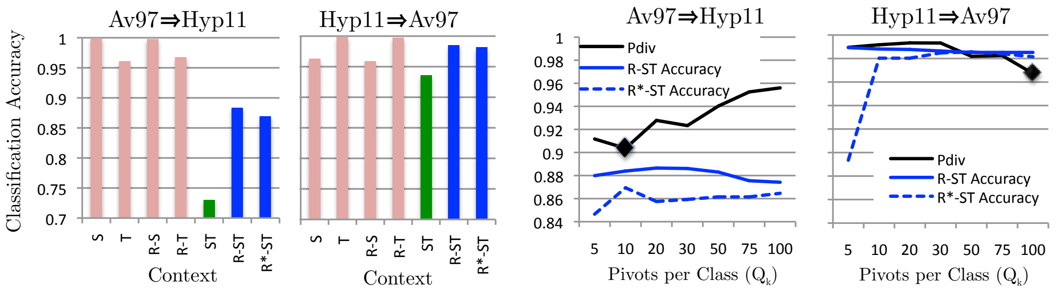

Figure 3: Classification accuracies for contexts described in Section 2.2 (left two plots) and Pdiv scores vs. pivots/class Q_k (right two plots) for Av97⇒Hyp11 and Hyp11⇒Av97 scenarios. Black diamonds indicate the best Pdiv score for the R^*-ST context, with accuracy shown in the right two plots.

For the Av97⇒Hyp11 scenario, $Q_k = 10$ attains the minimum Pdiv value, where we observe the maximum R^*-ST classification accuracy. Also, Pdiv increases with Q_k while accuracy remains relatively constant, indicating that additional pivots determined by well-separated Av97 data do not improve domain adaptation. In the Hyp11⇒Av97 scenario, while we see a gradual decrease in Pdiv for increasing Q_k – with slight improvements in accuracy, the Av97 classes are well separated for mid-range Q_k values $\in \{10, \ldots, 50\}$. For small Q_k, we observed low accuracy in all of R-S, R-T and R-ST cases, indicating the pivot set is inadequate to describe the classificatin task. We can filter such degenerate cases by ensuring that the R-space accuracy on the source data (R-S) is approximately the same as in the original feature space (S) (an approach also described in [2]). This allows us to define a lower limit on the number of pivots necessary to define a feature space expressive enough for domain adaptation. We note that the accuracy on the within-domain cases (S, T) are approximately equivalent to their corresponding R-space cases (R-S, R-T) when Q_k is sufficiently large.

References

[1] M Balcan, A Blum, and S Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, pages 79–94, Jan 2006.

[2] S Ben-David. Inductive transfer via embeddings into a common feature space. In *Open House on Multi-Task and Complex Outputs Learning*, July 2006.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach Learn*, 79(1-2):151–175, May 2010.

[4] J Blitzer, R McDonald, and F Pereira. Domain adaptation with structural correspondence learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Jan 2006.

[5] C.C Chang and C.J Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[6] FA Kruse, JW Boardman, and JF Huntington. Comparison of airborne hyperspectral data and EO-1 hyperion for mineral mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 41(6):1388–1400, 2003.

[7] David R Thompson, Lukas Mandrake, Martha S Gilmore, and R Castaño. Superpixel endmember detection. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–19, Jun 2010.