# 问题二的建模与求解

## 假设与前提

## 问题重述

问题二在问题一的基础上，要求给出题目给出的某种疾病最有可能的一个或几个致病位点。在问题一的建模后，数据变为了便于进行数据分析的格式，为该问题做了事前的准备。

在该问题中，一共有1000个样例数据，其中每个样例数据有是否患有该疾病，以及9445个位点的编码信息。这在全基因组关联性分析中，是典型的病例对照研究（Case-Control Study）。

本题的问题在于找出疾病最有可能的致病位点。而业界目前采取的病例对照研究可以较好地得到问题的结果。

## 2.1 问题分析与建模

传统的重大疾病的研究模式，是以假说为导向，即首先假设某基因为致病候选基因，再对其进行深入研究，这样的方式无疑不适合本次题目的要求。而现代的、工业化的重大疾病的研究模式，是以数据为导向，对物种或个体进行基因组水平、大规模化研究，即在掌握基因组全局的情况下，发现影响疾病的基因位点。[[1]](#endnote-1)

目前，基于基因数据的分析有两个方向，一个是连锁分析，一个是关联分析。这两类方法在实验设计、研究对象、采用的技术手段以及对结果的解释等方面均不同。连锁分析是从分析但基因疾病发展起来的，它利用基因组上的微卫星作为遗传标记。在患病家系中观察是否存在某些遗传标记的传递模式显著偏离预期。

而关联分析通常根据某些间接线索、比如基因的位置和基因的功能，选定一个或几个候选基因，通过直接测序或基于PCR 的方法如等位基因特异性扩增等在病例和对照中比较候选基因的序列差异，然后来确定这些候选基因与患病状态或数量性状之间的关系。

通过分析已有的数据，我们发现，找出疾病最有可能的致病位点，可以重新表述为找到与疾病关联性最大的一个或几个位点。尽管在逻辑上，关联性并不能完全等同于因果性。但是因为该问题是从统计与学习的角度出发，从结果出发进行病例对照的研究，进而分析出最有可能的致病位点。因此疾病与位点之间的因果关系不可能通过已有的数据得出，因此我们将问题转变为对于疾病与位点之间的关联性的分析。这也是病例对照研究的基本思想与思路。

在关联分析与连锁分析之后，我们之所以选择了关联分析作为我们主要的分析方式，一方面因为关联分析能够分析出多个位点或基因相互影响的情况下的多位点致病因素。而且同样因为题目给定的数据不能支撑我们对其进行连锁分析。

因此，我们的重点放在了如何进行病例对照研究中疾病与位点之间的关联性分析。这其中主要涉及了位点的关联分析方法，以及机器学习分类的算法。

2.1.1 位点关联性分析

位点的关联分析方法，是指利用统计学等知识，对位点集合与疾病的关联性进行分析，进而得到与疾病最相关的位点集合。在位点关联性分析中，主要有两方面会影响其结果。一方面是关联性评价方法，另一个方面是位点集合搜索方法。

其中，关联性评价是评价疾病与位点的关联性的函数。位点与疾病的关联性越大，关联性评价就会越高，因此位点就更有可能是疾病的致病位点。目前，有很多被广泛使用的关联性评价方法。比如互信息以及卡方检验。其中卡方检验是一个统计学中最常用的假设检验。

2.1.1．1 卡方检验

卡方分布是指，如果有个相互独立的随机变量,,,它们都服从正态分布，即

则称统计量

服从自由度为的分布，即卡方分布，其中自由度为右式中所含有的独立变量的个数。

而检验是一种基于卡方分布的假设检验。如果要进行卡方检验，首先要根据实际问题提出一个假设：，它被称作零假设；，被称作备择假设。

随后给定一个显著水平，假设零假设成立，随后根据零假设以及实际问题的描述，构造一个符合卡方分布的检验统计量，并求出拒绝域。根据观察值决定接受零假设还是拒绝零假设。

若某一数据有两个不同的特征与，这两个特征之间的关联性求法如下。

设此两个特征的取值均为离散，特征的取值范围为,,，特征的取值范围为,,。假设此数据中，在两特征中取值的样本数为，在特征取值样本的样本数为，在特征取值样本的样本数为，即有如下公式：

设所有样本总数为。此数据的观察值列连表如表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | … |  |  |
|  |  | … |  |  |
| … | … | … | … | … |
|  |  | … |  |  |
|  |  | … |  |  |

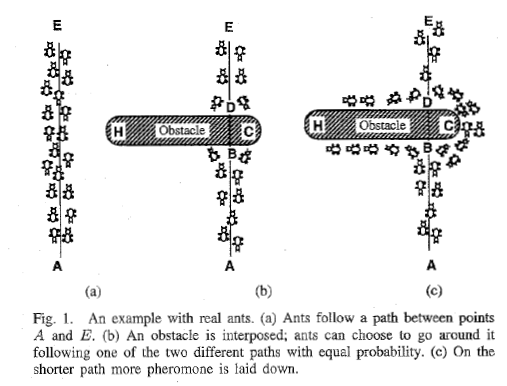
提出零假设：特征与特征无关联性。备择假设：特征与特征有关联性。如果为真，说明特征与特征相互独立，则可以进行值检验，值越大，零假设越可信，两个特征的关联性越小。

SNP 关联分析算法要求的是位点集与样本类标之间的关联性。在应用卡方检验时，只需要计算位点集基因型与样本集表现型之间的列连表，再根据列连表便可计算出值，并根据此来评价关联性大小。[[2]](#endnote-2)

卡方分布是常用的评价关联性的函数，在搜索位点集合的算法中，常用智能启发式算法来搜索合适的位点集合。其中的蚁群算法模型是一个实现简单，效率尚可的启发式搜索算法。

2.1.1.2 蚁群算法

蚁群算法是意大利学者Dorigo M等人于 1991 年提出的一种模拟进化算法，最初称之为蚁群系统(Ant Colony System, ACS)[[3]](#endnote-3)。其受到蚂蚁觅食时，总是能找到巢穴与食物地点之间的最短路径的启发。吸收了真实蚂蚁的行为特征，通过模拟蚁群寻找食物的过程来完成对问题的求解。该方法被广泛应用于数据分析、机器人协作、电力、通信、交通等领域。



在图中，有一群蚂蚁，假如是蚁巢，是食物源（反之亦然）。这群蚂蚁将沿着蚁巢和食物源之间的直线路径行驶。假如在和之间突然出现了一个障碍物（图（b）），那么，在点或点的蚂蚁将要做出决策，由于一开始路上没有前面蚂蚁留下的信息素（pheromone），蚂蚁朝着两个方向行进的概率是相等的。但是当有蚂蚁走过时，它将会在它行进的路上释放出信息素，并且这种信息素会以一定的速率散发。信息素是蚂蚁之间交流的工具之一。它后面的蚂蚁通过路上信息素的浓度，做出决策，而沿着短边的的路径上信息素将会越来越浓（图1（c）），从而吸引了越来越多的蚂蚁沿着这条路径行驶。

蚁群算法通过多次迭代，更新信息素的方式，模拟蚁群的决策方式，来搜索位点集合。

AntEpiSeeker模型

AntEpiSeeker[[4]](#endnote-4)模型是在2010年由Yupeng Wang在BioInformatics上提出的一个分析位点关联性的算法模型，AntEpiSeeker模型可以用来进行多个位点的致病因素关联分析。

AntEpiSeeker模型使用了卡方检验作为其评价函数。通过蚁群算法在可行解空间中搜索。AntEpiSeeker 有对边缘效应鲁棒性高的优点，但它受蚁群算法参数设置的影响较大，而这些参数往往靠经验设置。另外，AntEpiSeeker 还采用了两轮蚁群算法，对这两轮迭代过程，还要分别设置参数。

AntEpiSeeker需要指定所用蚂蚁数目，蒸发率，初始信息素，寻找位点集合中位点的数目，第一轮搜索位点数目，第二轮搜索位点数目，以及每轮迭代次数。在一轮迭代中，每只人工蚂蚁会按顺序选择一组位点，这组位点便是一个可行解。假设共有个位点。某只蚂蚁在决策时是否选择位点的概率公式类似于下式：

其中是第轮第个位点上的信息素。其中 ，其物理含义为每个蚂蚁在进行新的选择时只考虑信息素的影响，而不关心评价函数。这样是出于提高算法效率的考虑。如果每次蚂蚁在进行新的选择时都会考虑评价函数，则运算的时间复杂度太高，会影响算法的效率，对系统的CPU等运算单元有更高的要求。

在迭代结束后，每一个蚂蚁都会从个位点中选择出个位点，这个位点组成了一个集合,其中代表蚂蚁的编号。其更新的公式为：

其中是与疾病之间的。

为了提高蚁群算法在位点的搜索中的效果，AntEpiSeeker模型采取了两个阶段的搜索策略。在第一阶段中，AntEpiSeeker模型用蚁群算法搜索含有多于给定位点数目的位点子集，选出卡方统计量最大的前个子集作为候选子集，再选出信息素最大的前个位点作为候选位点，在候选子集与候选位点中，通过穷举搜索含有给定数目个位点的集合，评价函数为卡方统计量的值。候选子集的选取在探测边缘作用较小的因素中，起到了缩小解空间的作用。候选位点则是为了探测边缘作用比较大的位点。

重复上面两个阶段的搜索，AntEpiSeeker 采用了两轮搜索策略。在第一轮中，搜索包含位点数为的候选子集，第二轮中，搜索所含位点数为的候选子集。对两轮结果进行统计，选择卡方检验中值小于一个阈值的解为计算结果。

因为两轮搜索后的计算结果中可能包含有相同的位点，所以AntEpiSeeker对结果采用了最小化假阳率(Minimizing false positive)处理。将最小化假阳率前的结果用表示，初始化为两轮搜索后的结果，用表示最小化假阳率后的结果，初始化为空。对中的每一个元素执行以下操作：若中没有元素与含有同样的位点，则将移至中；若中有元素与含有至少一个相同的位点，则比较两者的值，将较大的保留在中，另外一个将会被丢弃。AntEpiSeeker模型计算的步骤如下所示。

## ../../../../Desktop/屏幕快照%202016-09-19%20下午2.59.13.

AntEpiSeeker模型相比于连锁分析等其他方法，具有很多优点。首先AntEpiSeeker模型充分考虑了上位作用。上位作用是指对等位基因受到另一对等位基因的制约，并随着后者不同前者的表型有所差异，后者即为上位基因（epistatic gene）。这一现象称为上位效应(epistasis)。其他诸如连锁分析等，都只是将位点作为独立的变量进行分析，而AntEpiSeeker模型则充分考虑了该作用的影响。其实际表现为两个位点各自与疾病的关联并不大，但是两个位点组合起来就有可能与疾病有很大关联，而AntEpiSeeker模型就可以发现这样的位点组合。

随机森林

随机森林，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，来决定这个样本应该属于哪一类。

在建立每一棵决策树的过程中，有两点需要注意，即采样与完全分裂。首先是两个随机采样的过程，随机森林对输入的数据要进行行、列的采样。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为个，那么采样的样本也为个。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现过拟合的现象。然后进行列采样，从个feature中，选择个。之后对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。一般很多的决策树算法都一个重要的步骤，那就是剪枝，但是随机森林并不需要剪枝的过程，由于之前的两个随机采样的过程保证了随机性，所以就算不剪枝，也不会出现过度拟合的情况。

随机森林的算法步骤如下所示。

## 问题求解

在本题中，我们使用了AntEpiSeeker模型分别从单个位点到多个位点进行计算，得到与该疾病关联最大的位点集合。并辅以随机森林和逻辑回归的结果进行综合考量，最后采取随机森林的方法训练分类器，进行结果的检验。

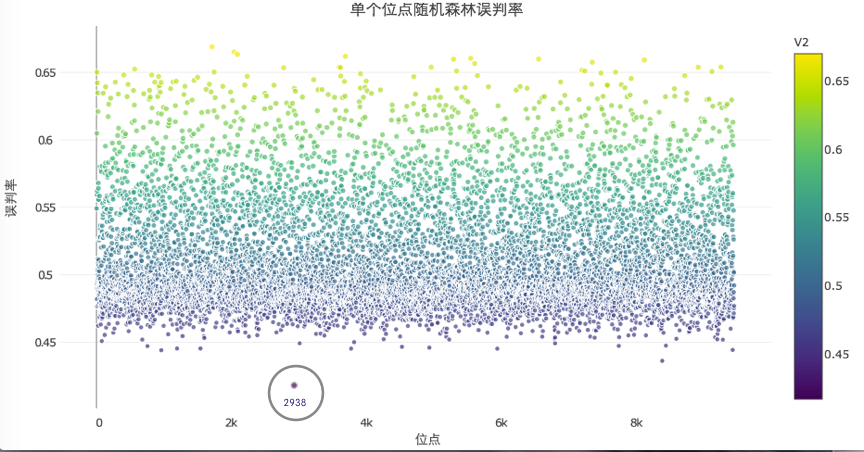
因为AntEpiSeeker模型需要给定一些初始参数，而这些参数对于结果会有影响，因此在使用由AntEpiSeeker论文作者提供的数据进行参数调优后，我们采取了如下的初始参数：

|  |  |  |
| --- | --- | --- |
| 变量名 | 值 | 备注 |
|  | 2000 | 蚂蚁数量 |
|  | 150 | 第一次迭代次数 |
|  | 300 | 第二次迭代次数 |
|  | 1 | 信息素沉积的权重 |
|  | 1000 | 第一次迭代的最大数 |
|  | 200 | 第一次迭代的最大位点 |
|  | 0.05 | 信息素挥发速度 |
|  | 100 | 初始信息素水平 |
|  | 6 | 大单倍型的尺寸 |
|  | 3 | 小单倍型的尺寸 |
|  | 1-N | 位点集合中位点个数 |
|  | 0.00001 | p值阈值 |

修改，将其从1开始，依次得到与疾病最有关联的n个位点集合。因为机器配置以及运行时间所限，在本题中只得到了为1和2时的两组解，解如下所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 位点集合 | 卡方 | P值 |
| 1 | 2938(rs2273298) | 28.8362 | 5.47405e-07 |
| 2 | 2938(rs2273298) 6077(rs11573253) | 54.3938 | 5.78981e-09 |

## 结果检验

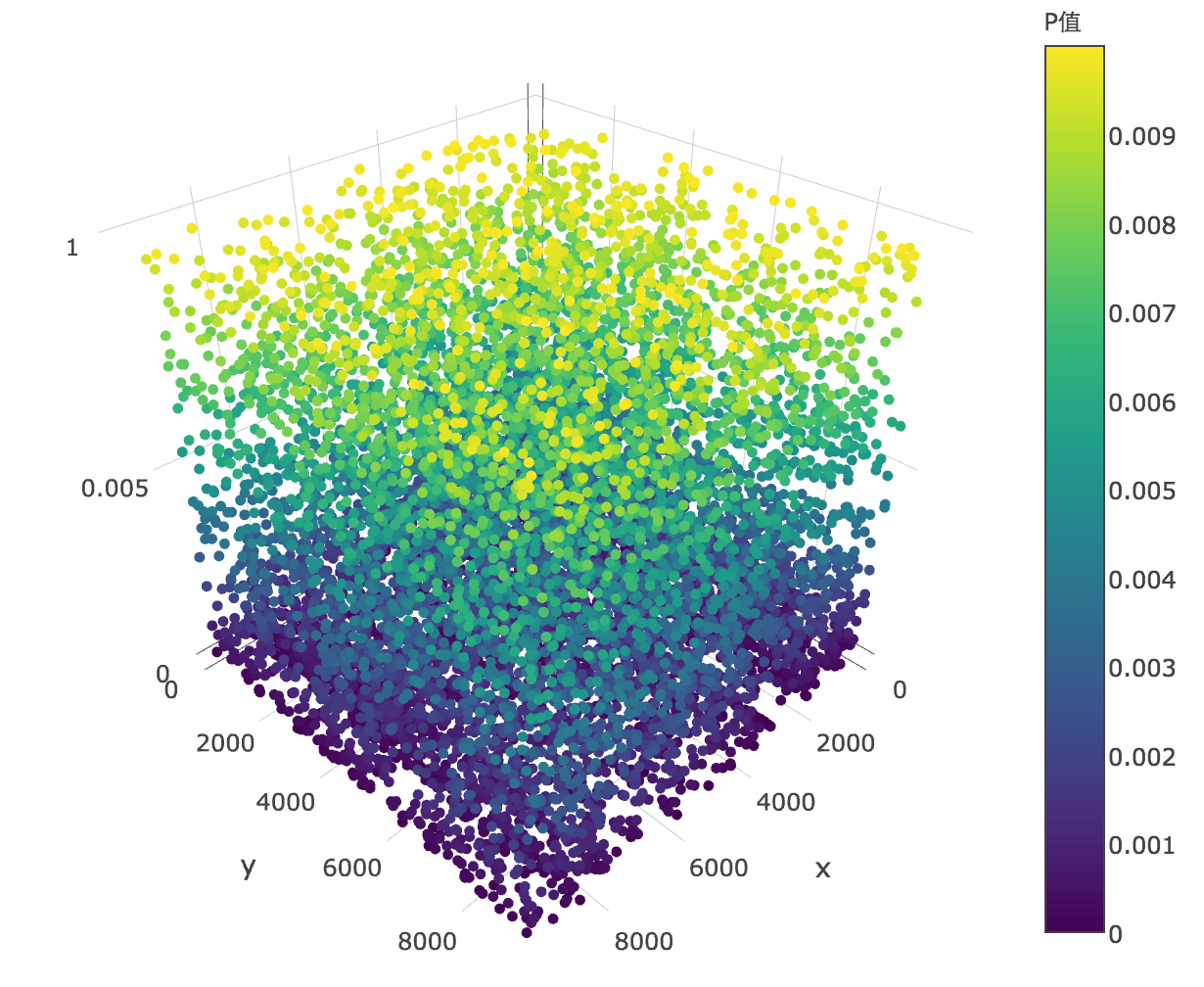
在结果检验时，我们首先将9445个位点数据分别作为变量，与样例的患病情况一起进行随机森林训练，并对其进行测试，其中误判率与位点之间的关系如图所示：

在图中，X轴为位点的ID，从0开始，一直到9445，而y轴是误判率。图中的点代表了由当前位点作为变量，所训练出的随机森林的分类器误判率。

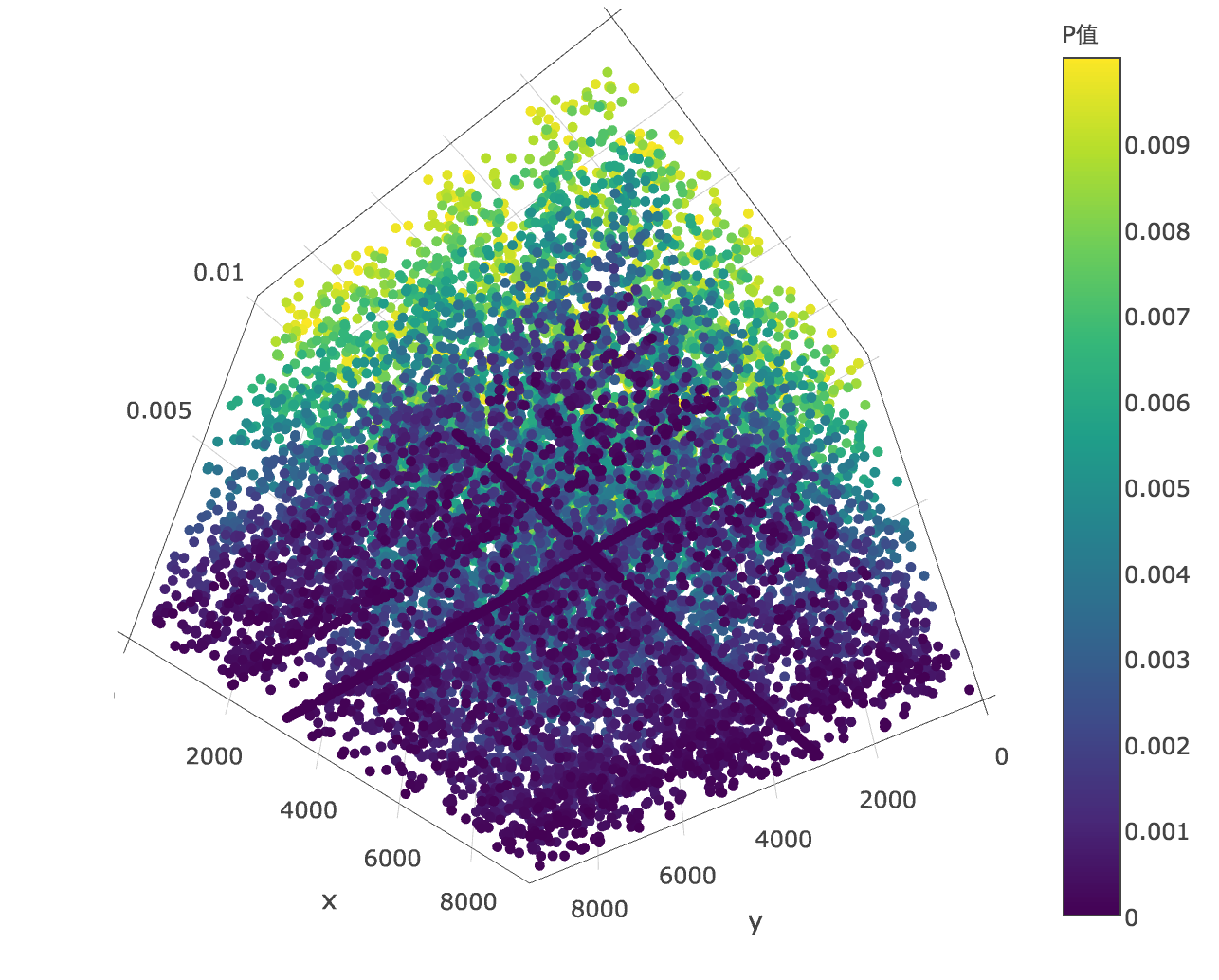
在其中可以看出，误判率最低的是由第2938个位点，即rs2273298训练出的分类器，其可以达到接近40%的误判率，在所有位点中是最低的。

因此可以看出，第2938个位点即rs2273298，是疾病最有可能的致病位点。

而在多位点的分析中，因为AntEpiSeeker模型考虑了位点组合的相互影响，因此相比于其他将位点作为独立变量来考虑的算法而言有着更好的效果。而因为位点两两组合的情况一共有713664200种可能，因此我们采取了随机采样的方式来进行验证。



上图是双位点的AntEpiSeeker运行结果，其中X，Y轴分别是两个位点的ID，而Z轴是值。由数据的随机性可以看出相邻的位点间并没有直接的关联。



在图中一处值得注意的地方在于，凡是有第2938个位点，即rs2273298出现的位点集合里，值都非常低，这也从另一方面佐证了该位点与疾病有最大的相关性。

# 附录

1. 叶道军. SNP疾病模型的性质研究及检出性能的比较[D].西安电子科技大学,2011. [↑](#endnote-ref-1)
2. 高原英. 基于稳定性的多因素SNP关联分析算法[D].西安电子科技大学,2012. [↑](#endnote-ref-2)
3. 蚁群算法 [↑](#endnote-ref-3)
4. Wang Y, Liu X, Rekaya R. AntEpiSeeker2. 0: extending epistasis detection to epistasis-associated pathway inference using ant colony optimization[J]. 2012. [↑](#endnote-ref-4)