Project and Internship

Mauro Sozio

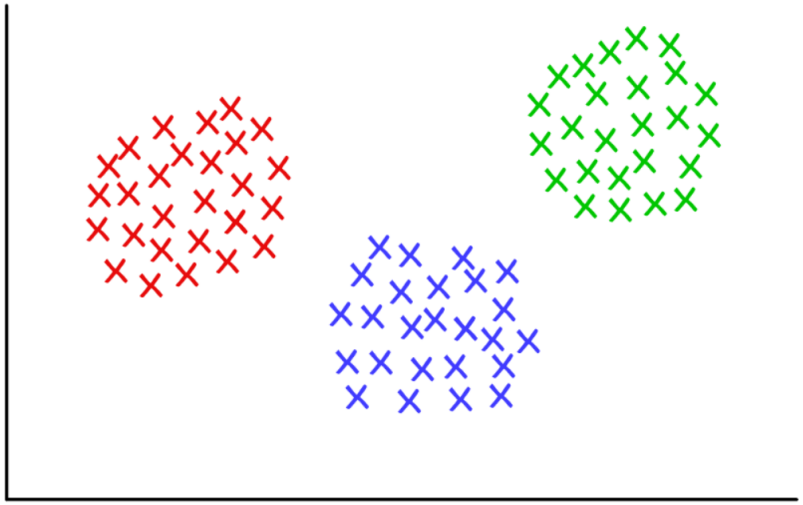
Telecom ParisTech

December 11, 2018

Project

Definition (k-Center Clustering)

We are given a set S of points with $|S| \geq k$ equipped with some metric d and an integer $k > 0$. We wish to find a set $C = \{c_1, \ldots, c_k\}$ of k points (centers) to minimize $\max_{x \in S} d(x, C)$, where $d(x, C) = \min_{c \in C} d(x, c)$.



A 2-approximation algorithm

There is a 2-approximation algorithm for k-center clustering.

1. $X \leftarrow c$ where c is chosen arbitrarily from S;
2. While $(|X| < k)$
   a. select a point c at maximum distance $d(c, X)$, where $d(c, X) := \min_{x \in X} d(c, x)$ (breaking ties arbitrarily).
   b. $X \leftarrow X \cup \{c\}$
   c. $S \leftarrow S \setminus \{c\}$
3. return X

Theorem

The previous algorithm is a 2-approximation algorithm for k-center. Moreover, any approximation algorithm (with polynomial running time) with an approximation factor better than 2 implies $P = NP$.

Strongly well-separated Clusters

We say that k clusters $C_1, \ldots, C_k$ are strongly well-separated if $d(x, y) < d(w, z)$, for every points x, y, w, z such that x, y belong to a same cluster, while w, z belong to two different clusters.

Strongly well-separated Clusters

We say that k clusters $C_1, \ldots, C_k$ are strongly well-separated if $d(x, y) < d(w, z)$, for every points x, y, w, z such that x, y belong to a same cluster, while w, z belong to two different clusters.

Let $d_{\max}$ be the maximum diameter of a cluster in $C_1, \ldots, C_k$. $C_1, \ldots, C_k$ are well separated if and only if the distance between two points in two different clusters is larger than the maximum diameter of a cluster.

Lemma

If there are k well-separated clusters in S, the 2-approximation algorithm for k-center will find those clusters.

k-center with outliers

Definition (k-Center Clustering with outliers)

We are given a set S of points with $|S| \geq k$ equipped with some metric d and integers $k > 0, z \geq 0$. We wish to remove a set Z (outliers) from S, $|Z| \leq z$, and find a set $C = \{c_1, \ldots, c_k\}$ of k points (centers) from $S \setminus Z$ so as to minimize $\max_{x \in S \setminus Z} d(x, C)$, where $d(x, C) = \min_{c \in C} d(x, c)$.

There is a simple 3-approximation algorithm for k-center clustering with outliers [4] (Section 3) and a constant approximation algorithm in streaming [2].

(Default) Project

1. Implement the static approximation algorithm for k-center clustering with outliers [4].
2. Implement the streaming algorithm k-center clustering with outliers [2].
3. Conduct an evaluation of the algorithm proposed on a collection of geotagged tweets that we will provide to you. Evaluate running time and quality of the solution.

The project should be implemented in either C/C++, Java or Python. Efficient codes in C/C++ will receive extra points.

Deadline: February 15th 2019

Weakly well-separated clusters

We say that k clusters $C_1, \ldots, C_k$ are weakly well-separated if $d(x, y) < d(x, z)$, for every points x, y, z such that x, y belong to a same cluster, while x, z belong to two different clusters.

Example of k weakly well-separated clusters which are not strongly well-separated?

Weakly well-separated clusters

We say that k clusters $C_1, \ldots, C_k$ are weakly well-separated if $d(x, y) < d(x, z)$, for every points x, y, z such that x, y belong to a same cluster, while x, z belong to two different clusters.

Example of k weakly well-separated clusters which are not strongly well-separated?

Algorithm for for finding k weakly well-separated clusters?

Axiomatic approach for clustering

Let X be a set of points equipped with a metric d. We wish that a clustering algorithm A have the following properties.

- scale-invariance. Suppose we run the algorithm A on X and we obtain a clustering $\mathcal{C}$, that is, a partition of X. Now divide all distances by any constant α obtaining d'. Run A on X with metric d'. We should obtain the same clustering $\mathcal{C}$.
- richness. A should be able to produce in output every possible partition of X.
- consistency. Suppose we run the algorithm A on X and we obtain a clustering $\mathcal{C}$. We now modify d as follows: we decrease the distance between points in a same cluster in $\mathcal{C}$, while we increase the distance between points in different cluster. Let d' the new distance so obtained. If we run A on X with metric d' we should obtain the same clustering $\mathcal{C}$.

Axiomatic approach for clustering: Results

Kleinberg [6] showed that there cannot be any clustering algorithm satisfying scale-invariance, richness and consistency.

Consistency has been criticized for long time, see for example [7]. In [7], a new weaker property (and more natural) has been defined. Authors then show that there are algorithms satisfying scale-invariance, richness, and the weaker variant of consistency.

In view of the criticism raised in [7] and other papers, Kleinberg's approach is perhaps not so interesting. However, it has the merit of studying clustering in a principled way.

Internship topic 1

The idea of defining good properties for clustering and then show positive/negative results is appealing.

Can we consider other properties? Such as being able to find weakly/strongly separated clusters? Other properties?

The variant with outliers has not been studied in this context, yet. Can we show positive/negative results in this case as well?

Internship topic 2

An axiomatic approach for graph clustering (aka community detection) would be very interesting.

The problem is perhaps more challenging, in that, the distance between nodes in a same cluster (aka community) depends on the nodes belonging to that cluster.

Which properties make sense in this context? Can we obtain positive/negative results?

Fully Dynamic Clustering: Adversarial Model

We wish to study the k-center clustering in a fully dynamic setting where points can be added or removed (almost) arbitrarily.

Fully Dynamic Adversarial Model. The adversary fixes a (possibly countably infinite) sequence O of operations, such as add or remove a point. The algorithm does not know the sequence O in advance, however, any randomness used by the algorithm is generated after the adversary fixes O.

Fully Dynamic k-center clustering

Can we maintain a $2+\epsilon$-approximation as points are added and removed, while requiring a small number of operations on average?

Fully Dynamic k-center clustering

Can we maintain a $2+\epsilon$-approximation as points are added and removed, while requiring a small number of operations on average?

Naïve solution: Whenever a new point is added or removed, recompute the current solution from scratch: $\Omega(|S_t| \cdot k)$ per update! (S_t the current set).

Theorem (from [5])

For any $\epsilon > 0$, for any sequence T of operations the fully dynamic clustering algorithm in [5] maintains a $2+\epsilon$-approximation algorithm for the k-center problem, while requiring $O(\frac{log\delta}{\epsilon} k^2 T)$ expected running time and $O(\frac{\log \delta}{\epsilon})N$ operations space under the fully dynamic model, where N is the largest number of points at any given time. The result holds also with high probability.

Internship Topic 3

Develop a fully dynamic clustering algorithm for k-center with outliers with:

- average number of operations per update being sublinear in the size of the input (ideally some polylog function)
- the approximation guarantee should be constant.

References I

- Arnaud Guerquin, Hubert Chan,Mauro Sozio. Fully dynamic clustering, Technical Report, 2017.
- Richard Matthew McCutchen, Samir Khuller: Streaming Algorithms for k-Center Clustering with Outliers and with Anonymity. APPROX-RANDOM 2008:165-178.
- Mikkel Thorup: Worst-case update times for fully-dynamic all-pairs shortest paths. STOC 2005:112-119.
- Moses Charikar, Samir Khuller, David M. Mount, Giri Narasimhan:Algorithms for facility location problems with outliers. SODA 2001: 642-651
- T.-H. Hubert Chan, Arnaud Guerqin, Mauro Sozio: Fully Dynamic k-Center Clustering. WWW 2018: 579-587
- Jon M. Kleinberg: An Impossibility Theorem for Clustering. NIPS 2002: 446-453
- Clustering Redemption Beyond the Impossibility of Kleinbergs Axioms. Vincent Cohen-Addad, Varun Kanade, and Frederik Mallmann-Trenn (NIPS) 2018.