Time: 11:00am-12:00pm, Wednesday July 14, 2010
Place: #1 Science Building 1114, PKU (理科一号楼1114)

Talk: Multi-Sample Data Spectroscopic Clustering of Large Datasets using Nystrom Extension
Speaker: Tao Shi
Department of Statistics
Department of Computer Science and Engineering
Ohio State University

Abstract:

Spectral clustering algorithms have shown promising results in statistics, bioinformatics, genetic study, machine learning, and other scientific fields. These spectral algorithms cluster observations (of size n) into groups by investigating eigenvectors of an affinity matrix or its corresponding Laplacian matrix, both of which are size of n by n. However, the computation involved in eigen-decompostion of an n by n matrix is expensive or even infeasible when the sample size is large. To overcome the computation hurdle, subsampling techniques, such as Nystrom extension, have been used in approximating eigenvectors of large matrices. In this talk, we discuss some spectral clustering algorithms and the statistical properties of such approximations and their influence on the accuracy of those algorithms. We found that the perturbation of spectrum due to subsampling could lead to large discrepancy among clustering results based on different subsamples. In order to provide accurate and stable clustering results for large datasets, we propose a method to combine multiple sub-samples using data spectroscopic clustering and the Nystrom extension. In addition, we propose a sparse approximation of the eigenvectors to further speed up the computation. Simulation and experiments on real data sets showed that this multi-sample approach is fast and accuracy. The results in this talk involve joint works with Prof. Mikhail Belkin (Computer Science, OSU), Dr. Jared Schuetter (Statistics, OSU), and Prof. Bin Yu (Statistics, UC Berkeley).