Generalization of CCA via Spectral Embedding

Jagadeesh Jagarlamudi
jags@umiacs.umd.edu
Department of Computer Science
University of Maryland
College Park, MD 20742

Raghavendra Udupa
raghavu@microsoft.com
Microsoft Research India
Bangalore, India 560080

Hal Daumé III
hal@umiacs.umd.edu
Department of Computer Science
University of Maryland
College Park, MD 20742

1 Canonical Correlation Analysis

Given a multi-view data, Canonical Correlation Analysis (CCA) [3] is a technique to find the projection directions in each view so that the observations when projected along these directions are maximally aligned. Let X $(d_1 \times n)$ and Y $(d_2 \times n)$be the representation of data in both the views, then CCA finds the projection directions $\mathbf{a}$ and $\mathbf{b}$ such that:

$$\arg\max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^T X Y^T \mathbf{b}}{\sqrt{\mathbf{a}^T X X^T \mathbf{a}}\sqrt{\mathbf{b}^T Y Y^T \mathbf{b}}}$$

This objective function can be re-written as:

$$\arg\min_{\mathbf{a},\mathbf{b}} ||X^T\mathbf{a} - Y^T\mathbf{b}||^2 \qquad \text{s.t. } \mathbf{a}^T X X^T \mathbf{a} = 1 \;\; \& \;\; \mathbf{b}^T Y Y^T \mathbf{b} = 1 \tag{1}$$

2 Spectral Embedding

Given a similarity matrix of size W $(n \times n)$, Spectral Embedding [4, 5] involves finding a vector $\mathbf{u}$ which minimizes the following objective function:

$$\arg\min_{\mathbf{u}} \frac{1}{2}\sum_{ij} W_{ij}(u_i - u_j)^2 \tag{2}$$

with an appropriate length constraint on u. The objective function in Eqn. 2 can be rewritten as [1]:

$$\arg\min_{\mathbf{u}} \mathbf{u}^T L \mathbf{u} \quad \text{s.t.} \;\; \mathbf{u}^T D \mathbf{u} = 1 \tag{3}$$

where $L = D - W$ is the unnormalized Laplacian matrix corresponding to W and D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$. The above optimization function reduces to solving the generalized eigenproblem $L\mathbf{u} = \lambda D\mathbf{u}$ [5].

3 CCA via Spectral Embedding

Locality Preserving Projections (LPP) [2] is a technique that uses spectral embedding to find a lower dimensional representation of observations such that local neighbourhood is preserved. Here we propose a generalization of this technique to the multi-view data. It turns out that our new formulation generalizes CCA in many ways.

Let Z be a $(d_1 + d_2) \times 2n$ matrix representing the multi-view data, $\mathbf{p}$ be a vector of length $d_1 + d_2$ and W be a matrix of size $2n \times 2n$ defined in the following way:

$$Z = \begin{bmatrix} X & \mathbf{0} \\ \mathbf{0} & Y \end{bmatrix}_{(d_1+d_2)\times 2n} \; ; \quad \mathbf{p} = \begin{bmatrix} \mathbf{a}_{(d_1\times 1)} \\ \mathbf{b}_{(d_2\times 1)} \end{bmatrix} ; \quad W = \begin{bmatrix} \mathbf{0} & I \\ I & \mathbf{0} \end{bmatrix}_{2n\times 2n} \Rightarrow D = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}_{2n\times 2n} \tag{4}$$

and let $u_i = \mathbf{p}^T Z_i$ for $i = 1 \cdots 2n$. Substituting these definitions in Eqn. 2:

$$\begin{aligned} \frac{1}{2}\sum_{i,j=1}^{2n} W_{ij}(u_i - u_j)^2 &= \frac{1}{2}\sum_{i,j=1}^{2n} W_{ij}(\mathbf{p}^T Z_i - \mathbf{p}^T Z_j)^2 \\ &= \frac{1}{2}\Big(\sum_{i=1}^{n}(\mathbf{p}^T Z_i - \mathbf{p}^T Z_{i+n})^2 + \sum_{i=1+1}^{2n}(\mathbf{p}^T Z_i - \mathbf{p}^T Z_{i-n})^2\Big) \\ &= \frac{1}{2}\Big(\sum_{i=1}^{n}(\mathbf{a}^T X_i - \mathbf{b}^T Y_i)^2 + \sum_{i=1}^{n}(\mathbf{b}^T Y_i - \mathbf{a}^T X_i)^2\Big) \\ &= \sum_{i=1}^{n}(\mathbf{a}^T X_i - \mathbf{b}^T Y_i)^2 = \|X^T\mathbf{a} - Y^T\mathbf{b}\|^2 \end{aligned} \tag{5}$$

which is same as the CCA objective function in Eqn. 1 and similarly the constraint $\mathbf{u}^T D\mathbf{u} = 1$ reduces to the constraints in Eqn. 1. Thus, with the above definitions of Z, $\mathbf{p}$ and W, both CCA and Spectral Embedding solves the same optimization problem. Hence, CCA solution can also be obtained by solving the generalized eigenvalue problem:

$$ZLZ^T \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \lambda \; ZDZ^T \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \tag{6}$$

where L and D are the unnormalized laplacian and the diagonal matrices corresponding to the weight matrix W (as defined in Sec. 2).

The above formulation offers several advantages that are not trivial with the original definition (Eqn. 1). First of which is the ability to incorporate varying types of local neighbour hood, *i.e.*, by appropriately defining the diagonal block matrices of weight matrix W we can take the intra-view similarities into account. Moreover, the anti-diagonal block matrices can be perturbed to consider the ambiguous/noisy alignments in the multi-view data. The second main advantage is its ability to consider common features across multiple views. The $\mathbf{p}$ vector can be trivially modified to consider the shared features across different views.

References

[1] David R. Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning Journal*, 2011.

[2] Xiaofei He and Partha Niyogi. Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[3] H. Hotelling. Relation between two sets of variables. *Biometrica*, 28:322–377, 1936.

[4] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *Advances in Neural Information Processing Systems*, 2002.

[5] Ulrike von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 2007.

Topic: learning algorithms and learning theory
Preference: oral