On Spectral Clustering: Accelerating Computation and Improving Accuracy of Clusters

(to be updated)

Start: April 22, 2022

Abstract

Key words:

1 Introduction

2 The main framework for spectral clustering

The main structure of a spectral clustering algorithm includes three parts: 1. Form the affinity matrix and a Laplacian matrix; 2. Compute the outer eigenvalues and their associated eigenvectors of the Laplacian; 3. Apply k-means to the normalized rows of the eigenvectors for clustering.

Assuming the goal is to cluster n data points $S = \{s_1, \cdots, s_n\} \subset \mathbb{R}^L, n \geqslant L$ into k clusters. A more detailed spectral clustering algorithm is presented in Algorithm 2.1, it summarizes the three spectral clustering algorithms surveyed in [14]. A slight difference from [14] is that the normalization of the rows of eigenvectors as proposed in [19] are applied to all three Laplacians.

Algorithm 2.1: Spectral Clustering Algorithm

1. Form the affinity matrix $A \in \mathbb{R}^{n\times n}$, one common way is to set $A_{ij} = e^{-\|s_i - s_j\|^2/\mu^2}$ for a chosen $\mu \neq 0$, and $A_{ii} = 0$
2. Compute a Laplacian matrix from A, such as the random walk Laplacian $I - D^{-1}A$, or the symmetric Laplacian $I - D^{-1/2}AD^{-1/2}$, or simply $D - A$. Here D is the diagonal degree matrix, its i-th diagonal is the sum of degrees at vertex i: $d_i = \sum_{j=1}^{n} A_{ij}$
3. Compute the **largest** k eigenvalues and their associated eigenvectors of the chosen Laplacian
4. Normalize the n rows of the eigenvector matrix to unit length, treat the n rows as n points on the unit hyper-sphere in $\mathbb{R}^k$, apply k-means to cluster these points into k clusters
5. Assign the same grouping of the rows of the eigenvectors to the original data, i.e., s_i is assigned to cluster j if and only if row-i of the eigenvector matrix is assigned so.

In [19], the scaled affinity matrix $D^{-1/2}AD^{-1/2}$ is used, thus the smallest k eigenvalues and their associated eigenvectors are computed for clustering.

There exist rather extensive literature that tries to extend the power of spectral clustering to larger data. They can be grouped into two categories, 1. Building the affinity matrix utilizing some kNN techniques; 2. Computing the eigenvectors only approximately, either by filtering or utilizing some localized techniques.

We focus on the second category. Our method can utilize any advance that have been made for the first category.

3 Existing acceleration schemes

For large dataset, the adjacency matrix as well as the Laplacians are of high dimension. Computing the eigenvectors using standard eigen-algorithm would suffer the $O(n^3)$ complexity, methods of lower than cubic-order complexity are needed.

Fowlkes *et al.* proposed the Nyström method to accelerate the computation of the eigenvectors [6].

Tremblay *et al.* proposed applying polynomials that approximate a low-pass step function to filter out unwanted spectrum of the adjacency matrix [24]. Note that the kept lower part of the spectrum of the adjacency matrix would translate to the higher part of the spectrum of a Laplacian.

[28] (haven't read yet)

4 Propose acceleration schemes

The subspace filter method proposed in [33, 32] has led to order of magnitude speedup in first principle density function theory calculations [23, 3, 31, 17, 18, 7].

The acceleration technique in [33, 32] seems naturally fit to speedup the computation of eigenvector for spectral clustering.

(more details to be added)

5 (Things to do or to figure out)

- Which part of the spectrum to keep?

  In [14] is was reasoned that it is the dominant part of the spectrum of the Laplacian matrix that are important for clustering. That is, keep the larger eigenvalues and their associated eigenvectors of the Laplacian matrix .

  While in [29] a different view was held, namely that larger magnitude of the eigenvalues of a Laplacian do not necessarily imply more importance for clustering.

  In [2] it was argued that it is the eigenvectors that are more important for clustering.

  Construct some realistic example to figure out which of the above is closer to the truth: I.e., which part of the spectrum (eigenvalues and/or eigenvectors) is more essential for finding the 'right' clusters in a given dataset?

- Study the models or Obtain the datasets that have been used as benchmarks for clustering and grouping:

  The stochastic block model (SBM) [8, 27, 1] that have been extensively used in spectral clustering studies [21, 22, 24, 4].

  The LFR benchmarks [11]. C++ code https://github.com/eXascaleInfolab/LFR-Benchmark_UndirWeightOvp Also in the networkx package (python) [1].

[1] https://networkx.org/documentation/stable/reference/generated/networkx.generators.community.LFR_benchmark_graph.html

More realistic datasets: smaller IsoRank PPI Network Alignment Based Ortholog Database [2], larger Amazon, DBLP, LiveJ, YouTube, Orkut datasets from SNAP [3] [12]; publicly available ACM dataset, and preprocessed DBLP and IMDB datasets [26, 20]; BrainNet, 20news, DBLP, Flickr [13].

SNAP datasets are said to have many small communities. While the social network Flickr has big communities.

- Learn and understand the different functions used to measure the quality of a clustering result, such as the F1 score –defined as the harmonic mean of precision and recall, the normalized mutual information (NMI) [10, 5], the variation of information (VI) [16][4], the within-cluster sum of squares (WCSS), the adjusted rand index (ARI) [9], ...

- For k-means clustering, the quality can be improved by seeding it using the clustering result obtained from the column pivoted QR [4], as pioneered in [30]. (Is there a way to avoid k-means altogether? Especially for the cases where the conditions in [30] are not met so that column pivoted QR is not expected to perform well.)

- Most paper use t-SNE [25] to visualize the clustering. Study UMAP [15], which is developed for dimension reduction and visualization of much larger data. Replace t-SNE by UMAP.

- Spectral clustering has the known issue of crowding, namely that some part of the clusters are crowded together (not well-separated). What is the root course of this issue? What can be done to alleviate such issue if clusters indeed share no common features?

6 Numerical results

7 Concluding Remarks

References

[1] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. IEEE Transactions on Information Theory, 62(1):471–487, 2016.

[2] Matthew Brand and Kun Huang. A unifying theorem for spectral embedding and clustering. In Christopher M. Bishop and Brendan J. Frey, editors, Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, volume R4 of Proceedings of Machine Learning Research, pages 41–48. PMLR, 03–06 Jan 2003.

[3] J. R. Chelikowsky, Y. Saad, T.-L. Chan, M. L. Tiago, A. T. Zayak, and Y. Zhou. Pseudopotentials on grids: Application to the electronic, optical, and vibrational properties of silicon nanocrystals. J. Comput. Theoret. Nanoscience, 6(6):1247–1261, 2009.

[4] Anil Damle, Victor Minden, and Lexing Ying. Simple, direct and efficient multi-way spectral clustering. Information and Inference: A Journal of the IMA, 8(1):181–203, 2018.

[2] http://cb.csail.mit.edu/cb/mna/isobase/

[3] http://snap.stanford.edu/

[4] https://handwiki.org/wiki/Variation_of_information

[5] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005(09):P09008, 2005.

[6] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(2):214–225, 2004.

[7] Swarnava Ghosh and Phanish Suryanarayana. SPARC: Accurate and efficient finite-difference formulation and parallel implementation of Density Functional Theory: Isolated clusters. Comp. Phys. Comm., 212:189–204, 2017.

[8] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social Networks, 5(2):109–137, 1983.

[9] L. Hubert and P. Arabie. Comparing partitions. J. Classification, 2:193–218, 1985.

[10] T.O. Kvalseth. Entropy and correlation: Some comments. IEEE Transactions on Systems, Man, and Cybernetics, 17 (3):517–519, 1987.

[11] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. Physical Review E, 78(4), 2008.

[12] Yixuan Li, Kun He, David Bindel, and John E. Hopcroft. Uncovering the small community structure in large networks: A local spectral approach. In Proceedings of the 24th International Conference on World Wide Web, WWW '15, pages 658–668. ACM, 2015.

[13] D. Luo, J. Ni, S. Wang, Y. Bian, X. Yu, and X. Zhang. Deep Multi-Graph Clustering via Attentive Cross-Graph Association, pages 393–401. Association for Computing Machinery, 2020.

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

[15] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. 2020.

[16] Marina Meilă. Comparing Clusterings by the Variation of Information, pages 173–187. Springer, 2003.

[17] P. Motamarri and V. Gavini. Higher-order adaptive finite-element methods for Kohn-Sham density functional theory. J. Comput. Phys., 253:308–343, 2013.

[18] P. Motamarri and V. Gavini. A subquadratic-scaling subspace projection method for large-scale Kohn-Sham density functional theory calculations using spectral finite-element discretization. Phys. Rev. B, 90:115127, 2014.

[19] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. Proc. NIPS, 2001.

[20] Chanyoung Park, Donghyun Kim, Jiawei Han, and Hwanjo Yu. Unsupervised attributed multiplex network embedding. arXiv, abs/1911.06750, 2019.

[21] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. The Annals of Statistics, 39(4):1878–1915, 2011.

[22] Geoffrey Schiebinger, Martin J. Wainwright, and Bin Yu. The geometry of kernelized spectral clustering. The Annals of Statistics, 43(2):819–846, 2015.

[23] M. L. Tiago, Y. Zhou, M. Alemany, Y. Saad, and J. R. Chelikowsky. Evolution of magnetism in iron from the atom to the bulk. Phys. Rev. Lett., 97:147201, 2006.

[24] Nicolas Tremblay, Gilles Puy, Remi Gribonval, and Pierre Vandergheynst. Compressive spectral clustering, arXiv:1602.02018, 2016.

[25] L. van Der Maaten and G. Hinton. Visualizing data using t-SNE. Journal of Machine Learning, 9:2431–2456, 2008.

[26] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, Philip S. Yu, and Yanfang Ye. Heterogeneous graph attention network. arXiv, abs/1903.07293, 2019.

[27] Yuchung J. Wang and George Y. Wong. Stochastic blockmodels for directed graphs. Journal of the American Statistical Association, 82(397):8–19, 1987.

[28] Lingfei Wu, Pin-Yu Chen, Ian En-Hsu Yen, Fangli Xu, Yinglong Xia, and Charu Aggarwal. Scalable spectral clustering using random binning features. In Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2018.

[29] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems 17 (NIPS 2004), pages 1601–1608, 2005.

[30] H. Zha, C. Ding, M. Gu, X. He, and H.D. Simon. Spectral relaxation for K-means clustering. In Neural Information Processing Systems, volume 14 of NIPS '01, pages 1057–1064, 2001.

[31] Y. Zhou, J. R. Chelikowsky, and Y. Saad. Chebyshev-filtered subspace iteration method free of sparse diagonalization for solving the Kohn-Sham equation. J. Comput. Phys., 274:770–782, 2014.

[32] Y. Zhou, Y. Saad, M. L. Tiago, and J. R. Chelikowsky. Parallel self-consistent-field calculations using Chebyshev-filtered subspace acceleration. Phys. Rev. E, 74(6):066704, 2006.

[33] Y. Zhou, Y. Saad, M. L. Tiago, and J. R. Chelikowsky. Self-consistent-field calculation using Chebyshev-filtered subspace iteration. J. Comput. Phys., 219(1):172–184, 2006.