PHD DISSERTATION DEFENSE

Computer Science

Clusterability, Model Selection and Evaluation

PhD Candidate: Kaixun Hua

Directed by: Prof. Dan A. Simovici

Date: Thursday, April 4th 2019, 12:30 PM – 1:30 PM

Location: IT Lab, Science Building (S-3-143)

Abstract

Clustering is a central topic in unsupervised learning and has a wide variety of applications. However, the increasing needs of clustering massive datasets and the high cost of running clustering algorithms poses difficult problems for users, while to select the best clustering model with a suitable number of clusters is also a primary focus. In this dissertation, we mainly focus on determining whether a data set is clusterable, and what is the "Natural" number of clusters of in a dataset.

First, we approach data clusterability from an ultrametric-based perspective. A novel approach to determine the ultrametricity of a dataset is proposed via a special type of matrix product, and via this measure, we can evaluate the clusterability of it. If a dataset has a unimodal or poorly constructed structure, its ultrametricity will be lower than other datasets with the same cardinality. Also, we show that by promoting the clusterability of a dataset, a poor clustering algorithm will perform better on the same dataset.

Secondly, we present a technique grounded in information theory for determining the "Natural" number of clusters existent in a data set. Our approach involves a bi-criterial optimization that makes use of the entropy and the cohesion of a partition. Additionally, the experimental results are validated by using two relatively distinct clustering methods: the k-means algorithm and Ward hierarchical clustering and their contour curves. We also show that by modifying the parameter, our approach can handle dataset with heavily imbalanced clustering structure, which is further complicated in practice.