Spectral clustering for extremely large-scale datasets (U-SPEC and U-SENC)

Introduction

Data clustering is a fundamental problem in the field of data mining and machine learning. Historically there have been a number of clustering algorithms that were developed and research is still happening in this space.

Spectral clustering in recent years has been gaining increasing attention due to its promising ability in dealing with nonlinearly separable datasets. However, a critical limitation to conventional spectral clustering lies in its huge time and space complexity, which significantly restricts its application to large-scale problems.

Spectral clustering consists of two phases, affinity matrix construction and eigen-decomposition. It takes O(N2d) time and O(N2) memory to construct the affinity matrix, and takes O(N3) time and O(N2) memory to solve the eigen-decomposition problem where N is the data size and d is the dimension. As the data size N increases, the computational burden of spectral clustering grows dramatically. For example, given a dataset with one million objects, the N×N affinity matrix alone will consume 7450.58 GB of memory.

In this article we will discuss two novel large-scale algorithms, namely, ultra-scalable spectral clustering (U-SPEC) and ultra-scalable ensemble clustering (U-SENC).

Spectral Clustering

Given a dataset of N objects, conventional spectral clustering first computes an N×N affinity matrix, in which each entry corresponds to the similarity of two objects according to some similarity metrics. Then, the eigen-decomposition is performed on the graph Laplacian of the affinity matrix to obtain the k eigenvectors associated with the first k eigenvalues. By embedding the datasets into the low-dimensional space via the obtained k eigenvectors, the final clustering can be achieved via k-means or some other discretization techniques.

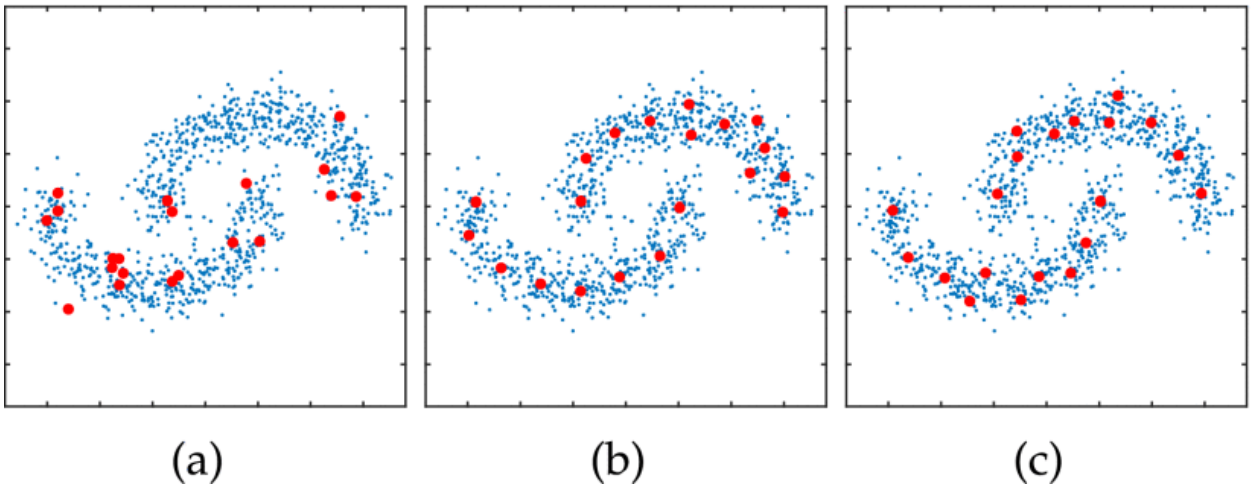
Although spectral clustering has shown promising advantages in finding clusters of arbitrary shapes from complex data, its O(N3) time complexity and O(N2) space complexity significantly restrict its application in large-scale tasks. To alleviate the huge computational cost, some researchers sparsified the affinity matrix by considering K-nearest neighbors or ϵ-neighbors, and then solved the eigen-decomposition problem by some sparse eigen-solvers , which, however, still requires the computation of all the entries in the original affinity matrix. The sub-matrix based approximation has emerged as a powerful and efficient tool for spectral clustering to avoid computation of full affinity matrix.The sub-matrix construction takes O(Npd) time and O(Np) memory, which is much lower than the full affinity matrix construction.The O(Np) memory cost of the sub-matrix construction can still be a critical bottleneck when dealing with very large datasets

Ultra-Scalable Spectral Clustering (U-SPEC) Approach

The proposed U-SPEC algorithm complies with the sub-matrix based formulation and aims to break through the efficiency bottleneck of previous algorithms via three phases. In the first phase, we present a hybrid representative selection strategy to strike a balance between the efficiency of the random selection and the effectiveness of the k-means based selection. In the second phase, we develop a coarse-to-fine method to efficiently approximate the K-nearest representatives for each data object, and construct a sparse affinity sub-matrix between the N objects and the p representatives. In the third phase, the N×p sub-matrix is interpreted as a bipartite graph, which can be efficiently partitioned to obtain the final clustering result.

Hybrid Representative Selection

A hybrid representative selection strategy, which is designed to find a balance between the efficiency of random selection and the effectiveness of k-means based selection.

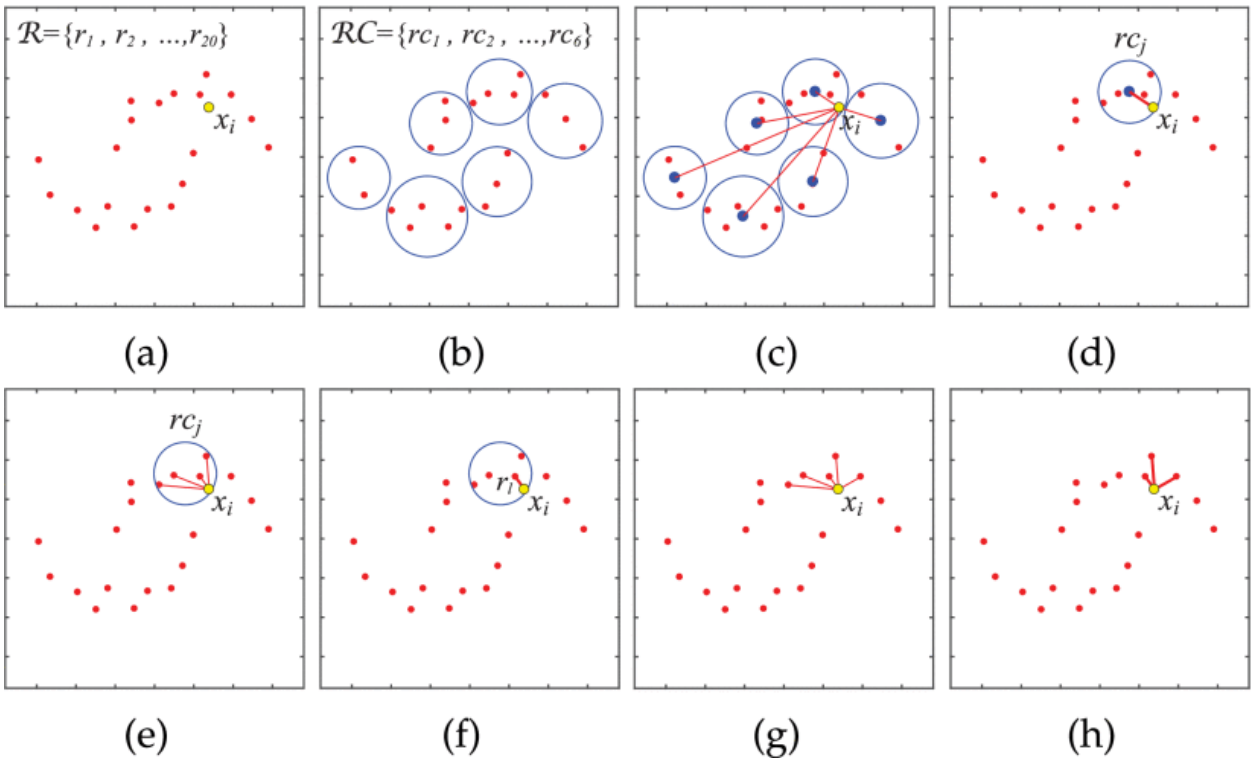


Comparison of the representatives produced by (a) random selection, (b) k -means based selection, and (c) hybrid selection.

Hybrid selection strategy first random samples a set of p′ candidate representatives such that p<p′≪N. Then, upon the p′ candidates, we perform the k-means method to obtain p clusters and exploit the p cluster centers as the set of representatives. By introducing an intermediate stage of random pre-sampling, the computational complexity of the k-means based selection is reduced from O(Npdt) to O(p2dt).

Approximation of K-Nearest Representatives

The next objective is to encode the pair-wise relationship of the entire dataset via the small set of representatives.a new K-nearest representative approximation method based on the coarse-to-fine mechanism, and build the sparse affinity sub-matrix with O(Np12d) complexity. The main idea of our K-nearest representative approximation is to first find the nearest region, then find the nearest representative (denoted as rl) in the nearest region, and finally find the K-nearest representatives in the neighborhood of rl.



Approximate K-nearest representatives. (a) The representative set R and an object xi∈X. (b) Partition the representatives into several rep-clusters. (c) Compute the distances between xi and all the rep-cluster centers. (d) Find the nearest rep-cluster rcj. (e) Compute the distances between xi and all the representatives in rcj. (f) Find the nearest rl∈rcj. (g) Compute the distances between xi and the representatives in the K'-nearest neighborhood of rl (K'>K). (h) Obtain the approximate K-nearest representatives (K=3).

Bipartite Graph Partitioning

The affinity sub-matrix B reflects the relationship between the objects in X and the representatives in R, which can be naturally interpreted as a bipartite graph G={X,R,B}, where X∪R is the node set and B is the cross-affinity matrix (as shown in Fig. 4). By taking advantage of the bipartite graph structure, the transfer cut [6] can thereby be used to efficiently partition the graph and achieve the final clustering result.

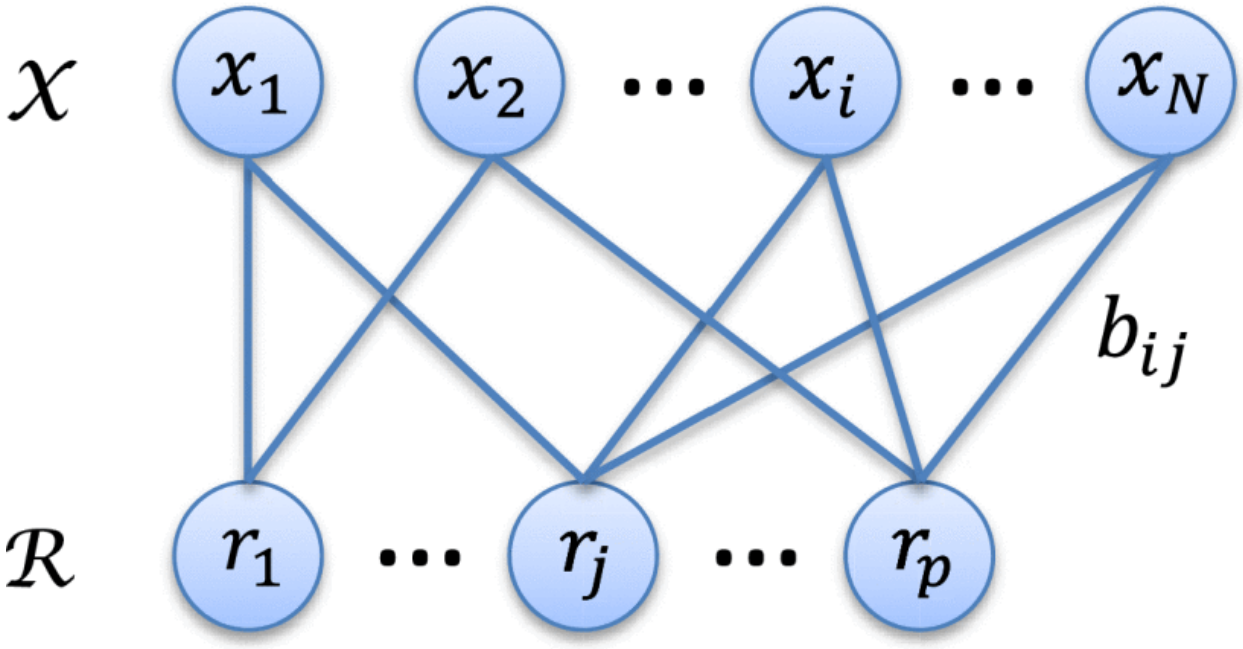


Illustration of the bipartite graph G.

Computational Complexity

The hybrid representative selection takes O(p2dt) time. The affinity construction takes O(N(p12d+Kd+K2)) time. The eigen-decomposition takes O(NK(K+k)+p3) time. The k-means discretization takes O(Nk2t) time. With consideration to k,K≪p≪N, the overall time complexity of U-SPEC is O(N(p12d+K2+Kk+Kd+k2t)), where O(Np12d) is the dominant term. Below table provides a comparison of time complexity of our U-SPEC algorithm against several other large-scale spectral clustering algorithms.

Method	Representative selection	Affinity construction	Eigen-decomposition
Nyström [3]	/	$O(Npd)$	$O(Np+p^3)$
LSC-R [4]	/	$O(Npd)$	$O(Np^2+p^3)$
LSC-K [4]	$O(Npdt)$	$O(Npd)$	$O(Np^2+p^3)$
U-SPEC	$O(p^2dt)$	$O(Np^{\frac{1}{2}}d)$	$O(NK(K+k)+p^3)$

Comparison of the Time Complexity of Several Large-Scale Spectral Clustering Methods

Ensemble Clustering

Ensemble clustering, also called consensus clustering, has been attracting much attention in recent years, aiming to combine multiple base clustering algorithms into a better and more consensus clustering.The existing ensemble clustering algorithms can be mainly classified into three categories.

- The first category is the pair-wise co-occurrence based methods. Fred and Jain proposed the evidence accumulation clustering (EAC) method, which makes use of the co-association matrix by considering the frequency of pair-wise co-occurrence among multiple base clusterings. With the co-association matrix treated as the similarity matrix, the agglomerative clustering algorithms were then performed to obtain the consensus clustering.
- The second category is the graph partitioning based methods. Strehl and Ghosh transformed the multiple base clusterings into a hypergraph representation, based on which three graph partitioning based ensemble clustering methods were presented.
- The third category is the median partition based methods , which cast the ensemble clustering problem into an optimization problem that aims to find a median clustering (or partition) by maximizing the similarity between this clustering and the multiple base clusterings.

These ensemble clustering algorithms have shown their advantages in improving clustering accuracy and robustness. However, due to the efficiency bottleneck, most of them are not suitable for very large-scale applications.

While there are two phases in ensemble clustering (i.e., ensemble generation and consensus function), these algorithms generally focus on the efficiency of the consensus function. In ensemble generation, they mostly exploited k-means to produce m base clusterings. Note that the time complexity of ensemble generation by k-means is O(Nmdkt), which can still be computationally expensive when dealing with very large-scale datasets. Moreover, the performance of k-means may significantly deteriorate when handling nonlinearly separable datasets, which has a critical influence on the robustness of the ensemble clustering algorithms.

Ultra-Scalable Ensemble Clustering (U-SENC) Approach

Unlike the common practice that typically exploits multiple k-means clusterers as base clusterers, the proposed U-SENC algorithm integrates a diverse set of large-scale U-SPEC clusterers into a highly efficient ensemble clustering framework, simultaneously tackles the scalability and nonlinear separability issues in both the ensemble generation and consensus function phases in ensemble clustering. The multiple base clusterings are incorporated into a new bipartite graph, which treats both objects and base clusters as graph nodes and is then efficiently partitioned to achieve the final consensus clustering

Conclusion

This article focuses on two large-scale clustering algorithms, termed ultra-scalable spectral clustering and ultra-scalable ensemble clustering, respectively. In U-SPEC, a new hybrid representative selection strategy is designed to strike a balance between the efficiency of random selection and the effectiveness of k-means based selection. Then a new approximation method for K-nearest representatives is presented to efficiently construct a bipartite graph between the original data objects and the set of representatives, upon which the transfer cut can be utilized to obtain the clustering result. Starting from the U-SPEC algorithm, we further integrate multiple U-SPEC clusterers into a unified ensemble clustering framework and propose the U-SENC algorithm. Specifically, multiple U-SPEC's are exploited in the ensemble generation phase to produce an ensemble of diverse and high-quality base clusterings. The multiple base clusterings are incorporated into a new bipartite graph, which treats both objects and base clusters as graph nodes and is then efficiently partitioned to achieve the final consensus clustering.

References :

D. Huang, C. -D. Wang, J. -S. Wu, J. -H. Lai and C. -K. Kwoh, "Ultra-Scalable Spectral Clustering and Ensemble Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1212-1226, 1 June 2020, doi: 10.1109/TKDE.2019.2903410.

U. von Luxburg, "A tutorial on spectral clustering", *Statist. Comput.*, vol. 17, no. 4, pp. 395-416, 2007

D. Huang, C.-D. Wang and J.-H. Lai, "Locally weighted ensemble clustering", *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460-1473, May 2018.

A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions", *J. Mach. Learn. Res.*, vol. 3, pp. 583-617, 2003.