A spectral approximation to the Indian Buffet Process

J Lloyd, Z Ghahramani, C Reed
Department of Engineering
Cambridge University

Abstract

Literature review of spectral clustering for maximum-likelihood IBP formulation.

1 Paper Summaries

1.1 Fast Approximate Spectral Clustering: Yan *et al.* [9]

Yan *et al.* [9] discuss fast approximate spectral clustering framework based on minimizing the distortion (misclustering) of the resulting clusters as compared to non-approximate spectral clustering. Specifically, they introduce a K-means and RP-trees preprocessing step that forms a set of representative points and then performs spectral clustering on the representative points. These algorithms are referred to as *KASP* and *RASP*, respectively. The complexity of KASP is $O(knt) + O(k^3)$ while RASP is $O(hn) + O(k^3)$, where k is the number of representative points, n is the number of data points, t is the number of iterations of the preprocessing algorithms, and h is the depth of the RP tree.

Yan *et al.* [9] use a perturbation analysis of the graph Laplacian to show that the misclustering of (K—R)ASP converges to 0 as the number of representative points increases. Furthermore, they show that their techniques perform much better than K-means and comparable to a Nyström approximation but with smaller memory requirements and runtime.

1.2 Non-redundant Multi-view Clustering via Orthogonalization: Cui *et al.* [1]

Cui *et al.* [1] propose general methods for forming multiple clustering solutions for a given dataset, where each clustering is in a different subspace. Their method operates iteratively by clustering the data using k-means, orthogonalizing the data to a new subspace that is not covered by the previous clusters, clustering the data in the new subspace, et cetera until reaching some form of convergence. The authors discuss two approaches to performing this clustering: orthogonal clustering and clustering in orthogonal spaces.

Orthogonal clustering: given current data $X^{(t)}$ and the clustering solution of $X^{(t)}$, $M^{(t)} = [\mu_1^{(t)}\mu_2^{(2)}\cdots\mu_k^{(t)}]$, Cui *et al.* [1] view orthogonal clustering as projecting the data points from the original data space to a compressed space (subspace) spanned by the mean vectors. They describe two variations for performing clustering in the space that is orthogonal to the compressed data space based on the hard and soft clustering views:

- hard clustering: each data point $x_i^{(t)}$ belonging to cluster j is projected onto the cluster center $\mu_j^{(t)}$ and $x_i^{(t+1)}$ is determined as $x_i^{(t)}$ projected onto the subspace orthogonal to the cluster centroids:
$$x_i^{(t+1)} = (I - \mu_j^{(t)}\mu_j^{(t)T}/\mu_j^{(t)T}\mu_j^{(t)})x_i^{(t)}$$
- soft clustering: since each data point can partially belong to each cluster, we can project $X^{(t)}$ onto all cluster means and compute $X^{(t+1)}$ as $X^{(t)}$ projected onto the subspace or-

thogonal to all the cluster centroids:

$$X^{(t+1)} = (I - M^{(t)}(M^{(t)T}M^{(t)})^{-1}M^{(t)T}X^{(t)}$$

Clustering in orthogonal subspaces: Cui *et al.* [1] view clustering in orthogonal spaces as projecting the data into a reduced dimensional space that discriminates the given classes: i.e. via latent discriminant analysis or PCA. In particular, given current data $X^{(t)}$ and the clustering solution of $X^{(t)}$, $M^{(t)} = [\mu_1^{(t)}\mu_2^{(2)}\cdots\mu_k^{(t)}]$, the authors find the PCA solution of $M^{(t)}$, keep the $k^{(t)} - 1$ eigenvectors to obtain the subspace $A^{(t)}$ that captures the current clustering, and then projects $X^{(t)}$ to the space orthogonal to $A^{(t)}$:

$$X^{(t+1)} = (I - A^{(t)}(A^{(t)T}A^{(t)})^{-1}A^{(t)T}X^{(t)}$$

Key observations: Cui *et al.* [1] used k-means for clustering and is therefore limited to convex clusters.v

1.3 Multiple Non-Redundant Spectral Clustering Views: Niu *et al.* [6]

Notation summary

- G: $G = \{V, E\}$ is the graph used to define the $NCut$ formulation of spectral clustering
- K: similarity matrix with elements k_{ij} measure the similarity between vertices i and j in G
- P_i: partition/cluster i
- d_i: degree of vertex i with $d_i = \sum_{j=1}^{n} k_{ij}$
- U: indicator matrix: assigns vertex (row) to cluster (c) — takes real values in relaxed spectral clustering formalism
- n: number of elements
- c: number of clusters
- W_q: $\mathbb{R}^{d\times l_q}$ transformation matrix for each view that transforms the data in the original space to the lower-dimensional space
- m: number of subspaces (number of multiple clustering solutions)
- q: indexes the subspaces
- x: the input data

The central idea of this work was to determine multiple non-redundant spectral clustering views by learning non-redundant subspaces and performing spectral clustering in these subspaces. Specifically, this was accomplished by using the NCut spectral clustering formulation over a similarity graph $G = \{V, E\}$ with disjoint partitions $P_1, \ldots, P_c$, and similarity matrix K:

$$NCut(P_1, \ldots, P_c) = \sum_{t=1}^{c} \frac{cut(P_t, VP_t}{vol(P_t)},$$

where

$$cut(\mathcal{A}, \mathcal{B}) = \sum_{v_i\in\mathcal{A}, v_j\in\mathcal{B}} k_{ij},$$

and degree $d_i = \sum_{j=1}^{n} k_{ij}$ so that $vol(\mathcal{A}) = \sum_{i\in\mathcal{A}} d_i$. By introducing the indicator matrix U and allowing this matrix to take on values in $[0, 1]$ (a relaxation of the original discrete optimization problem), cluster assignment optimization reduces to the well-known trace-maximization problem:

$$\max_{U\in\mathbb{R}^{n\times c}} \text{tr}(U^T D^{-1/2} K D^{-1/2} U)$$
$$\text{s.t. } U^T U = 1$$

Niu *et al.* [6] use the kernel similarity matrix K_q computed with the kernel function $k(W_q^T x_i, W_q^T x_j)$ where $W_q \in \mathbb{R}^{d\times l_q}$ transforms the original data to subspace q, where $q \in$

$\{1, \ldots, m\}$. The trick used by the authors is to introduce a penalty (regularization term) for redundant kernel spaces via the Hilbert-Schmidt Independence Criterion, so that our optimization problem becomes:

$$\max_{U_1,\ldots,U_m,W_1,\ldots,W_m} \text{tr}(U_q^T D_q^{-1/2} K_q D_q^{-1/2} U_q) - \lambda \sum_{q \neq r} \text{HSIC}(W_q^T x, W_r^T x)$$

$$\text{s.t. } U_q^T U_q = 1$$

$$(K_q)_{ij} = k_q(W_q^T x_i, W_q^T x_j)$$

$$W_q^T W_q = I$$

which is optimized in two steps:

1. Assume all W_q fixed, optimize U_q in each view (perform spectral clustering in the given subspace—simple)
2. Assume all U_q fixed, optimize W_q in each view —- used gradient ascent on the Stiefel manifold[1]

Finally, initialization is important in this algorithm. The authors initialize the W_q matrices by performing spectral clustering with a similarity matrix A that measures similarity between the original features of the dataset using HSIC and then clusters the features into m clusters (m is a parameter). Then for each feature assigned to cluster q, append a $d \times 1$ column to W_q with all entries 0 except the j^{th} element which is set to 1. This effectively creates m subspaces where each subspace clusters on a disjoint subset of the features. However, the final learned W_q can have features with weights in many views.

2 Literature to Investigate

- Griffiths and Ghahramani [2]
- Luxburg [5]
- Cui *et al.* [1]
- Niu *et al.* [6]
- Niu *et al.* [7]
- Niu *et al.* [8]
- Higham and Kibble [3]
- Yan *et al.* [9]
- Jain *et al.* [4]

References

[1] Cui, Y., Fern, X. Z., and Dy, J. G. (2007). Non-redundant Multi-view Clustering via Orthogonalization. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, **3**, 133–142.

[2] Griffiths, T. L. and Ghahramani, Z. (2011). The Indian Buffet Process : An Introduction and Review. *Journal of Machine Learning Research (JMLR)*, **12**, 1185–1224.

[3] Higham, D. and Kibble, M. (2004). A unified view of spectral clustering. *University of Strathclyde mathematics research . . .*, (53441), 1–17.

[4] Jain, P., Meka, R., and Dhillon, I. (2008). Simultaneous unsupervised learning of disparate clusterings. *Statistical Analysis and Data Mining*, **1**(3), 195–210.

[5] Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17**(4), 395–416.

[6] Niu, D., Dy, J. G., and Jordan, M. I. (2010). Multiple Non-Redundant Spectral Clustering Views. In *Proceedings of the International Conference on Machine Learning (ICML-*.

[7] Niu, D., Dy, J. G., and Jordan, M. I. (2011). Dimensionality Reduction for Spectral Clustering. In *Proceeding of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15.

[1] Colorado does not understand this optimization technique

[8] Niu, D., Dy, J. G., and Ghahramani, Z. (2012). A Nonparametric Bayesian Model for Multiple Clustering with Overlapping Feature Views. In *Proceedings of the International Conference on Artifical Intelligence and Statistics (AISTATS)*.

[9] Yan, D., Huang, L., and Jordan, M. (2009). Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916. ACM.