距离几何问题的理论，算法及应用

盛镇醴

摘要

距离几何问题, 也就是依据部分点与点之间的距离确定点的坐标的问题, 在最近几十年成为了一个跨学科的研究热点. 该问题在画图, 生物学, 环境监测以及传感器网络等领域有着广泛的应用. 一般来说, 距离几何问题是NP-难的. 尽管有很多算法存在, 该问题并未被彻底解决. 本文就是围绕这一问题, 针对其建模, 算法分析及设计进行研究.

距离几何问题通常都建模成一个优化问题来求解, 不同的文献提出了多个误差函数将其建模成不同的优化问题. 本文首先介绍各文献中使用的误差函数, 从可微性和鲁棒性等方面分析其优劣, 并在此基础上提出了几个新的误差函数, 新的误差函数具有克服点与点过于靠近的优点. 我们也介绍了正则项, 它有助于得到一个低秩的解, 但正则参数的选择是一个悬而未决的问题.

基于已有的几何构建算法, 我们提出改进的带误差函数优化的几何构建算法(GBEM-LLS 和GBEM-NLS). 我们的贡献主要在于两个方面. 首先, 我们分析了误差累积的成因, 并用小例子说明了计算顺序对算法结果的重要影响, 我们提出了新的规则谨慎选择每次加入的点, 我们用数值实验证明了新规则的有效性. 其次, 我们提出了一个新的算法框架, 将误差函数优化整合进了原有算法, 从而有效地控制了误差的累积, 这是原有算法最大的缺陷. 大量的数值实验表明, 新的算法能在相对短的计算时间内, 得到精确的定位结果. 例如, 给定原子间小于6Å的全部距离, 随机添加10% 的乘性噪音, 一个有8629 个原子的蛋白质能够在4分钟之内被定位, 得到的最小均方误差RMSD 为0.24Å. 已有的最新的几何构建算法在给定距离小于6Å的情况下最多只能容忍0.01%的噪音, 我们的新算法可以处理10%的误差, 对距离误差的鲁棒性大幅提高, 这使得新的算法能够真正用来解决实际问题.

研究发现,距离几何问题与非线性降维问题有着紧密的联系, 我们用一个简单的例子说明这种关系. 基于这种认识, 我们率先将求解非线性降维问题的拉普拉斯特征映射算法引入到距离几何问题的求解中. 我们利用稀疏的距离矩阵构造相应的拉普拉斯矩阵, 计算其最小的几个特征值对应的特征向量, 得到一个低维嵌入. 我们进一步根据距离信息将映射后的点放缩到合适的量级, 并以此为初值, 使用梯度下降算法进一步调整点的坐标, 得到算法LEEM. 数值实验表明, 该算法能够快速而精确地求解随机产生的传感器网络定位问题和原子分布比较均匀的蛋白质折叠问题. 例如, 给定原子间小于6Å的全部距离, 随机添加10%的噪音, 一个有3944个原子的蛋白质能够在33.5秒被定位, 得到的RMSD 为0.65Å.

除此之外, 我们还研究了针对距离几何问题的分布式算法. 分布式算法是求解大规模问题的常用手段, 但该问题有其特殊性, 因为我们已知的只有点之间的局部距离信息, 缺乏有效的空间信息, 如何有效地分解原问题是一大难点. 好的分块算法既要尽可能减少距离信息的损失, 又要保证拥有足够的重叠区域, 使得

各子块能够精确地拼接. 已有的分布式算法主要使用symrcm 顺序对距离矩阵指标进行置换重排, 使得矩阵非零元都集中在主对角线附近, 再基于置换后的矩阵分块. 我们另辟蹊径, 通过计算拉普拉斯特征映射得到原图的结构信息, 再使用k-means 聚类算法对整个图进行无重叠区域的分块, 并使用前文提出的选点规则构造重叠区域, 以及GBEM 算法进行子块定位, 最后将各块拼接起来, 得到一个新的分布式算法LSCDL. 初步的数值实验表明, 我们的算法能够非常合理地对原图分块.

关键词: 距离几何问题, 误差函数, 几何构建算法, 误差累积, 拉普拉斯映射, 分布式算法, 拉普拉斯分块