摘要

近年来，随着大数据技术的不断应用，数据的保护与隐藏越来越受到人们的重视。本文主要针对不同情况下的数据隐藏问题做出了模型的建立和求解，为数据的保护提供了一定的思路和方案。

对于问题1，我们对一般的二元数据表隐私保护问题进行了模型的建立和求解。我们引入了信息损失量、差异度距离等概念，通过建立使得信息损失量最小的最优化目标函数，来监督隐藏数据量最少。在求解过程中，我们使用了聚类分析中的bi-kmeans算法，把两条数据中的不同属性值的个数作为两条数据之间的距离，先找到簇中距离最远的两个点，最终形成新的簇。之后将隐匿信息较多的元组进行了重新聚类划分优化。最终得到附件表中二元1需要隐匿20个数据，二元2中需要隐匿358个数据。

对于问题2，属性取值由二元变为了多元。为了提高运算准确度，我们引入了贪心算法来为数据初步分组并利用模拟退火算法来进行模型的优化。首先，我们运用贪心算法，从寻找隐匿数据量为0的小组开始，依次组好隐匿数目最少的数据组，以此为局部最优目标进行迭代，最终实现对数据的初步分组，获得一个较好的初始解。其次，获得初始解后，再用模拟退火算法进行优化，以隐匿数据量总数最小为目标函数，按照自行设计的新解产生方法挑选三个数据组中的一个数据，再数据组之间进行互换，不断迭代最终得到全局最优解。最终得到附件表中多元1需要隐匿628个数据，多元2中需要隐匿7946个数据。

对于问题3，需要额外考虑个体的保护重数，即每个分组中至少需要包含（）个元组。首先，我们优化了贪心算法思路，把作为分组的约束条件之一，数据组内数据条个数必须大于。其次，我们引入了的k-隐匿概念。再次，我们修改了最优化目标函数的约束条件，对其进行重新求解。然后，我们继续使用第二问中的贪心算法和模拟退火算法。最终，得出表中四组数据的隐匿个数分别为, = 2 ：“56、458、806、9547”， = 5：“90，674，979，10373”， = 8 ：“102，801，1063，10414”，依次为二元1，二元2，多元1，多元2。

对于问题4，在情形1中，以隐藏数据量最少为目标函数，同时增加约束条件，即每个数据条的隐匿数量必须为0，1，或全隐。为此，我们要使全隐的数据条个数尽可能少。首先，针对不同的保护重数，使用贪心算法对原始数据集分组，具体步骤如题3中所示。其次，我们对模拟退火算法进行优化，当组中的数据条的隐匿数目大于1时，视作对所有数据隐藏数据，并以此设计代价函数。在情形2中，在算法分组时，我们忽略不能隐藏的数据，只关注能隐藏的数据并以此为约束条件，优化贪心算法和模拟退火算法。

关键字：模拟退火算法，贪心算法，聚类分析，