Title： Clustering on Imbalanced Data

Abstract: In many practical problems, the number of data forming difference classes can be quite imbalanced, which could make the performance of the most machine learning methods become deteriorate to a certain degree. In general, the problem of learning from imbalanced data is nontrivial and challenging in the field of data engineering and machine learning, which has attracted growing attentions in recent years. In the literature, most of the existing works are focusing on supervised learning only. As far as we know, imbalanced data clustering in unsupervised environment has yet to be well studied. In this talk, we will first formally describe and compare the 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 propose 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. Empirical studies show the promising results.