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。为了更好的分析问题，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，来使得隐藏数据量最少。在求解过程中，我们使用了聚类分析中的 bi-kmeans算法，将差异度距离最小的元组划分在一起，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x个数据，二元2中需要隐匿y个数据。

对于问题2，属性取值由二元变为了多元，此外数据总量也大幅增加。为了减小时间复杂度，提高运算准确度，我们引入的贪心算法和模拟退火算法来进行模型的求解。我们首先对第一问中的差异度距离进行更正，使用元组间距离来建立最优化目标方程。在进行模型求解是，首先使用贪心算法对所有数据进行初步分类获得局部最优解，之后使用模拟退火算法，通过设置合理的迭代次数与温度下降速率来获取全局最优解。最终得到附件表中多元1需要隐匿x个数据，多元2中需要隐匿y个数据。

对于问题3，需要额外考虑个体的保护重数，即每个分组中需要至少包含（）个元组。因此，我们引入的k-隐匿概念。同时修改最优化目标函数的约束条件，重新进行求解，我们继续使用第二问中的贪心算法和模拟退火算法，得出表中四组数据的隐匿个数分别为：

对于问题4，我们解决的是情形2，对于不能隐藏的数据，我们引入隐私数据来增加约束条件，使得至少两维数据被固定，无法进行隐藏操作，重新建立模型后，继续使用问题2的方法进行求解。

最后，我们进行了模型的评价与改进，通过对比聚类算法与贪心、模拟退火算法，发现后者的时间复杂度与计算准确度均有较大提高，但可能仍然无法得到最优的分类与最少的数据隐匿。

关键字：

1. 问题背景与重述

随着医疗系统、交通枢纽等流量平台信息化建设及其在商业调查和科学研究的需要，大量用户数据被收集并发布，其中包含了大量个人隐私信息，而信息共享又是不可避免的。通过适当的隐藏部分信息，就可以保证发布的数据中隐私信息不会被泄露，也能保证发布数据的有效性以供实际需要。因而，找到合适的隐藏信息的方式是亟待解决的问题。

随着大数据分析技术的迅猛发展，研究者以及各个商业公司迫切的需要从大数据中挖掘出有价值的信息。要想从大数据中挖掘信息，首先要有足够的可公开的数据，但是当大规模数据拥有者比如医院、政府、大数据公司等，对外发布数据时，不可避免的会涉及到公民的隐私问题。如果最大限度的保护公布数据的统计特征，又不泄露公民的隐私显得格外重要。

问题一：

针对附件中的两组二元数据，在保证每个个体的信息都能得到保护的前提下，分析如何进行信息隐藏，使得隐藏量最小，并建立二元数据表隐私保护的一般数学模型。

问题二：

研究附件中的两组多元数据，在保证每个个体的信息都能得到保护的前提下，给出一个隐藏数据量最少的方案，并建立多元数据隐私保护的一般数学模型。

问题三：

研究附件中的数据，考虑到每个个体能够隐藏在一个至少包含P个个体的数据组中，建立多重保护数据隐藏的数学模型，并分别给出p=3, 5, 8情形的最佳隐藏方案。

问题四：

对多元数据限定了隐藏模式，给出三种隐藏模式，分别对其进行数据隐藏，得到最佳隐藏方案。

1. 问题分析

2.1问题一分析

1. 模型假设

1.

2.

1. 符号说明

|  |  |
| --- | --- |
| 符号 | 意义 |
| X | 所有的数据 |
| Ai | 第i个数据 |
| AiBj | 第i个数据属性值为j的下标集合 |
| Ui | 等价类 |
| Ci | 等价类中的数据 |
| Ci\* | 隐匿后的数据 |
| Ti | 第i个等价类的数据隐匿量 |
| T | 数据隐匿总量 |
| P | 模拟退火算法中接受新解的概率 |
| α | 降温系数 |

1. **问题一模型建立与求解**

**5.1模型的准备**

在建立模型之前，我们首先要了解一些数据集中常见的概念，如下：

（1）标识符：可以直接确定一个个体，如：身份证号，姓名等。

准标识符集: 可以和外部表链接来识别个体的最小属性集，如：邮编，生日，性别等。

(2)链式攻击：为了保护用户隐私，不让用户的信息泄露，在发布信息时，可以删去标识符信息，试图达到保护隐私的目的。但是攻击者手上如果存在其他信息表，可以通过准标识符信息进行链接，就可以推断出标识符信息。这就是链式攻击

(3)隐藏处理：为了保护用户信息，我们需要对准标识符信息进行一定的处理，通过隐藏一些数据来使表中任意数据都可以找到与其相同的数据，即对标识符进行隐藏处理。

**5.2模型的建立**

在题目所给数据中，可以认为已经将标识符数据隐藏，所有数据都为准标识符数据，因此，只需要对准标识符数据进行隐藏处理。

为方便讨论，假定附件数据中，共有M维数据；即m行，每个数据均有k个属性，即k列；每个属性有n种取值，即n元数据。

记所有的数据为X，第i个数据为Ai，则有：

记AiBj为第i个数据属性值为j的下标集合，下标为该属性位于数据的第几列。为了使得隐藏数据量尽可能的少，现引入差异度距离和信息损失了概念。

（1）二元信息间距离的定义：

假设两个信息间，有n个属性不同，则认为这两个信息的间的距离为n，例如两个信息 Am和 An为间的距离记为dmn。

(2)等价类的定义

在附件数据中，属性取值相近的若干信息组成一个类，将这些信息隐匿掉部分数据，使其具有相同的属性值，便称该类为一个等价类，记为。假设一个等价类种包含t个数据，记为Ci。

(3) 等价数据的定义

经过数据隐藏组合后的某个等价类Ui，类中的数据经过数据隐藏之后都相同，这些信息统称为等价数据。

(4) 数据隐匿量的定义

记第i个等价类U数据隐匿量为Ti，指类中所有数据隐匿的总量，等价于数据属性不同的个数。记为X集合内的相同属性个数。

在题目所给数据中，可以认为已经将标识符数据隐藏，所有数据都为准标识符数据，因此，只需要对准标识符数据进行隐藏处理。为了使得隐藏数据量尽可能的少，现引入差异度距离和信息损失了概念。

对于表1与表2的二元组信息，数据取值只有0，1两种情况，可直接认为数据间的差异即为同数据的数值差异之和。

现给出元组的信息损失定义。给定元组 表示 在属性 上的取值, 则 泛化为 的信息损失量 定义为:

其中, 表示 泛化的精度损失权重.

根据定义概念，可知整个表的信息损失量为每个元组的信息损失量之和，因此给出整个表的信息损失量定义：

而题目中的要求即为整个表的信息损失量值最小化，于是，我们我们可以建立单目标规划模型：

**5.3模型的求解**

（一）初步分类

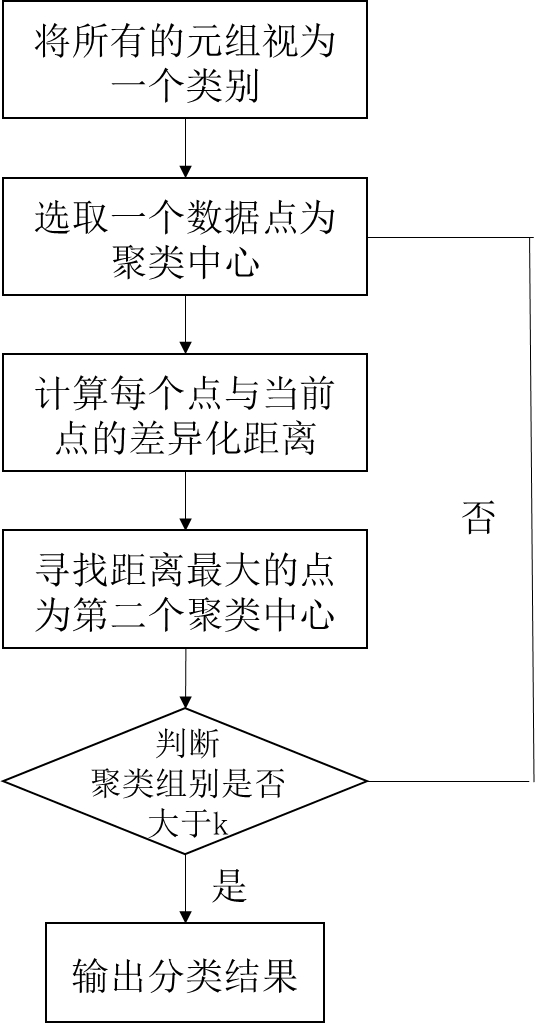
为了尽可能减小信息损失量，可以考虑根据相似度对数据进行分组，即将相似度较高的数据元组分在一组，因此可进行初步分组。

对于所要求的二元数据信息损失量值最小化，我们可以使用聚类算法来进行求解。

聚类(Clustering)是按照某个特定标准(如距离)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。也即聚类后同一类的数据尽可能聚集到一起，不同类数据尽量分离。

我们期望将所包含信息相近的元组尽可能的分在同一类中，因此考虑使用相异度距离作为聚类标准。我们根据元组间的相异度距离, 将性质相异的对象形成若干类, 最终形成一个聚类集. 显然地, 元组被分配到类簇中的原则是 与类簇 中质心 的相异度距离最近.

聚类算法流程如下；



（二）分类优化

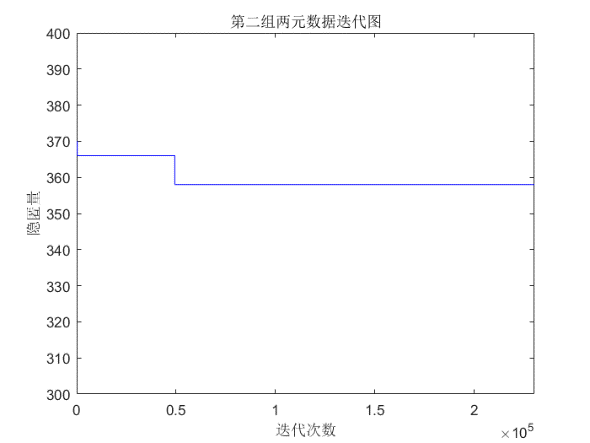
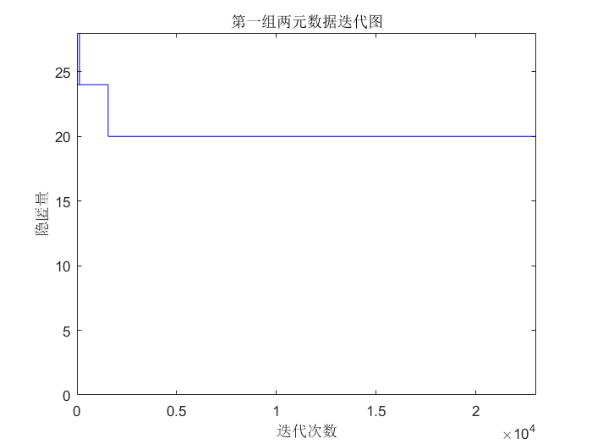
如果在聚类后，存在某些类的相异度距离过大，即需要隐匿过多的数据信息，则考虑将其重新进行归类，查找是否可以将其分在其他组中，以尽可能的减小隐匿信息量。

算法流程如下

**5.4结果展示**

**第一组分类优化前隐匿数据量为24，优化后20**

**第一组分类优化前隐匿数据量为364，优化后346**

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 第一组两元数据分组情况 | | | | |
| 组数 | 数据1 | 数据2 | 数据3 | 数据4 |
| 1 | 0 | 10 | 32 | 无 |
| 2 | 1 | 63 | 无 | 无 |
| 3 | 3 | 12 | 58 | 无 |
| 4 | 4 | 31 | 64 | 无 |
| 5 | 8 | 39 | 无 | 无 |
| 6 | 9 | 19 | 无 | 无 |
| 7 | 11 | 68 | 无 | 无 |
| 8 | 14 | 67 | 77 | 无 |

|  |  |  |
| --- | --- | --- |
| 第二组两元数据分组情况 | | |
| 组数 | 数据1 | 数据2 |
| 1 | 3 | 49 |
| 2 | 27 | 152 |
| 3 | 72 | 99 |
| 4 | 83 | 124 |
| 5 | 86 | 193 |
| 6 | 50 | 87 |
| 7 | 5 | 77 |
| 8 | 15 | 69 |

|  |  |
| --- | --- |
| 隐匿量 | 组数 |
| 2 | 1 |
| 1 | 8 |
| 0 | 27 |

1. **问题二模型建立与求解**

**6.1模型准备**

为了方便模型的建立，我们给出如下定义：

(1)等价类的定义

在附件数据中，属性取值相近的若干信息组成一个类，将这些信息隐匿掉部分数据，使其具有相同的属性值，便称该类为一个等价类，记为。假设一个等价类种包含t个数据，记为Ci。

(2) 等价数据的定义

经过数据隐藏组合后的某个等价类Ui，类中的数据经过数据隐藏之后都相同，这些信息统称为等价数据。

(3) 数据隐匿量的定义

记第i个等价类U数据隐匿量为Ti，指类中所有数据隐匿的总量，等价于数据属性不同的个数。记为X集合内的相同属性个数。

**6.2单目标优化模型建立**

**(一)目标函数**

依据题目要求，是所有信息都得到保护的前提下，确定合理的数据隐匿方式，使得隐匿数据的总量最小，即数据损失量最小，则优化目标为：

**(二) 决策变量**

考虑到隐匿量和约束条件都是关于信息属性AiBj的函数，因此将 AiBj 作为决策变量表示更加方便。

**(三) 约束条件**

在式(1)、(2)的约束下，增加约束条件:

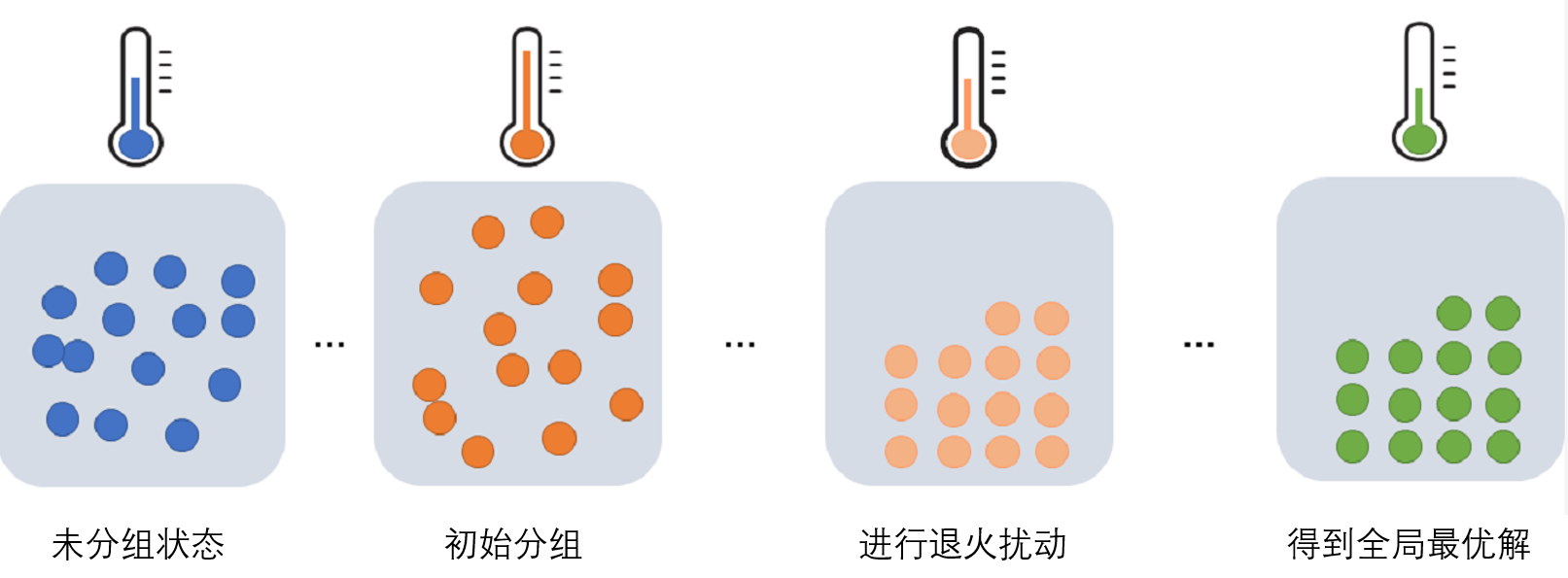
1. 每个组数据的量n要大于2，使得每个数据都能得到保护。

(2)每个数据隐匿的数据量一定大于等于不隐匿，小于等于全隐匿

(3)分类过程需要将信息分成多个等价类，使任意两个等价类相互没有共同信息。

**(四) 单目标优化模型确立**

**6.3模型求解**



（一）基于贪心算法对数据进行分组

与第二问相同，我们任然需要进行初始分组，由于多元数据的属性取值变为多元，并且数据量大幅增加，而原先的算法在复杂度和运算速度上已经不适合在进行第二问的计算，因此考虑寻找新的算法进行求解。

在进行初步分类是，使用贪心算法进行求解，它是寻找最优解问题的常用方法，这种方法模式一般将求解过程分成若干个步骤，但每个步骤都应用贪心原则，选取当前状态下最好/最优的选择（局部最有利的选择），并以此希望最后堆叠出的结果也是最好/最优的解。

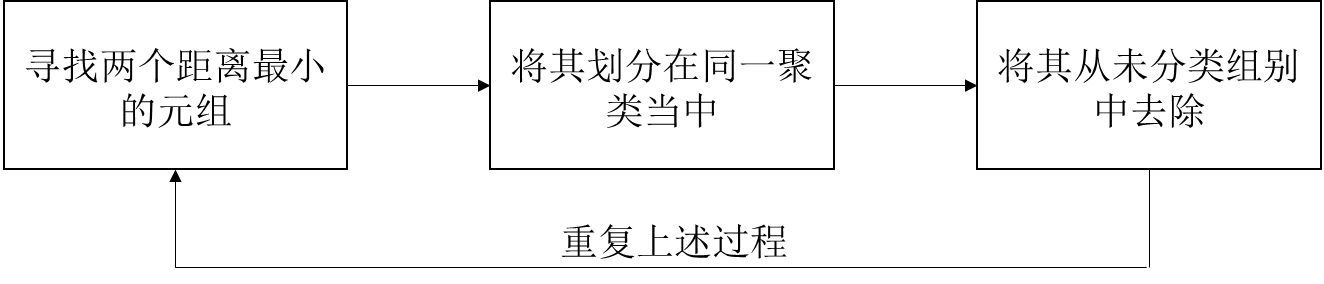
按照贪心算法的思想，按一定的规则进行遍历，步骤如下：

Step1：从第一个数开始遍历，首先**完全相同**的数据分为一类，如果没有，则先不分组。

Step2：再从未分组的数据开始遍历，将最先匹配到的**差异只有一个属性**的数据与其分为一类，如果没有，则先不分组。

Step3：与上一部相似，开始匹配差异更多属性的数据。

Step4：重复步骤3，直至所有数据都完成分组。



得到对数据的初步分组，并计算出此时的数据隐匿量

（二）利用退火算法对分组优化

使用贪心算法进行求解后，只能得到局部最优解，而第一问中的分类优化算法计算时间长，次数多，难以进行大量，多元数据的求解，因此采用模拟退火算法进行求解

模拟退火算法最早的思想是由N. Metropolis等人于1953年提出。它是基于迭代求解策略的一种随机寻优算法，其出发点是基于物理中固体物质的退火过程与一般组合优化问题之间的相似性。该算法的优点为求解速度快，可达到全局最优解。

模拟退火算法包括三函数两准则，即状态产生函数，状态接受函数，温度更新函数； 内循环终止和外循环终止准则[6] 。这些环节之间影响模拟退火算法的优化性能。

求解模拟退火算法描述如下：

（1）解空间。解空间S可表示为数据的所有分组形式的排列组合。

（2）目标函数。目标函数为数据隐匿总量。要求

（3）新解的产生。利用上一步迭代分好的组，每次选取三个组，每组内各挑选一个数据进行交换，得到新的分组。

选组要求为：1.每次选取三个不同的组；2.三个小组两两之间隐匿的个数不能完全相同。

(4) 代价函数差。假设原先的分组数据隐匿量为T1，新的分组数据隐匿量为T2。隐匿量差可表示为：

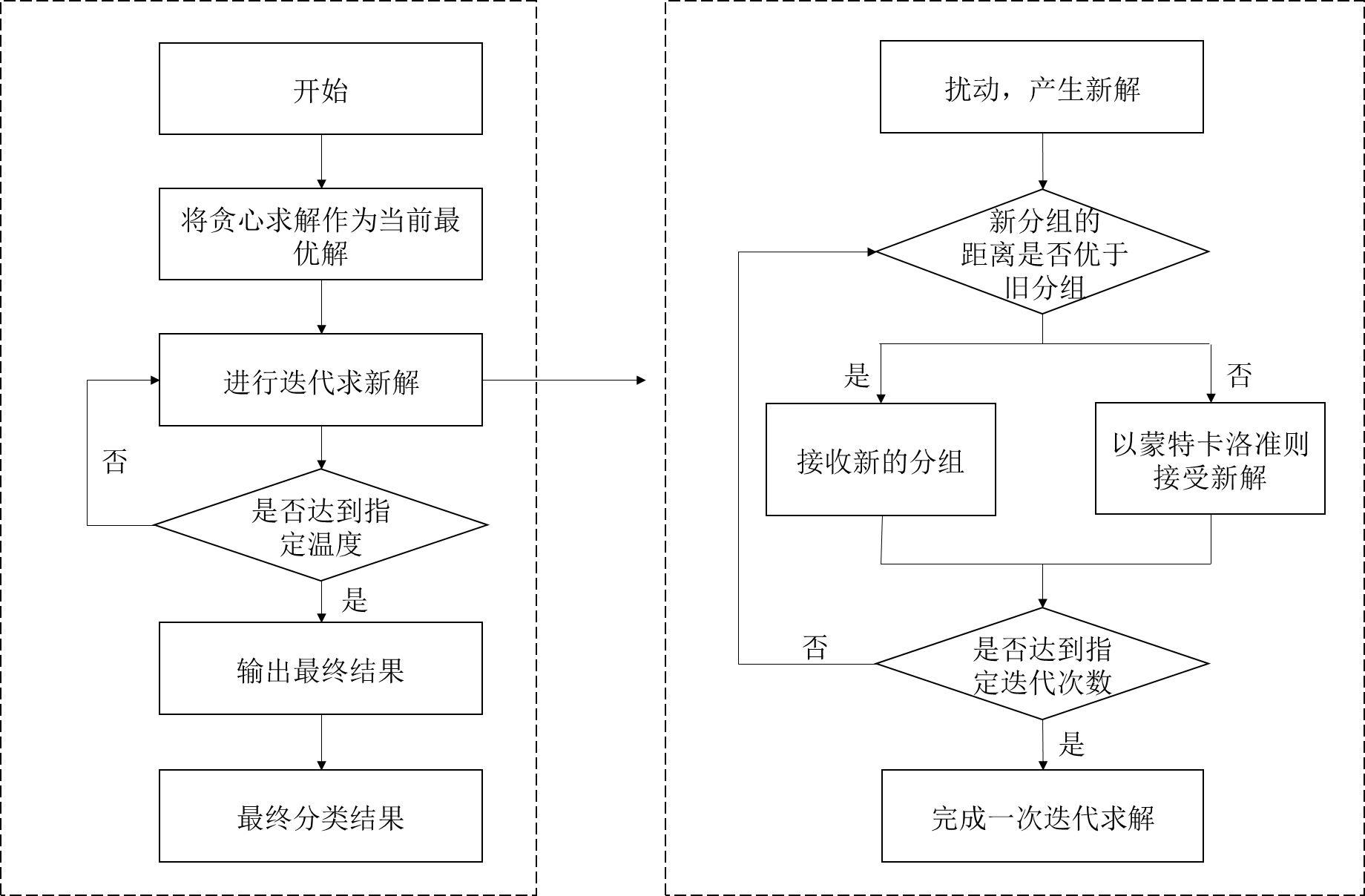
(5) 接受准则

如果，则接受新的分组；否则，以概率接受新的分组，即用计算机产生一个[0,1]区间上均匀分布的随机数rand，若则接受。

(6)降温。利用选定的降温系数α进行降温，取新的温度T为ΑT(这里T为上一步迭代的温度)，这里选α=0.999。

（7）结束条件。用选定的终止温度，判断退火过程是否结束。若T<e，则算发结束，输出当前状态。

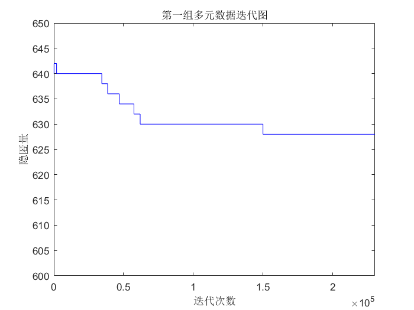
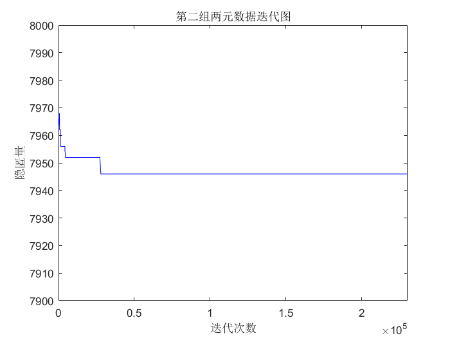
其算法的步骤具体为：



**6.4结果展示**

**第一组分类优化前隐匿数据量为640，优化后628**

**第一组分类优化前隐匿数据量为7964，优化后7946**



|  |  |  |  |
| --- | --- | --- | --- |
| 第一组多元数据分组情况 | | | |
| 组数 | 数据1 | 数据2 | 数据3 |
| 1 | 21 | 115 | 无 |
| 2 | 8 | 104 | 无 |
| 3 | 16 | 168 | 无 |
| 4 | 18 | 37 | 无 |
| 5 | 22 | 109 | 无 |
| 6 | 31 | 63 | 无 |
| 7 | 36 | 60 | 无 |
| 8 | 38 | 122 | 无 |

|  |  |  |
| --- | --- | --- |
| 第二多元数据分组情况 | | |
| 组数 | 数据1 | 数据2 |
| 1 | 6 | 492 |
| 2 | 20 | 259 |
| 3 | 36 | 715 |
| 4 | 39 | 575 |
| 5 | 76 | 337 |
| 6 | 79 | 165 |
| 7 | 81 | 232 |
| 8 | 127 | 197 |

1. **问题三模型建立与求解**

**7.1模型准备**

**K匿名技术**

通过K匿名技术，可以使攻击者无法知道某特定个人是否在公开的数据中；给定一个人，攻击者无法确认他是否有某项敏感属性；攻击者无法确认某条数据对应的是哪个人。这就要求使每个个体能够隐藏在一个至少包含k个个体的数据组中，这样的问题成为k匿名问题，就等价于P重保护问题。

将附件数据分成P个等价类，每个等价类C至少包含P个等价元组，附件数据满足 k-匿名，此时所有数据都得到了P重保护。

在对于匿名组的划分上，匿名化数据集中包含的匿名组越多，匿名组的规模约小，数据集的可用性就越高。因而，在P重保护时，最好的结果是将信息分为n/P组。

**7.2单目标优化模型建立**

**(一)目标函数**

依据题目要求，是所有信息都得到保护的前提下，确定合理的数据隐匿方式，使得隐匿数据的总量最小，即数据损失量最小，则优化目标为：

**(二) 决策变量**

考虑到隐匿量和约束条件都是关于信息属性AiBj的函数，因此将 AiBj 作为决策变量表示更加方便。

**(三) 约束条件**

在式(1)、(2)、 的约束下，增加约束条件:

（1）每个组数据的量n要大于P，使得每个数据都能得到P重保护。

**(四) 单目标优化模型确立**

**7.3模型求解**

（一）基于贪心算法对数据进行分组

与第二问模型求解类似，我们仍然用贪心算法进行分组，但是由于要做到P重保护，即每个组的数量至少要为P个，所以要对贪心策略进行改进。

按照贪心算法的思想，按改进的规则进行遍历，步骤如下：

Step1：从第一个数开始遍历，首先**完全相同**的数据分为一类，如果没有，则先不分组。

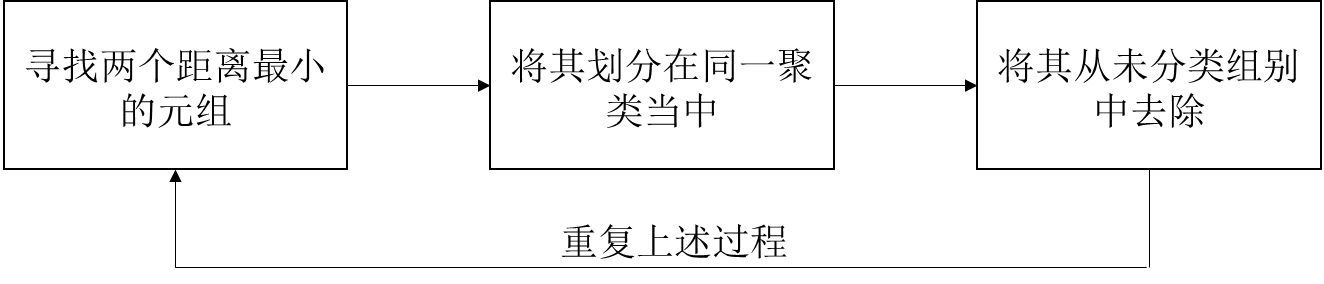
Step2：从分好的第一个类开始遍历，在未分类的数据中，寻找与类差异最小的数据放入类中的，直至类中的数据到达P个。

Step3：如果剩余未分组数据大于P个，再从未分组的数据开始遍历，将最先匹配到的**差异只有一个属性**的数据与其分为一类，如果没有，则先不分组；如果剩余数据未分组数据小于P个，寻找与其最相近的类并加入，并结束算法。

Step4：重复步骤2。

Step5：与上一部相似，开始匹配差异更多属性的数据。

Step6：重复步骤3~5，直至所有数据都完成分组。



得到对数据的初步分组，并计算出此时的数据隐匿量

（二）利用退火算法对分组优化

模拟退火算法包括三函数两准则，即状态产生函数，状态接受函数，温度更新函数； 内循环终止和外循环终止准则。这些环节之间影响模拟退火算法的优化性能。为了能使算法求解结果尽量靠近全局最优解，模拟退火算法可以接受较差解，当当前迭代产生的解比目前最优解更优时，算法以概率 1 接受该解作为新的最优解；而当前迭代产生18的解比目前最优解差时，算法则会exp(−∆E/T) 的概率接受该差解作为新的最优解其中T 为当前温度，∆T 为当前解所对应的目标函数值与最优解所对应的目标函数值的差值。可以看出，当初始温度足够高，降温足够慢时，每一温度下抽样足够长、最终温度趋近于 0 时，最终降低概率 1收敛到全局最优解。

**7.4结果展示**

1. **问题四模型建立与求解**
2. 模型改进
3. 模型评价

参考文献