Title: Clustering on Imbalanced Data

Abstract:
In many practical problems, number of data form difference classes can be quite imbalanced, which could make the performance of the most machine learning methods become deteriorate to a certain degree. As far as we know, the problem of learning from imbalanced data continues to be one of the challenges in the field of data engineering and machine learning, which has attracted growing attentions in recent years. However, most of the researches in the area focus on supervised learning, and imbalanced data clustering in unsupervised environment has yet to be well studied. In this talk, we will first formally describe and compare the class imbalance problem on supervised and unsupervised learning setting. Then, we describe the key challenge of the problem of clustering on imbalanced data, which is called uniform effect. Accordingly, we have proposed a solution called SMCL for this problem. The advantages of SMCL are three-fold: (1) It inherits the advantages of competitive learning and meanwhile is applicable to the imbalanced data clustering; (2) The self-adaptive multi-prototype mechanism uses a proper number of subclusters to represent each cluster with any arbitrary shape; (3) It automatically determines the number of clusters for imbalanced clusters. Finally, some challenging problems in this topic are explored as well.