Number-On-Forehead Communication Complexity of Data Clustering with Sunflowers

Fabricio Mendoza-Granada, Marcos Villagra

[1]Universidad Nacional de Asunción
NIDTEC, Campus Universitario, San Lorenzo C.P. 2169, Paraguay

Abstract. *We study the problem of performing data clustering in a distributed setting, which is a problem that may arise in many practical areas such as machine learning and data analysis. The way in which the sites communicate and the way data is allocated define a model of communication. We develop a protocol to compute distributed clustering in the Number-on-Forehead model of communication complexity. In our model, we requiere that each site is aware of all clusters in its own data and all data allocated among sites define a sunflower. We show that there exists a two round communication protocol for data clustering where each site knows an approximation to all clusters. The cost of our protocol is at most $O\left(\log\left(\frac{n}{\epsilon^2}\sqrt{1-\lambda}\right)\right)$ bits of communication, where n is the number data points, ϵ is an approximation factor and λ is a ratio of common data points among sites.*

1. Introduction

In several situations, algorithms need to work with data that is not centralized and allocated in different sites. One way to deal with this situation is to design communication protocols so that the sites can communicate among them. In our days where data analysis is becoming more relevant in industry and academia, *clustering* is one of the main tools for understanding data.

In clustering, the data set is often represented as points in $\mathbb{R}^d$. One way to identify clusters in data points is to represent them as a weighted graph $G = (V, E, w)$ with a cost function w. The goal is to find a partition of the vertex set of G, which can be seen as a *multicut problem* [von Luxburg 2007]. Clustering has been studied previously in distributed models like the *coordinator model* and *blackboard model* [Chen et al. 2016].

Let $E_1, E_2, \ldots, E_s$ be a collection of data and let $P_1, P_2, \ldots, P_s$ be a collection of sites. Each site P_i has data E_i assigned to it. In this work, we study a communication model known as *Number-on-Forehead* (or NOF), where a site P_i knows all data except its own E_i. This is a well studied model in communication complexity because of its relevance in proving lower bounds in circuit complexity [Håstad and Goldmann 1991]. Our main goal is to have all sites compute the clusters of the vertex set of the input graph G so that all of them can know to which cluster is own data belongs. We also assume that data is allocated is such a way that the collection $E_1, E_2, \ldots, E_s$ form a *sunflower* [Erdös et al. 1961]. By exploiting the structure of the sunflower, we showed that all sites can compute the clusters using a communication protocol that exchange at most $O\left(\log\left(\frac{n}{\epsilon^2}\sqrt{1-\lambda}\right)\right)$ bits of communication, where n is the total number data points, ϵ is an approximation factor and λ is a ratio of common data points among sites. To achieve this upper bound we used a well-known technique of spectral sparsification of graphs [Batson et al. 2009, Lee and Sun 2018] and developed a technical lemma that allows us to compute spectral sparsifiers in a distributed setting.

2. Preliminaries

We will introduce some standard notations from graph theory and communication complexity which can be found in [Kushilevitz and Nisan 2006]. In the NOF model there are s sites $P_1, P_2, \ldots, P_s$ and each one has its own input on the set $\{0,1\}^r$. Let X_j be the set of possible inputs for the site P_j, and we want to jointly compute a function $f : X_1 \times X_2 \times ... \times X_s \to Z$ for some finite codomain Z. Each site can only see the others sites's input but cannot see its own input. Hence, a site P_j has access to the input $(x_1, ..., x_{j-1}, x_{j+1}, ..., x_s)$. The communication among the sites is written on a blackboard, where everyone can see it. This is the so-called *blackboard model* of communication. The maximum number of bits exchanged in the protocol over the worst-case input is the *cost* of the protocol. The *deterministic communication complexity* of the function f is the minimum cost over all protocols which compute f.

Let $G = (V, E)$ be an input data graph. On the NOF model we let E_j denote the set of edges that belong to P_j. Also, all sites know the vertices of G. Let $F_j = \{E_1, E_2, ..., E_{j-1}, E_{j+1}, ..., E_s\}$ be the set of edges which P_j can see from the other sites. Given a site P_j, the *symmetric difference on* P_j, denoted Δ_j, is defined as the symmetric difference between all sets in F_j.

A *sunflower* or Δ-*system* is a family of sets $F = \{A_1, ..., A_t\}$ where $(A_i \cap A_j) = \bigcap_k^t A_k$ for all $i \neq j$. A *weak* Δ-*system* is a family F with sets of size ℓ such that $|A_i \cap A_j| = \lambda$ for all $i \neq j$ for some λ [Kostochka 2000]. It is known that if F is a weak Δ-system and $|F| \geq \ell^2 - \ell + 2$, then F is a sunflower [Deza 1974].

Finally, we will introduce some standard notations of spectral sparsification techniques which can be found in [Chen et al. 2016, Batson et al. 2009, Lee and Sun 2018]. Every undirected and weighted graph G has a positive semidefinite matrix associated called its *Laplacian* with the form $L_G = BWB^T$ where B is an signed vertex-edge incidence matrix and W is the diagonal edge-weighted matrix. We say that a subgraph $H \subseteq G$ is an ϵ-spectral sparsifier of G if $(1-\epsilon)x^T L_G x \leq x^T L_H x \leq (1+\epsilon)x^T L_G x$ for all $x \in \mathbb{R}^{|V|}$. If L is a graph Laplacian we say that $x^T L x$ is the *quadratic form* of L. Spectral sparsifiers with approximation factor $\epsilon > 0$ can be constructed in time $\widetilde{O}(\frac{qmn^{5/q}}{\epsilon^{4+4/q}})$ with a number of edges $O(qn/\epsilon^2)$, where n is the number of vertices, m is the number of edges, and $q \geq 10$ a constant [Lee and Sun 2018].

3. Results

In this section we present a communication protocol among s sites in the blackboard NOF model for clustering. We model our data set using a complete undirected weighted graph $G = (V, E, w)$ with n vertices where the edges are allocated among sites.

We define an overlapping coefficient of the edges of G which can be seen as a measure of how well spread out are the edges around the sites.

Definition 1 *The overlapping coefficient on site P_j is defined as $\delta(j) = \frac{|\bigcap_{i \neq j} E_i|}{|\bigcup_{i \neq j} E_i|}$ and the greatest overlapping coeficient is defined as $\delta = \max_{j \in [s]} \delta(j)$.*

In order to perform clustering with high accuracy, we need to make sure that each sites knows at least a large part of the data graph G. In the following theorem we present the analysis of a simple protocol that takes into account the greatest overlapping coefficient δ and makes use of the sunflower organization of data.

Theorem 1 *Let P_j be a site and let $\mathcal{E} = \{E_i\}_{i \neq j}$ be a weak Δ-system with each $|E_k| = \ell$ for $k = 1, 2, \ldots, s$, with a kernel of size λ. Suppose that $s \geq \ell^2 - \ell + 3$. If site P_j sends all the edges in Δ_j, then every other site will know the entire graph G. The number of edges this communication protocol sends is at most $|\bigcup_{i \neq j} E_i|(1 - \delta) + \ell$.*

Proof. We will prove this lemma by showing how each site constructs the graph G. First, a given site P_j computes Δ_j and writes it on the blackboard. Since $s \geq \ell^2 - \ell + 3$, by the result of [Deza 1974], we known that $\mathcal{E}$ is a sunflower with kernel A. At this point all sites $i \neq j$ know Δ_j, therefore, they can construct G using the kernel A of $\mathcal{E}$. In one more round, one of the sites $i \neq j$ writes E_j so that site P_j can also construct G.

In order to compute the communication cost of the protocol, first notice that $\delta = \lambda/(|\bigcup_{i \neq j} E_i|) = \lambda/(|\Delta_j| + \lambda)$, where we used the fact that the union of all edges in every site equals the union of the symmetric difference and the kernel A. Then we have that $\delta|\Delta_j| = \lambda - \delta\lambda$, which implies $|\Delta_j| = \frac{\lambda - \delta\lambda}{\delta} = |\bigcup_{i \neq j} E_i||(1 - \delta)$, where the last equality follows from the fact that $|\bigcup_{i \neq j} E_i| = \lambda/\delta$. Finally, after E_j was sent to the blackboard the communication cost is $|\bigcup_{i \neq j} E_i||(1 - \delta) + \ell$. ■

Corollary 1 *The communication complexity of the protocol of Theorem 1 is $O(\log(\ell\sqrt{s(1 - \delta)}))$*

Proof. First, a site P_j sees $s - 1$ sites and $|E_i| = \ell$ for all $i \neq j$. Then $|\bigcup_{i \neq j} E_i| \leq \sum_{i \neq j} |E_i| \leq s\ell$. Replacing the last result in Theorem 1 we get a total communication cost of $c \leq log(s\ell(1 - \delta)) + \log \ell = 2\log(\ell\sqrt{s(1 - \delta)})$. ■

In the following, we will slightly modify the protocol of Theorem 1 to improve its communication cost together with an application of spectral sparsification. Note that the number of optimal clusters or the optimal multicut in a graph depends on the spectrum of the graph Laplacian [von Luxburg 2007], and therefore, it is important that all sites have a good approximation in spectrum of the graph. We will use the following lemma (with a short sketch of its proof) to construct a sparse graph that approximates the spectra of the original graph so that we can perform clustering in a distributed manner.

Lemma 1 *Let $G = (V, E, f)$ be a weighted undirected graph with cost function f and $E_1, \ldots, E_l \subseteq E$ for some fixed l where $\cup_i E_i = E$. Let $G_i = (V, E_i, f_i)$ be an induced subgraph of G. If $H_i = (V, \hat{E}_i, h_i)$ is an ϵ-spectral sparsifier of G_i, then $H = (V, \bigcup_i \hat{E}_i, h)$ is an ϵ'-spectral sparsifier of G where $h(e) = \frac{1}{c_1 c_2} \sum_i h_i(e)$ and c_1, c_2 denote the minimum and maximum number of sites in which an edge appears and $\epsilon' \geq \frac{c_1 - 1 + \epsilon}{c_1}$.*

Proof sketch. Let L_{G_i} be the Laplacian matrix of G_i. To prove the lemma we showed that $\sum_{i=1}^{s} L_{G_i}$ can be written as a linear combination of graph Laplacians $\{L_{G'_j}\}_{j \geq 0}$ with coefficients in the discrete interval $[c_1, c_2]$. Then we showed that the quadratic form of this linear combination can be bounded from below and above by $(1 - \epsilon)/c_2$ and $(1 + \epsilon)/c_1$ times the quadratic form of L_G, respectively. Finally using ϵ' we obtain that H is a spectral sparsifier of G. ■

Theorem 2 *Let P_j be a site and let $\mathcal{E} = \{E_i\}_{1 \leq i \leq s}$ be a weak Δ-system with each $|E_k| = \ell$ for $k = 1, 2, \ldots, s$, and suppose that $s \geq \ell^2 - \ell + 3$. There exists a communication protocol where after two rounds of communication every site knows an ϵ-spectral sparsifier of the entire graph G with communication cost $O\left(\log\left(\frac{n}{\epsilon^2}\sqrt{1-\delta}\right)\right)$.*

Proof. From [Deza 1974] we know that $\mathcal{E}$ is a sunflower with a kernel A of size λ. First, a site P_j computes a spectral sparsifier $H_j = (V, \hat{\Delta}_j)$ of the induced subgraph $G_j = (V, \Delta_j)$ using the spectral sparsification algorithm of [Lee and Sun 2018]. This way we have that $|\hat{\Delta}_j| = O(n/\epsilon^2)$ where $0 < \epsilon \leq 1/120$. Then site P_j writes $\hat{\Delta}_j$ on the blackboard. Any other site $i \neq j$ constructs an ϵ-spectral sparsifier $H_i' = (V, \hat{E}_j)$ of $G_i' = (V, E_j)$. By Lemma 1, the graph $H = (V, \hat{\Delta}_j \cup \hat{E}_j)$ is a ϵ'-spectral sparsifier of G. In a second round, a given site P_i writes $\hat{E}_j$ on the blackboard. Finally, site P_j receives $\hat{E}_j$ and by Lemma 1 it can also construct an ϵ'-spectral sparsifier for G. Finally, the communication complexity is upper-bounded by $O\left(\log\left(\frac{n}{\epsilon^2}(1-\lambda)\right) + \log\left(\frac{n}{\epsilon^2}\right)\right) = O\left(\log\left(\frac{n}{\epsilon^2}\sqrt{1-\lambda}\right)\right)$. ∎

Acknowledgment. This work is supported by Conacyt research grant PINV15-208 and POSG17-62.

References

Batson, J. D., Spielman, D. A., and Srivastava, N. (2009). Twice-ramanujan sparsifiers. In *Proceedings of the 41st annual ACM symposium on Theory of computing (STOC)*, pages 255–262.

Chen, J., Sun, H., Woodruff, D., and Zhang, Q. (2016). Communication-optimal distributed clustering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, pages 3727–3735.

Deza, M. (1974). Solution d'un problème de Erdös-Lovász. *Journal of Combinatorial Theory, Series B*, 16(2):166–167.

Erdös, P., Chao, and Rado, R. (1961). Intersection theorems for systems op finite sets. *Quarterly Journal of Mathematics*, 12(1):313–320.

Håstad, J. and Goldmann, M. (1991). On the power of small-depth threshold circuits. *Computational Complexity*, 1(2):113–129.

Kostochka, A. (2000). Extremal problems on Δ-systems. *Numbers, Information and Complexity*, pages 143–150.

Kushilevitz, E. and Nisan, N. (2006). *Communication complexity*. Cambridge University Press.

Lee, Y. T. and Sun, H. (2018). Constructing linear-sized spectral sparsification in almost-linear time. *SIAM Journal on Computing*, 47(6):2315–2336.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.