DISSERTATION PROPOSAL

Yuyan Wang

“Foundations of Clustering: New Models and Algorithms”

Friday, December 4, 2020
3:00pm EST
Zoom: https://cmu.zoom.us/s/94262620698

In this dissertation, we study clustering, one of the most common unsupervised learning problems. This dissertation covers recent developments in both clustering theory and machine learning practice. In particular, it explores how to bridge the gap between theory and practice by making them benefit from each other.

In the first and second chapters, we study how to build the mathematical optimization framework for one classic clustering problem: hierarchical clustering. The first chapter focuses on new objective function design for hierarchical clustering on point inputs in a Euclidean space. It provides theoretical guarantees for a popular heuristic commonly used in practice. The second chapter studies fairness in hierarchical clustering. It defines fair hierarchical clustering trees and discusses how to design algorithms that find fair solutions for previous hierarchical clustering objectives established by the community.

In the third and fourth chapters, we give new algorithmic designs that could be used to speed up famous clustering algorithms in scenarios where they are known to be inefficient.

The third chapter considers average-linkage, which is one of the most commonly used hierarchical clustering approach. We introduce a new technique named "clustering embedding", which maps clusters into points in Euclidean space. The points are then used as surrogates for the clusters, to facilitate approximate nearest neighbor search. By iteratively doing this, we reduce the previous quadratic bound on running time to only slightly super-linear, making average-linkage scalable to big datasets.

In the fourth chapter, we turn to a new data input format other than the conventional sample-feature matrix - relational database, which is a memory-efficient way to store data. The naive way of running conventional ML algorithms requires, as a data preprocessing step, recovering the original dataset from the relational database representation. This approach loses the memory efficiency of relation database. We propose k-means algorithms that could directly work on any relational database without recovering the sample-feature matrix. We show how to implement the famous k-means++ algorithm and find a constant approximation for the optimal k-means solution.

The fifth and sixth chapters propose future work which will use machine-learned models to provide advice to clustering algorithms. Potentially, the algorithm could benefit from the advice it is given and thus produce solutions with higher performance and/or efficiency.

Finally, the seventh chapter explores an application-based design of clustering algorithms. We introduce a novel streaming clustering task where we have a stream of periodically-created particles traveling along a trajectory. We observe sporadic, noisy reports from the particles with spatial-temporal information that keep arriving, without knowing which particle is being observed. We seek a streaming algorithm that maintains accurate estimates of the number and locations of the particles that are currently en route, by clustering the data reports online, where ideally every single cluster represents a moving object.