具有遗传性疾病和性状的全基因组关联性分析

摘要

随着人类所获取的海量生物dna信息量的不断增长，针对各类复杂的遗传性疾病的遗传机理及性状与相关的关联基因以及关联的位点的研究成为遗传学领域的研究热点之一，如何中在大规模dna序列信息中找到关联位点及对应基因带来了很大挑战。本文中，我们通过贝叶斯推断模型找到致病位点与疾病关联关系。

对于问题一，由于一个基因位点只存在三种不同的编码方式，并且呈离散状态，因此用0,1,2对三种编码方式进行重新编码。这方便使用程序进行统计分析，比较有利于贝叶斯推断模型的实现。

针对问题二，首先，我们分析了单个位点与疾病关联性，但是估算发现单个位点无法解释该疾病的遗传性。由于生物个体的复杂性，遗传性疾病及性状更可能受到多个位点的影响。基于此，我们通过贝叶斯推断方法检验每个位点或者位点集合是否与疾病显著相关。

针对问题三，通过贝叶斯推断模型，产生基因位点与疾病A的后验概率分布，取概率较大的一些位点组成集合*S*。假设基因*1-N*，包含位点集合为*-*，*S*与的交集包含位点个数越多，包含的位点后验概率越大，则该基因与疾病A的关联性更强。

针对问题四，对十种性状，建立贝叶斯推断模型，分别计算出在该性状下，样本数据中基因位点的后验概率分布。设在中单边影响或关联影响较大的位点集合，则这些位点集合的交集中的位点是与10个性状有关联的位点。

**关键字**： 遗传统计学 全基因组关联性分析 位点

1. 问题重述

1.1引言

疾病在很大程度上影响了人类健康，由于特定位置的单个核苷酸发生变异引起dna的多态性，我们称之为位点也可称为SNP第三代遗传标记，针对某些疾病，单个独立位点有可能代表该疾病发病机理中的主要遗传作用因素，而往往大部分疾病由多个位点联合作用产生主要影响，通常多个位点对疾病及其性状产生起关键影响的相互关联作用，即一个基因对疾病的影响很可能受到其它一个或多个基因的影响。

1.2问题的提出

针对遗传性疾病的全基因组关联性问题的研究，本文依次提出如下问题：

1. 为了更方便对基因数据进行处理分析，选取合适的编码方式将数据集中每个位点的碱基（A,T,C,G）编码方式转化为数值编码方式。
2. 对于给定的样本数据以及可能致病的染色体片段上的位点上的编码信息，找到某种疾病与一个或者多个致病位点的关联关系。
3. 由给出基因与位点关系的数据集，可将对应基因看作若干位点组成的集合，由问题二可知某种疾病可能与一个或多个位点存在关联关系，我们可以找到遗传疾病与基因的关联性，可以由基因中包含的所有位点或者部分位点表现出来，找到与疾病关联的一个或多个基因。
4. 通常，实际研究过程中，相关科研人员会将有关联的性状或疾病看作一个整体，以便找到与它们相关的位点或者基因。根据所给出的样本与性状相关性的信息以及所有位点的编码信息，找到给出10个性状有关联的位点。
5. 模型假设
6. 忽略实际加工误差对设计的影响；
7. 木条与圆桌面之间的交接处缝隙较小，可忽略；
8. 钢筋强度足够大，不弯曲；
9. 假设地面平整。
10. 符号说明

|  |  |
| --- | --- |
| 符号 | 意义 |
|  | 第n根木条开槽长度（cm） | |
|  |  | |

1. 问题分析

4.1问题一分析

题目要求给出便于数据分析的碱基数值编码方式，然而从材料中可知各个位点虽然碱基的组合不同，但也都分别只有三种不同编码。由于每个位点相当于样本的一个离散性特征,所以我们可以使用离散值来表示位点的编码,比如0,1,2,并且所有的位点的编码都可以使用一套离散值来表示。

4.2问题二分析

题目给出1000个样本在某条有可能致病的染色体片段上的9445个位点的编码信息和样本患有遗传疾病A的信息，要求找出某种疾病最有可能的一个或几个致病位点。这是一个相关性分析问题，需要分别找出单独对疾病关联的位点和交互作用与疾病关联的位点，即需要对染色体中的致病位点或其组合进行定位。我们发现样本的维度相当高，如果找出所有单独对疾病关联的位点和交互作用与疾病关联的位点，则会面临计算复杂度高，难于对全部可能致病位点或其组合进行计算等复杂问题。对于高维度的问题，通常可以先使用降维的方法减少维度，然后在使用合适方法找出位点与疾病之间的关系。其中一个特别有名的方法为多因素维数约减方法（Multifactor Dimensionality Reduction, MDR）。MDR 方法试图识别出那些通过互关联作用而非单个位点所具有的边缘效应作用对疾病等模型输出变量产生影响的变异位点集合。MDR 通过将高维的多位点模型约减为一维模型从而减少了需要进行二分类的预测变量的维数。MDR面临一个重要的问题即它不可被用于分析、学习预测变量数目过多的数据集。MDR 方法只适合应用在最多含有几百个变异位点的基因组数据集上。然而，材料提供的是一个含有9445个位点的样本集，所以MDR 方法不适用。递归分区方法（Recursive Partitioning Approaches, RPA）生成树的过程中通过将其上的每个路径与某些预测变量的

特定值组合相对应进而考虑这些预测变量之间所可能具有的互关联作用及这种

互关联作用所可能具有的对输出变量的影响。递归分区方法的一个缺陷即在于它在第一阶段条件依赖于各个变量的边缘效应并在其后的各阶段条件依赖于先前已经被选中的变量的主效应。贝叶斯模型选取方法提供了一种新的可选择、学习出那些对所研究的如疾病等生理型特征起某种预测作用的遗传变异位点及它们间的相互关联作用的方法。在贝叶斯方法中对未知回归参数的先验分布及模型

中的维度参数进行先验指定。基于已经观测到的数据，这些参数的一个后验概率分布可以通过一个马尔可夫链蒙特卡罗（Markov Chain Monte Carlo, MCMC）仿真方法计算得出。本题我们采用贝叶斯模型选取方法，并用MCMC计算后验概率。

4.3问题三分析

4.4问题四分析

1. 模型建立与解决

5.1 问题一的解决

设定出现频率最高的基因为T,通常称为主要等位基因，则t可以表示出现频率较低的另外一个等位基因，可知，对于一个某个染色体的碱基对信息，某个特定的碱基对位置，可能会出现TT,TC,CC三种基因型状态，同理，对于染色体其它位置上面的碱基的组合可能不同，但是也只有三种状态，据此，对应的数值编码我们设定为0,1与2这三个离散数值分别表示这三种状态。

对于样本中所给出疾病特征，为离散型数值特征，设定1表示患病状态，0表示健康状态，据此，对于给出表1中基因信息，比如位点rs100015,我们分别用0,1,2表示三个不同的编码TT、TC，CC，其他位点的编码使用相似的方式表示，可以将给定样本的基因组信息处理为如下表2所示。

表1. 在对每个样本采集完全基因组信息后，一般有以下的数据信息

(以6个样本为例，其中3个病人，3个健康者)：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 样本编号 | 样本健康状况 | 染色体片段位点名称和位点等位基因信息 | | | |
| rs100015 | rs56341 | ... | rs21132 |
| 1 | 1 | TT | CA | ... | GT |
| 2 | 0 | TT | CC | ... | GG |
| 3 | 1 | TC | CC | ... | GG |
| 4 | 1 | TC | CA | ... | GG |
| 5 | 0 | CC | CC | ... | GG |
| 6 | 0 | TT | CC | ... | GG |

*注：位点名称通常以rs开头。*

表2. 全基因组信息中数值编码状态信息

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 样本编号 | 样本健康状况 | 染色体片段位点名称和位点等位基因信息 | | | |
| rs100015 | rs56341 | ... | rs21132 |
| 1 | 1 | 0 | 1 | ... | 1 |
| 2 | 0 | 0 | 0 | ... | 0 |
| 3 | 1 | 1 | 0 | ... | 0 |
| 4 | 1 | 1 | 1 | ... | 0 |
| 5 | 0 | 2 | 0 | ... | 0 |
| 6 | 0 | 0 | 0 | ... | 0 |

5.1 问题二的模型建立和解决

我们的算法通过马尔科夫链蒙特卡洛方法进行贝叶斯上位性推断。算法首先通过贝叶斯划分模型将位点空间分为三组：组 0 包含那些与所研究疾病无关联作用的位点集合；组 1 包含那些只对所研究疾病呈现边际效应作用的位点集合；而组 2 则包含那些相互关联起来以对疾病产生影响的位点集合。然后假设三类中基因型对应的病例数服从多项式分布，其中参数服从狄利克雷分布，根据贝叶斯推断可以求得三种分类的后验概率表达式，再运用马尔科夫链蒙特卡洛方法可以求得三种分类的后验概率。

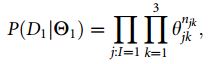
**定义：**

假设有患者和个正常人被标记了个位点。令病人基因为，其中表示病人令常人基因为,其中表示正常人。这个个位点被划分成三组：组 0 包含那些与所研究疾病无关联作用的位点集合；组 1 包含那些只对所研究疾病呈现边际效应作用的位点集合；而组 2 则包含那些相互关联起来以对疾病产生影响的位点集合。令分别用0,1和2来标识位点所属组别。我们的目标是推理出与疾病有关联的位点的集合，即集合{}。令分别表示每组中位点的数量（），并且令分别表示在组0，1,2中的病人基因.

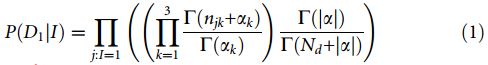
**贝叶斯位点划分模型:**

当与正常人基因相比时，在相关联位点上的患病基因应该表现出不同的分布。具体来说，我们描述似然模型假设位点独立在对照人群。

令是组1中每个等位基因位点的基因型频率，我们把的似然写为：

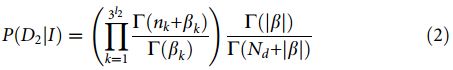


其中是组1中每个位点的基因计数。假设的先验概率分布为，其中，我们整合出，获得边缘概率：

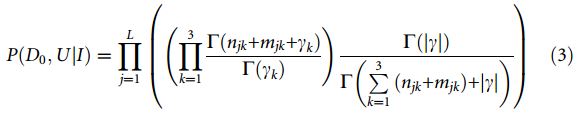


这里符号表示在中所有元素之和。

组2中位点通过相互作用影响疾病风险。因此，在这个组的个位点的每个基因型组合都是一个候选关联。共存在个频率为可能的基因型组合。令是基因组合在中的数量。假设的先验概率分布是，其中，我们整合出，因此



剩下的数据包含的位点服从和正常人相同的分布。令表示个位点在正常人的基因频率，令和是个体中分别在和中位点基因型的数量。假设的先验概率，我们整合出，获得



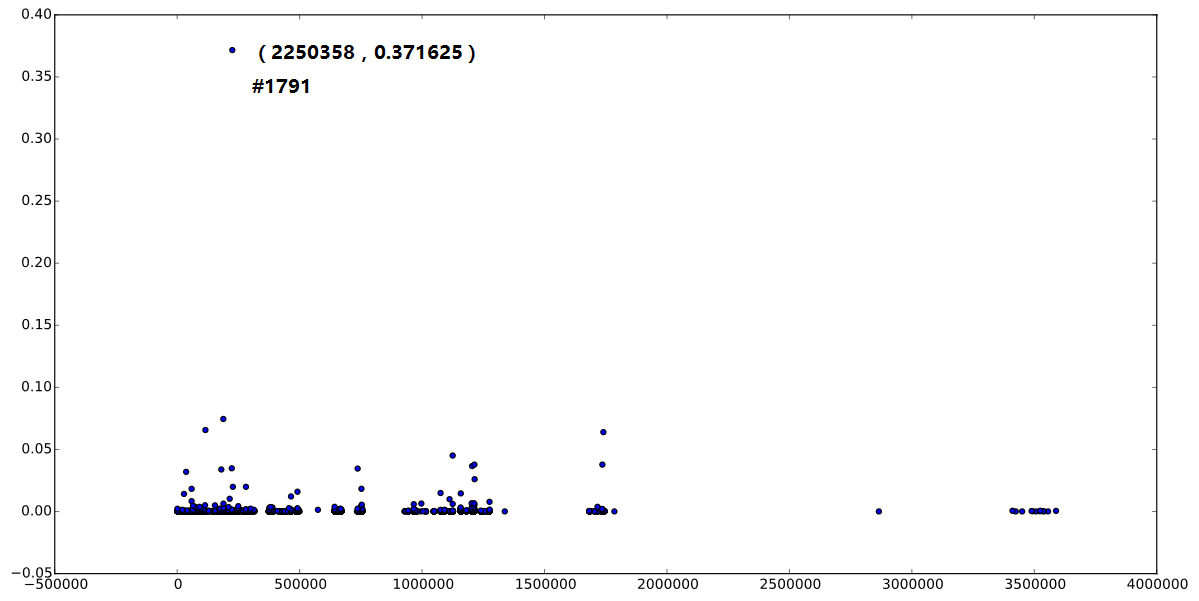
结合公式（1）、（2）和（3），我们获得后验分布为：

D:\其他\比赛\2016年第十三届全国研究生数学建模竞赛\全国竞赛\2016试题\B\论文\(4)-1.JPG

注意，决定了的结构，我们令，这个可以修改来反映出关于每个与疾病关联的位点的先验知识。默认的，我们设置。

**MCMC采样：**

我们的目标是从分布（4）中找到划分器。我们按照先验概率初始化，使用Metropolis-Hastings (MH)算法来更新。使用两类方法：（i）随机改变一个位点群成员（ii）随机交换组0，1,2中的两个位点。按照MH率接受移动，这个比率是一个函数。输出是位点和相互作用的后验分布。为了提高采样效率，我们首先给组2中位点数量设置一个较低的边界值，然后不断减少这个边界值直到0。这强迫算法去搜索高阶关联空间。



横坐标为位点标记rs2250358中2250358，纵坐标为后验概率

5.1 问题三的模型建立和解决

通过问题二模型计算出，属于组1集合位点的后验概率（即单个位点具有边缘效应的后验概率），进行排序之后（如图2），取后验概率大于0.019905（位点rs2807345）的位点组成集合*S*=（'rs2273298','rs12145450','rs364642','rs1802353'， 'rs7368252','rs2229579','rs12042240','rs12136961''rs17361679','rs11249209','rs17401924','rs1152984''rs1883567','rs2250358'）

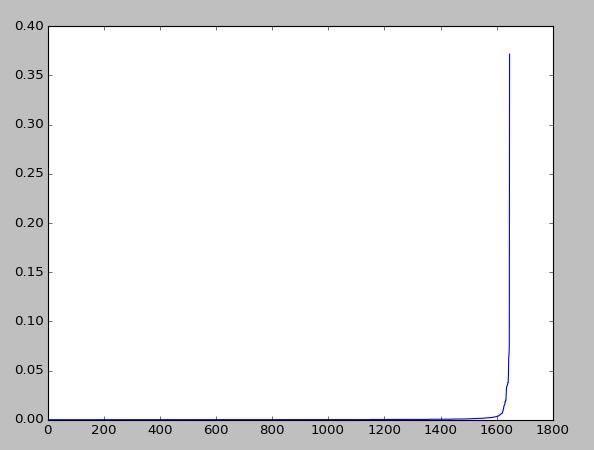


图2 单个位点边缘效应后验概率排序

将300个基因位点集合与*S*求交集，记交集个数为，后验概率和为，设评判标准,较大的基因与疾病A的相关性强。通过findGene.py计算出基因编号为[44，61，101，102，188，230，240，297 ]相关性较强。

5.4 问题四的解决

分别对十种性状，建立贝叶斯推断模型，计算出在该性状下，样本数据中基因位点的后验概率。使用程序2(findRS.py)选取设在中单边影响或关联影响较大的位点100个组成集合，对这些集合取交集得到rs=set(['rs7513908', 'rs10909989', 'rs16824089', 'rs10157835', 'rs6663452', 'rs951805', 'rs17376524', 'rs10927473', 'rs12746773', 'rs4949516'])

1. 模型评价及改进

6.1模型评价

6.2模型改进

七、参考文献

[1] 权晟, 张学军, 全基因组关联研究的深度分析策略 [J]. 遗传, 2011. 33(2): 100-8.

[2] STADLER Z. K., THOM P., ROBSON M. E., et al., Genome-wide association studies of cancer [J]. Journal of Clinical Oncology, 2010. 28(27): 4255-67.

[3] JOSTINS L., BARRETT J. C., Genetic risk prediction in complex disease [J]. Human molecular genetics, 2011. 20(2): 182-8.

[4] EVANS D. M., VISSCHER P. M., WRAY N. R., Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk [J]. Human molecular genetics, 2009. 18(18): 3525-31.

[5] CHORK N J., FALLIN D., LANCHBURY J. S., Single nucleotide polymorphisms and the future of genetic epidemiology [J]. Clinical genetics, 2001. 58(4): 250-64.

[6] MITRA S., ACHARYA T., Data Mining: multimedia, soft computing, and bioinformatics [M]. Wiley-Interscience, 2003. 44(2): 637-89.

[7] FREITAS A. A., Understanding the crucial role of attribute interaction in data mining [J]. Artificial Intelligence Review, 2001. 16(3): 177-99.

[8] MOORE J. H., Computational analysis of gene-gene interactions using multifactor dimensionality reduction [J]. Expert review of molecular diagnostics, 2004. 4(6): 795-803.

[9] CHO Y., RITCHIE M., MOORE J., et al., Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus [J]. Diabetologia, 2004. 47(3): 549-54.

[10] 尚军亮. 全基因组单核苷酸多态性交互作用研究[D]. 西安电子科技大学,

2013.

[11] 胡晓菡. CUDA 平台下的复杂疾病全基因组基因-基因相互作用研究[D]. 上海交通大学, 2010.

[12] 袁敏. 关联分析中的统计方法研究[D]. 中国科学技术大学, 2009.

[13] 韩建文,张学军, 全基因组关联研究现状[J]. 遗传, 2010, 33(1): 25-35.

### 源程序引索

#### 问题二源程序

#### 问题三源程序

#### 问题四源程序