Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Distributed Spectral Clustering on the Coordinator Model

Fabricio A. Mendoza Granada[1]
Núcleo de Investigación y Desarrollo Tecnológico, Universidad Nacional de Asunción, Paraguay
Marcos Villagra[2]
Núcleo de Investigación y Desarrollo Tecnológico, Universidad Nacional de Asunción, Paraguay

Clustering is a popular subject in non-supervised learning. Spectral clustering is a method for clustering that reduces dimensionality of data and guarantees a faster convergence to almost optimal clusters. Given a set of n points in $\mathbb{R}^d$, let $G = (V, E)$ be an undirected graph where each vertex represents a point in $\mathbb{R}^d$ and there is a nonnegative weighted edge between each pair of vertices (u, v) representing the similarity between u and v. The spectral clustering algorithm takes G as input and finds an almost optimal partition of the vertices of G by manipulating the spectra of the graph Laplacian of G.

A *Laplacian matrix* of G is defined as $L_G = D_G - A_G$, where D_G is the weighted degree matrix of G and A_G is its adjacency matrix. If k is the optimal number of clusters, spectral clustering works by first finding the first k eigenvalues and eigenvectors in ascending order of L_G. Then it runs any conventional clustering algorithm, like k-means, on the rows of a matrix constructed with the first k eigenvectors as column vectors.

In real world situations data is not always centralized, but spread among several sites. Achieving clustering in a decentralized setting is thus an interesting research subject.

A distributed protocol for spectral clustering was first proposed by Chen et al. [3]. In the work of [3], each site knows a graph $G_i = (V, E_i)$, where $\{E_i\}$ is a partition of the edge set and V is the entire set of vertices of G. Their protocol works as follows. Every player builds a $(1+\epsilon)$-*spectral sparsification* of its graph G_i. A graph H is a $(1+\epsilon)$-spectral sparsifier of G if $(1-\epsilon)x^T L_G x \leq x^T L_H x \leq (1+\epsilon)x^T L_G x$, where $x \in \mathbb{R}^n$, x^T is the transpose of x, and L_H is the graph Laplacian of a graph H which is constructed by random sampling over the edges with respect to some carefully selected probability distribution over the edges of G. Then, after each player construct its spectral sparsifier $H_i = (V, \tilde{E}_i)$, it sends H_i to the coordinator. As a final step, the coordinator computes $\tilde{G} = (V, \cup_{i=1}^{s} \tilde{E}_i)$ and applies the spectral clustering algorithm on $\tilde{G}$.

Chen et al. [3] showed that the total amount of communication between the players and the coordinator is $\tilde{O}(ns \log^c ns)$, where $c \geq 1$ is a real constant. They also showed a lower bound of $\Omega(ns)$ for the total amount of communication in the coordinator model for any randomized protocol.

The work of Chen et al. [3] studied the case where each player knows the vertices of the data graph G, but only a subset of the edges. In this work, we will study the

[1]fabromendoza95@gmail.com
[2]mvillagra@pol.una.py

communication complexity in the more extreme case where the vertex set is completely partitioned. Let $G = (V, E)$ be graph of the data points. In the coordinator model we have s sites where site i, with $1 \leq i \leq s$, knows a graph $G_i = (V_i, E_i)$, where $\{V_i\}$ and $\{E_i\}$ are partitions of the vertex set and the edge set of G, respectively. Each site can communicate with the coordinator with messages but no site can communicate with another site. Then after a finite number of rounds of communication, the coordinator computes an optimal partition of V. The goal is to find a protocol where an optimal partition of V can be computed using the minimum amount of communication. Formally there are s players and one coordinator and the coordinator wants to compute some function $f : X_1 \times ... \times X_s \to Z$ where X_i is the set of available inputs for player i. A *protocol* Π is defined as a sequence of binary strings sent by every player to the coordinator and back.

Furthermore, as a midpoint between Chen et al.'s work, we will study the case where there is an overlapping between the data among sites. These works are the first steps towards a communication-efficient full distributed protocol for graph clustering.

Acknowledgement. This work is supported by Conacyt research grant PINV15-208.

References

[1] J. Batson, D. A. Spielman, N. Srivastava, and S. H. Teng, Spectral sparsification of graphs: theory and algorithms. *Communications of the ACM*, 2013. 56(8), 87-94.

[2] M. Braverman, F. Ellen, R. Oshman, T. Pitassi and V. Vaikuntanathan, A tight bound for set disjointness in the message-passing model. *Foundations of Computer Science (FOCS)*, 2013 IEEE 54th Annual Symposium on. IEEE, 2013.

[3] J. Chen, H. Sun, D. Woodruff and Q. Zhang, Communication-optimal distributed clustering.*In Advances in Neural Information Processing Systems*, 2016. (pp. 3727-3735).

[4] Z. Huang, B. Radunovic, M. Vojnovic and Q. Zhang, Communication complexity of approximate matching in distributed graphs. *In LIPIcs-Leibniz International Proceedings in Informatics*, volume 30, 2015. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[5] J. R. Lee, S. O. Gharan and L. Trevisan, Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 2014. 61(6), 37.