数据分析中的优化模型与算法

陈雅丹

摘 要

在大数据和人工智能时代, 随着数据规模不断扩大, 如何分析和处理数据中隐藏的丰富的信息越来越受到人们的关注. 本文从模型、理论与算法方面研究基于数据图的聚类问题, 应用于聚类和特征选取的正交非负约束优化问题以及大规模线性响应特征值问题的求解.

对于谱聚类问题, 我们考虑了一个等价于 Ratio Cut 问题的模型, 将一个离散问题转化为一个连续优化模型. 在该模型中, 每个数据点归属于哪一类可以直接由模型的解所决定. 特别地, 我们设计了一个分裂技巧, 将原模型的变量复制成三块, 使得新的约束优化问题的目标函数和约束只与部分变量相关, 且模型具有很好的分块结构. 同时, 我们提出了一个基于交替方向乘子法的框架求解分裂模型, 在此算法框架下, 每个子问题都易求解且均有显式解. 除此之外, 我们证明了在一定假设条件下子列的收敛性. 针对所提出的算法, 我们通过数值实验验证了其有效性. 与此同时, 与已有的聚类方法相比, 基于所提出算法的模型在处理包括文本数据集, 手写数字集以及人脸图像集在内的来自于不同领域的真实数据集的聚类问题时, 有明显的优势.

针对一般的正交非负约束优化问题, 我们提出了一个基于块坐标下降法的非精确增广拉格朗日方法求解此问题. 该方法可以看作是增广拉格朗日方法和块坐标下降法的结合, 包括内迭代和外迭代两部分组成, 内迭代计算原始变量使其对应的增广拉格朗日函数的次微分满足指定的精度, 外迭代更新对偶变量和自适应参数. 除此之外, 我们也给出了所提出算法在一定的假设条件下子列的收敛性结果. 将正交非负约束优化模型具体应用在谱聚类问题, 正交非负矩阵分解问题以及无监督特征选取问题中, 数值实验结果也显示出所提出算法对于正交非负约束优化问题的求解是有效的.

对于在计算电子和分子的能量激发态有重要应用的大规模线性响应特征值问题, 通常, 这类问题的变量规模非常巨大. 传统的特征问题求解器可能并不适用于此问题. 基于此, 我们提出了一个基于罚函数的迹极小化无约束模型, 同时也采用梯度型算法和 Barzilai-Borwein 步长来求解, 并提出了一个重启策略, 使得算法能够以更快的速度和更高的精度收敛. 所提出的模型避免了正交约束, 这大大节省了算法的计算时间. 除此之外,数值实验也展示了我们所提出的算法和重启策略是有效的.

关键字: 谱聚类, 正交约束, 非负约束, 增广拉格朗日法, 块坐标下降法, 特征选取, 线性响应特征值问题