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. Using these "pivot samples," we map both domains into a common feature space, allowing 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.

1 Structural Correspondence Learning for Continuous Feature Spaces

We extend the Structural Correspondence Learning (SCL) algorithm of Blitzer et al. [4] to 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. We extend this concept to continuous domains where the objects we classify are continuous-valued functions, making SCL 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 between continuous feature spaces. They illustrated that designing 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 a similar vein, we define our correspondence mapping using distances to canonical samples, or *pivot samples*. These distances become the pivot features we use to reconcile differences between the source and target domains.

Determining a mapping between domains is closely related to the topic of manifold alignment. Most manifold alignment algorithms assume knowledge of the target domain in the form of paired (source to target) correspondences [11], [13] or a number of labeled target examples [8], to define a transformation that reconciles the feature spaces, but recent work (e.g., [12]) determines the correspondence mapping automatically by matching local geometric properties across feature spaces.

This work presents Multiclass Continuous Correspondence Learning (MCCL): a domain adaptation technique for high-dimensional continuous data. In previous work [5], [6], we demonstrated the feasibility of a similar domain adaptation technique for continous data – specifically, hyperspectral imagery. In this work, we show that by exploiting structured relationships between a diverse set of source classes, we can automatically select a set of pivot samples to reconcile differences between source and target domains.

1.1 Domain Adaptation and Classification with MCCL

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). The distributions share a set of classes with labels $Y = \{1, \ldots, K\}$. Without loss of generality we assume the samples $\mathbf{x}$ in both domains are F-dimensional vectors, where F is the number of features. We use the following transformation 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 distance relationships between source pivots (Step 1). Next, we train a multiclass predictor using the transformed source samples (Step 2), in order to classify the transformed target samples (Step 3). We evaluate

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 $\ell = \underset{i}{\operatorname{argmin}} \, \|R(\mathbf{p}_i^S, P^S) - R(\mathbf{x}_i^T, P^S)\|, \; i \in \{1, \ldots, M\}$
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 $Y^T = (p(R(\mathbf{x}_i^T, P^T)))_{i=1}^M, \; \mathbf{x}_i^T \in X^T$.

the quality of the correspondence mapping defined by the pivot set using a technique inspired by the H-divergence [3]. The (empirical) H-divergence measures the difference between two distributions by finding a classifier which separates samples drawn from either. Low H-divergence scores indicate we cannot distinguish between samples drawn from either domain, so 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), each of 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
$H_k = \frac{1}{2Q_k} \left(\sum_{i=1}^{Q_k} \mathcal{I}(h(\mathbf{p}_i^S, P^S) = y_i) + \sum_{i=Q_k+1}^{2Q_k} \mathcal{I}(h(\mathbf{p}_i^T, P^T) = y_i) \right)$
5: **return** $H = \frac{1}{K} \sum_{i=1}^{K} H_k$

2 Evaluation on Synthetic Data, Hyperspectral Imagery

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. In the naive class knowledge transfer context (ST), we simply train a classifier on the (whitened) source data to classify the (whitened) target data, which gives a lower bound we expect to improve. Next, we calculate accuracy using the R-space transform defined by Q_k pivots per class sampled from *labeled* source and target data (R-S, R-T, and R-ST, respectively). This 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) which selects target pivots using labeled source data only. In the R-space cases, we select the 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 [7], with 5 fold cross-validation. We select slack parameter C via grid search over $\{10^{-2}, \ldots, 10^{2}\}$.

Synthetic Data: We first provide an illustrative example on a synthetic dataset, shown in Figure 1 (left two plots). 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 (pivots ordered by class membership). Visually, the R-space class means appear better reconciled than in the original

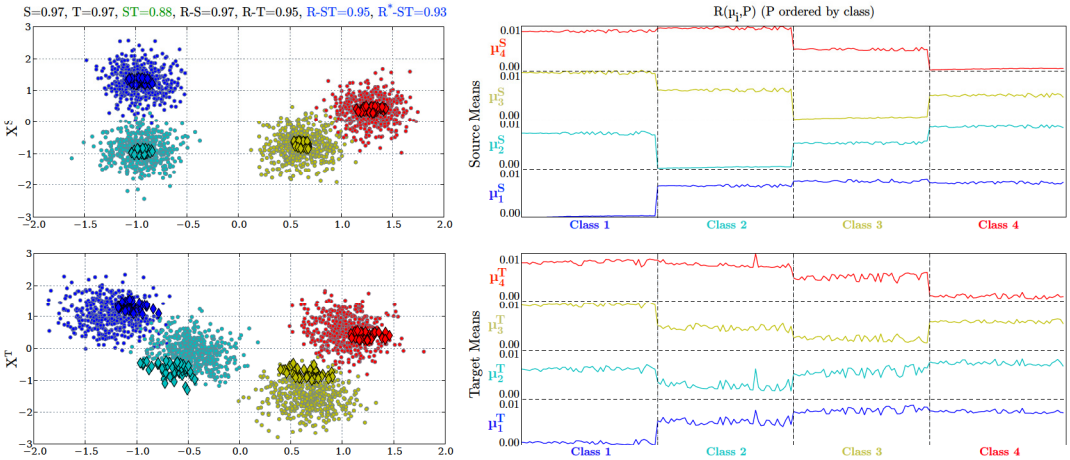

Figure 1: Left: 4 class synthetic source (top) and target (bottom) data. Right: source class means (top) and target class means (bottom) in the R-space $R(\mu_i, P)$

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=0.95 and R*-ST=0.93) over the baseline (ST=0.88).

Hyperspectral Imagery: We next evaluate our algorithm on hyperspectral image data. Here we address the task of classifying a set of mineralogical samples taken from one image using training data from another image captured under different conditions – a problem highly relevant to global hyperspectral mapping and analysis tasks. Our data consists of five mineralogical classes manually labeled by an expert geologist from two images of the Cuprite mining district in Cuprite, NV. Image *Av97* was captured in June 19, 1997 by the AVIRIS instrument, consists of 512×614 pixels, and was studied in detail in [9]. Image *Hyp11* was acquired 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. The domain adaptation task is particularly challenging for these images, as many unique mineralogical signatures appear in this region. We preprocess the images with atmospheric calibration (i.e., conversion from spectral radiance to surface reflectance) and illumination normalization (i.e., scaling each pixel by its L^2 norm). The smallest image consists of over 300,000 pixels, so we also segment each image using the technique described in [10]. We select the target pivots $\mathbf{p}_i^T \in P^T$ from the set of means of the resulting segments.

Identical classes appear differently in each image due to differences in sensor type, environmental conditions, capture dates, and atmospheric calibration. Scaling each sample by its L^2 norm accounts for some scale differences, and whitening filters further reconcile these scenes. Figure 2 shows the whitened class means in each image. However, as we show in subsequent sections, these steps alone are insufficient for robust class knowledge transfer between images. We consider two domain adaptation scenarios. First we train a classifier using the Av97 image as the source data and test the classifier on target data from the Hyp11 image. 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). Selecting pivots

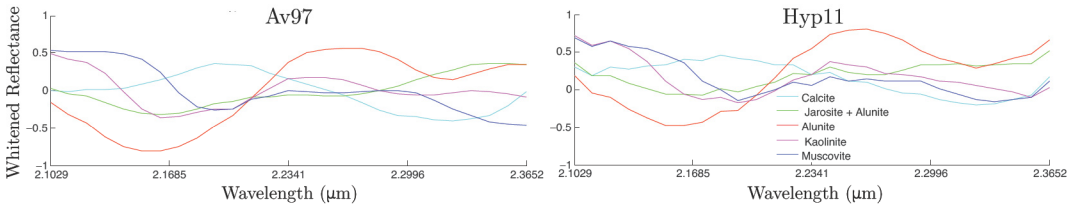

Figure 2: Whitened class means for Av97 (left) and Hyp11 (right) images. Sample counts for each class are as follows: Calcite: 1076, Jarosite+Alunite: 55, Alunite: 336, Kaolinite: 382, Muscovite: 425.

using Algorithm 1 (R*-ST) yields comparable results to using labeled pivots (R-ST) for domain adaptation. However, in the Av97⇒Hyp11 scenario, we see worse domain adaptation performance along with a larger gap between the R-ST and R*-ST results. Recall that the mapping between domains is defined by the source pivots, so if the classes are better separated in the target domain then in the source (e.g. the Hyp11⇒Av97 scenario), the mapping performs well. However, if the target data is less separable than the source (e.g. the Av97⇒Hyp11 scenario), then the source pivots may not discriminate ambiguous target classes.

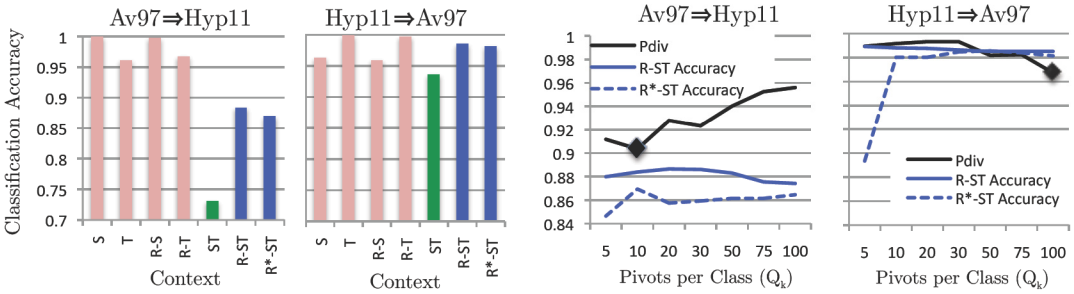

Figure 3: Classification accuracies for contexts described in Section 2 (left two plots) and Pdiv scores vs. pivots/class Q_k (right two plots) for Av97⇒Hyp11 and Hyp11⇒Av97 scenarios. Black diamonds indicate the best Pdiv score for the R*-ST context yielding the classification accuracy in the left two plots.

For the Av97⇒Hyp11 scenario, $Q_k = 10$ attains the minimum Pdiv value, where we also observe the maximum R*-ST classification accuracy. Also, Pdiv increases with Q_k while the R*-ST accuracy remains relatively constant, indicating that additional pivots determined by the Av97 source 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 classification 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 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 ($Q_k \geq 10$). We also note that when target labels are available for domain adaptation (R-ST), we achieve relatively high accuracy even for small Q_k.

3 Conclusions and Future Work

In this paper, we provided an extension to structural correspondence learning in continuous domains built upon our previous work in domain adaptation [5], [6], and provided a methodology to automatically select pivot samples to reconcile differences between domains. We show empirically that when between-class distances are preserved across domains, our automated pivot sample selection technique performs competitively to the case when labeled target samples are available to define a mapping between domains. In future work we will investigate the theoretical relationship between the implicit kernel mapping described in [1] to the R-transform (Equation 1) in the contexts of multiclass classification and domain adaptation.

Acknowledgements: A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, with support from an AMMOS Multimission Operations Systems technology development program. Copyright 2011 Rice University. All Rights Reserved. U.S. Government Support Acknowledged.

References

[1] M-F Balcan, A Blum, and S Vempala. On kernels, margins and low-dimensional mappings. *Proc. of Algorithmic Learning Theory 2008*, pages 1–12, Apr 2008.

[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] B Bue and E Merényi. Using spatial correspondences for hyperspectral knowledge transfer: evaluation on synthetic data. *Proceedings of the 2rd IEEE WHISPERS*, Jun 2010.

[6] B Bue, E Merényi, and B Csathó. An evaluation of class knowledge transfer from real to synthetic imagery. *Proceedings of the 3rd IEEE WHISPERS*, Jun 2011.

[7] 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.

[8] J Ham, D Lee, and L Saul. Semisupervised alignment of manifolds. *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence, Z. Ghahramani and R. Cowell, Eds*, 10:120–127, 2005.

[9] 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.

[10] 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.

[11] C Wang and S Mahadevan. Manifold alignment using procrustes analysis. *Proceedings of the 25th international conference on Machine learning*, pages 1120–1127, 2008.

[12] C Wang and S Mahadevan. Manifold alignment without correspondence. *Proceedings of the 21st International Joint Conferences on Artificial Intelligence*, 2009.

[13] Deming Zhai, Bo Li, Hong Chang, Shiguang Shan, Xilin Chen, and Wen Gao. Manifold alignment via corresponding projections. *Proceedings of the British Machine Vision Conference*, pages 3.1—3.11, Jan 2010.