谱聚类问题的一种连续优化模型

张瑞

摘 要

聚类问题是数据挖掘和统计分析等重要领域的主要研究课题之一, 聚类分析被广泛应用于机器学习、模式识别、图像分析、信息恢复、数据压缩和计算机绘图等众多领域. 谱聚类方法处理聚类问题得到的结果效果好且易于实现, 已经发展成为最成功的聚类方法之一. 然而现有的谱聚类模型也有一些缺点, 比如, 谱聚类模型的解并不能显式地给出聚类结构, 需要增加一个后处理过程才能将聚类结构显示出来.

通常, 聚类问题可以化归为图划分问题. 本文综述了谱聚类和图划分问题的进展, 提出了一类新的连续优化模型,不仅从理论上分析了新模型的性质, 给出了块坐标下降算法, 还通过数值实验验证了新模型及算法的有效性.

具体地, 第 2 章约定了一些数学符号和定义,给出了图的加权邻接矩阵、连通分支等概念,介绍了图的拉普拉斯矩阵的一些重要性质,包括图的拉普拉斯矩阵零特征值的重数等于图的连通分支数.

第 3 章先介绍了谱聚类与图划分的联系,接着介绍了经典的 k-means 聚类模型和几种常见的谱聚类模型的建模过程及其算法框架. 其中, k-means 模型将类内部的平均相似度尽可能大作为优化目标,谱聚类模型将类之间的相似度尽可能小、类的规模尽可能均衡作为聚类目标. 这两类模型要求预设准确的类的划分个数,这在实际中是个难题.

第 4 章构造了一个对聚类数目有较好鲁棒性、以聚类以后得到的图（称为结果图）的 0-1 邻接矩阵为解的谱聚类新模型. 新模型不需要准确的类的划分个数, 代之以一个上界估计. 新模型可以找到不同类的对象之间相似度的和尽可能小、类的数目尽可能大的聚类方案.本文用原图和结果图的邻接矩阵计算不同类的对象之间相似度的和, 用图的拉普拉斯矩阵零特征值的重数等于图的连通分支数这个性质计算类的数目,将两者的线性组合作为优化目标函数,得到一个秩最小化问题. 经适当地松弛, 最终得到一类新的连续优化模型. 通过对新模型进行分析,我们得到了全局最优解的一些重要性质.

第 5 章介绍了求解谱聚类新模型的块坐标下降算法, 分析了算法的全局收敛性, 通过几个数值例子,验证了新模型和块坐标下降算法的有效性.

关键词：聚类, 图划分, 谱聚类, 图的拉普拉斯矩阵, 块坐标下降算法