Graph Mining (MPRI): Streaming Algorithms for k-center Clustering with Outliers

Mauro Sozio
Télécom ParisTech, Paris, France
sozio@telecom-paristech.f

December 14, 2018

General Information. The main goal of this project is to implement a theoretical paper and analyze its performances (running time, quality of the solution) on real data. You are asked to implement two algorithms for k-center clustering with outliers, both in a streaming and static environment (see below). The algorithm can be implemented in C, Java or Python. More efficient codes will receive more points. We recall that the maximum number of points for this project is 4. The project can be implemented in groups of maximum 2 people, however, we will be more demanding for group projects. You are allowed and encouraged to exchange ideas with other students and the professor, however, everybody should implement the algorithm on his/her own. In case we suspect there has been some misconduct, all students involved will receive zero points for this project. The project is not strictly required to pass the course, however, in case the project is not completed the maximum grade is capped at 16. The deadline for this project is **February 15th, 2019**.

Tasks.

1. Implement the algorithm presented in [CKMN01] (Section 3) for the static case. The paper is available at this link `https://www.cs.umd.edu/~mount/Papers/soda01-outlier.pdf`.

2. Implement the Algorithm 3.1 in [MK08]. The paper is available at `https://www.researchgate.net/profile/Eli_Ben-sasson/publication/221462630_Tensor_Products_of_Weakly_Smooth_Codes_Are_Robust/links/0f31753b2d18823cbf000000.pdf#page=176`. In step 3 of the algorithm, you can use the algorithm implemented in the previous point. You should set $\alpha = 4, \beta = 8, \eta = 16$ as described in the analysis of the algorithm. The algorithm has also a parameter m, you can ignore this parameter (i.e. set $m = 1$).

3. Evaluate the algorithm developed in point 2 on a collection of geotagged tweets provided at the link on the course webpage. The dataset contains for each tweet a timestamp when the tweet is posted and some geographic information (see the readme file describing the file format). Your task is to cluster the tweets that are provided to you. The algorithm should be implemented in a streaming setting, where data is stored in a file and you have only sequential access to the file while the total amount of main memory is $O(kz)$. In particular, you should:
   (a) Evaluate the approximation ratio of the algorithm on the data provided to you when $k = 20$ while the number of outliers $z = 10$. That is, you should try to estimate the ratio between the maximum radius of the clusters computed by your algorithm and the maximum radius of an optimum solution. To this end, you should try to give an upper bound on the optimum solution. Describe in the report how you computed such an upper bound. How does the approximation guarantee on the data compare with the worst-case approximation guarantee?
   (b) consider all values for z in $\{10, 20, 30, 40, 50\}$. Include in the report the following two plots: approximation ratio of the algorithm on the data as a function of z, running time (in secs) as a function of z. Discuss the plots you obtained.

What to send. You should send the code you wrote, as well as a short report (max $3-5$ pages) with the two plots discussed above as well as a discussion on the results. You should discuss in the report whether the plots you obtained are expected and why. Any implementation issue you encountered and how you solved it should also be discussed. In the case you obtain something unexpected you should try to understand why and discuss this in the report. You should also discuss how you computed a lower bound on the optimum solution. You should send all the material to `sozio@telecom-paristech.f` specifying in the email subject: "MPRI 2019 project".

References

[CKMN01] Moses Charikar, Samir Khuller, David M. Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the Twelfth Annual Symposium on Discrete Algorithms, SODA*, pages 642–651, 2001.

[MK08] Richard Matthew McCutchen and Samir Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 165–178. Springer, 2008.