Bibliography Notes

Geoffrey Iyer

January 23, 2018

1 Wei Feng Spectral Multiplicity Tolerant Inexact Graph Matching [?]

Title says it all really. I don't think it's useful to me. At least not for a while.

2 Conte Survey Article on Graph Matching [2]

No real notes as of yet. There are 5 papers here on spectral methods in inexact matching but they are relatively old (like year 2000). This whole survey article was released in 2004 so it's hard to get super excited about it.

3 Sedighin Microphone Source Separation [6]

Multimodal audio visual. Idea is that two people are speaking in a single mic. How to separate?

4 Ali Multimodal Imaging Graph Cuts [1]

I didn't read carefully. They do something like what we're doing but the do it locally, and there's a like a flow thing going on???

5 Notes for Möller paper [5]

I think we're not using this at all

6 Notes for Wang paper [7]

Have K different input data sets $X_1, \ldots, X_K$. Not necessarily related (not same features or dimension). Some data is labeled but not all. Idea: map all X_k to the same "latent space". Likely lower dimensional.

For each X_k, let $V_k = \{v_1, \ldots, v_{l_k}\}$ be the set of labeled data. Create a similarity matrix W_S for *all* data points (from every set X_j). Similarity is 1 if elements are labeled as same class, 0 otherwise (including unlabeled elements). Next make W_d where similarity is 1 if elements are from different classes. Lastly make W_k for $1 \leq k \leq K$ the standard similarity matrix on X_k.

Cost function is $\frac{A+C}{B}$. Here A comes from W_s, C comes from W_k, and B comes from W_d. Idea is minimizing A represents keeping classes together. Minimizing C represents preserving topology from individual X_k, and maximizing B keeping different classes apart. Can do standard graph laplacian stuff to find a solution.

7 Notes for Yeh paper [9]

Overall idea: take two sets X^s, X^t, and project them into a lower dimensional latent space. Here we are working under the assumption that we have a lot of labels in X^s, and few to no labels in X^t. So the goal is to take our understanding of X^s and transfer it over to X^t.

Update 11/10/2016: The method assumes that X^s, X^t contain the same number of data samples, and that there is some kind of correspondence between them (ex: sampled at the same time using different methods). In particular, the order of the data points matters. If X^t is a permuted version of X^s you will get wild results.

Fundamental equation: trying to find projection vectors u^s, u^t to maximize:

$$\rho = \frac{{u^s}^T X^s {X^t}^T u^t}{\sqrt{{u^s}^T X^s {X^s}^T u^s}\sqrt{{u^t}^T X^t {X^t}^T u^t}}.$$

Can turn this into an eigenproblem. Large eigenvalues correspond to good correlation. Number of eigenvectors chosen = dimension of latent space (called correlation subspace).

This model is entirely linear. To deal with nonlinearity can apply a nonlinear kernel function to X^s and X^t. This gives sets K^s and K^t. Perform same maximization on those. I understand the rough idea here, but I don't understand the details of the choice of kernel function. How do I pick the kernel function? Do I need to first do some machine learning on each individual X^s and X^t? This paper uses Gaussians, and uses a k-means to train the σ.

Computing kernel can be expensive for large sets. Use "Reduced Kernel". Same idea as Nyström approximation.

Once we've created the projection maps, apply SVM to proj(X^s) (recall: X^s is the set that has labels). Get a binary classifier. Use this classifier on proj(X^t). But instead of standard SVM, add a term to the energy function that encourages the algorithm to use the features with a high correlation coefficient ρ.

In the example experiments near the end, it seems like they first perform feature extraction on the individual X^s and X^t, then afterwards use the algorithm proposed in this paper.

8 Notes for Meng paper [4] (Andrea's group)

This paper has some algorithms for graph-based machine learning methods. They don't mention multimodal data at all.

For semi-supervised learning, this paper uses the energy

$$E(u) = \text{Graph-Laplacian} + \text{Double-Well-Potential} + \text{Fidelity-Term}.$$

Energy is minimized by alternating between the two steps

- Step 1: Take a gradient-descent step (partially implicit partially explicit)
- Step 2: Threshold to vertices of simplex

For unsupervised learning, the paper replaces the fidelity term with a term that encourages u to be piecewise constant with $\hat{n}$ total pieces (where $\hat{n}$ is the desired number of classes in the segmentation). The energy-minimization step is pretty similar to the semi-supervised case.

The rest of the paper talks about algorithms for computing (Nystrom Extension, Parallelization). They can handle 329 frames, each of which is a 128 × 320 hyperspectral image (with 129 bands), in about 3 minutes. That's pretty impressive.

9 Notes for Lafon paper [3]

This paper has a lot of material so I split it into sections.

9.1 Diffusion Distance

Given a data set X, form the usual Graph Laplacian

$$p_1(x,y) = \frac{w(x,y)}{d(x)}.$$

Think of this as a pmf for a random walk (chance to move from x to y is $p_1(x,y)$). Now think of Markov Chains and have $p_t(x,y)$ be the t-th power. Let

$$\phi_0(y) = \lim_{t\to+\infty} p_t(x,y)$$

the limiting distribution (note $\phi_0(y) = \frac{d(y)}{\sum_z d(z)}$). Then we define the *diffusion distance* between points x, z as

$$D_t^2(x,z) = \sum_{y\in\Omega} \frac{\left(p_t(x,y) - p_t(z,y)\right)^2}{\phi_0(y)}.$$

Can do the standard graph-laplacian trick and represent the data by only the first few eigenvectors. Even better, as t grows fewer eigenvectors are needed (spectrum decay).

The also define here Ψ_t the "diffusion map", which is just the projection on to eigenvectors of graph laplacian

$$\Psi_t(x) = \lambda_1^t \psi_1(x) + \cdots + \lambda_{m(t)}^t \psi_{m(t)}(x).$$

Here ψ_j are ordered eigenvectors, and $m(t)$ is the number needed to accurately represent your data.

9.2 Data Density

In section 2.2 they make a very interesting point about the density of the data. Depending on the sampling method, it's possible that the data will have nonuniform density in your manifold. This will affect the graph-laplacian feature extraction process. If we have two different sensors (with different manifold densities) this could really affect our data merging. The paper suggests a change in the weight function when building the graph laplacian. Can prove that it has good limiting behavior wrt density.

9.3 Nyström Extension and Multiscale Extension

Note: This isn't the same as the Nyström that I'm used to. They're taking the embedding map Ψ_t defined on the data Ω and extending it to a bigger map $\tilde{\Psi}_t$ defined on $\tilde{\Omega} \supseteq \Omega$. But I honestly didn't understand this section. They are assuming $\Omega \subseteq \mathbb{R}^d$, then using a Gaussian kernel to extend their representation from Ω to all of $\mathbb{R}^d$. There are some issues with the scale of the Gaussian (which made no sense to me), and the authors solve it by doing the work at multiple different scales until an appropriate one is found (try $\sigma = \sigma_0$. If it doesn't work, do the work on $\sigma_1 = \sigma_0/2$. Repeat).

9.4 Data matching algorithm

Suppose we have sets Ω_1, Ω_2, and suppose we have landmarks $x_1, \ldots, x_k \in \Omega_1$, $y_1, \ldots, y_k \in \Omega_2$ where there is a known correspondence. Map both sets into $\mathbb{R}^{k-1}$ using the above ideas. Use an affine translation to match each x_j to y_j. This gives a correspondence between Ω_1 and Ω_2 by matching them up in the latent space. It is assuming a lot of linearity though.

10 Notes for Yang paper [8]

Setup: X^S, and X^T are our two sets. Create a distance matrix D_G via

$$D_G = \begin{bmatrix} D_{X^S,X^S} & D_{X^S,X^T} \\ D_{X^T,X^S} & D_{X^T,X^T} \end{bmatrix}.$$

Distance within one set is normal geodesic distance. For distance between sets we have to do some thinking. Suppose we have correspondence pairs $(x^S_{c^i}, x^T_{x^i})$. First we rescale D_{X^T,X^T} to match D_{X^S,X^S}. Choose a μ minimizing $\left\|D_{X^S_c,X^S_c} - \mu D_{X^T_c,X^T_c}\right\|$ and use this to rescale. Then define $dist(x^S_i, x^t_j) = \min(dist(x^S_i, x^S_{c^k}) + dist(x^T_{c^k}, x^T_j))$. In other words, glue the correspondence pairs together, then use geodesic distance still.

The paper also has some ideas of how to choose correspondence pairs. Geographical nearest neighbor and spectral nearest neighbor are both interesting. The paper suggest picking corresponding pairs by minimizing.

$$\left\|x^S_p - x^T_q\right\| + a\,(\text{physical distance})\,.$$

Here a is a balancing parameter.

References

[1] A. M. Ali and A. A. Farag. A novel framework for n-d multimodal image segmentation using graph cuts. In *2008 15th IEEE International Conference on Image Processing*, pages 729–732, Oct 2008. 4

[2] D. CONTE, P. FOGGIA, C. SANSONE, and M. VENTO. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004. 2

[3] Stéphane Lafon, Yosi Keller, and Ronald R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1784–1797, May 2006. 9

[4] Zhaoyi Meng, Ekaterina Merkurjev, Alice Koniges, and Andrea L Bertozzi. Hyperspectral image classification using graph clustering methods. Note: This is a preprint, 2017. 8

[5] Michael Möller, Todd Wittman, Andrea L Bertozzi, and Martin Burger. A variational approach for sharpening high dimensional images. *SIAM J. Imaging Sciences*, 5(1):150–178, 2012. 5

[6] Farnaz Sedighin, Massoud Babaie-Zadeh, Bertrand Rivet, and Christian Jutten. Two Multimodal Approaches for Single Microphone Source Separation. In *24th European Signal Processing Conference (EUSIPCO 2016)*, pages 110–114, Budapest, Hungary, September 2016. 3

[7] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 6

[8] Hsiuhan Lexie Yang and Melba M. Crawford. Learning a joint manifold with global-local preservation for multitemporal hyperspectral image classification. In *IEEE International Geoscience and Remote Sensing Symposium*, 2013. 10

[9] Yi-Ren Yes, Chun-Hao Huang, and Yu-Chiang Frank Wang. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Transactions on Image Processing*, 23(5):2009–2018, May 2014. 7