Week 20

对于聚类的多视角自步学习

摘要：利用多视角中的信息可以提高聚类准确率。然而现有的大部分聚类算法是非凸的，因此倾向于陷入局部最优，尤其是在有异常值和缺失值的情况下。为了解决这个问题，我们提出了MSPL（多视角自步学习）。通过从简单到复杂的样本，以及简单到复杂的视角，来学习多视角模型。不是将样本或视角二分地划分为简单或复杂，我们设计了一个新的概率光滑加权方案。将多视角应用于聚类和复杂度定义，被证明有利于得到最优解。

介绍：聚类的目标是在无监督模式下找到有意义的样本组，来用于数据分析。单独地应用每个视角会使聚类不准确，因为单独的视角不能综合地描述所有的样本。因此通过多视角来开发它们之间的联系是有利的。

现有的大部分算法致力于非凸的目标函数，容易导致局部最优，对于噪声和异常值尤其敏感。一个启发式的方法是用不同的初始化条件多次运行算法，然后从结果中选取最优解。

这个策略比较耗时，而且在非监督的条件下很难实现，因为对于模型选择没有明确的准则。

通过模拟人类的学习，自步学习首先试图在简单样本上训练一个模型，然后逐渐将复杂的样本考虑进去。这已经被证明有利于避免局部最优解，而且提高了一般的结果。与单独视角中样本的复杂度相似，多视角样本中也有简单和复杂的视角，而且对于特定的多视角样本，简单和复杂视角之间的区别可能不一样。例如GIST在识别自然景物图像时达到了高准确率，而CENTRIST在分类室内环境图像时表现很好。

在本文中我们提出了用于聚类的多视角自步学习，通过同时考虑视角和样本的复杂度，来学习多视角的模型。我们设计了一个光滑的加权方案，继承了逻辑斯蒂函数的优点，提供了概率权重。

问题描述：

传统的k均值聚类算法只适用于单视角的数据，因此用多视角来提升聚类表现很有必要。

k均值聚类用K个原型向量来刻画数据并最小化均方误差损失函数。给出n个样本，k均值聚类的目标可以表示为：



其中是中心矩阵，是第个簇的聚类中心。B是聚类分配矩阵，其中值为0或1。

2.1多视角自步学习

一般的自步学习模型由一个施加在所有样本上的权重损失项和一个施加在样本权重上的正则项组成。在模型选择过程中，通过逐渐增大惩罚项，更多的样本通过一个自步方法被自动从简单训练集包含进复杂训练集。不仅样本有简单复杂之分，视角也有。通过同时考虑样本和视角的复杂度，我们提出了对于聚类的多视角自步学习：





是第个视角下的权值向量，。表示用来决定样本和视角选择的正则项。之前的是



因此第个样本在第个视角下的最佳权值为。其中是第个样本在第个视角下的重构误差。参数控制着模型学习新样本的速率，在优化过程中通常迭代性地减小。

上面的是硬加权。现实应用中软加权更能反映样本的真实重要性。因此我们为自步学习提出一个新的正则项：

（1）

第个样本在第个视角下的最佳权值可以通过解下式确定：



得到。

注意到（1）式可看做适应的逻辑斯蒂函数，因此它继承了所有逻辑斯蒂函数的优点：无穷可微，严格凸，利普希茨连续。而且提供了一个权重的概率解释。

3优化

在B,C固定的情况下，W可以如下优化：



4理论分析