参赛密码_____

（由组委会填写）

“华为杯”第十三届全国研究生数学建模竞赛

学校	上海对外经贸大学
参赛队号	10273004
队员姓名	1. 詹德勇
	2. 段 伟
	3. 谢灵艳

参赛密码_____

（由组委会填写）

题目：　具有遗传性疾病和性状的遗传位点分析（B 题）

摘要：

本文根据 1000 个样本的遗传病和性状信息，以及 9445 个位点上的遗传信息，利用多种统计分析、优化软件，进行大批量的数据处理和数据挖掘，主要完成了以下几个方面的工作：

对于问题 1，传统的基于碱基的编码方式是以碱基为基本单位，对 C,T,A,G 四个碱基对应编码 0(00),1(01),2(10),3(11)四个数字。由于本文所有数据最小分析单元为位点，每个位点的观测对应两个碱基，因此我们以位点为基本单位，对等位基因 TT,TC,CC 编码 0(00),1(01),2(10)，相比传统的的编码方式，基于位点的碱基对数值编码方式能有效减少内存，便于数据分析。

对于问题 2，首先对位点数据进行预处理，考虑到基因的遗传必须满足基本的传统统计特征,因此,在一定的显著性水平阈值下,基于最小等位基因频率和 Hardy-Weinberg 平衡定律对题目所给的全基因组进行分析，从而剔除了 97 个不满足条件的位点。在预处理之后，为寻找与疾病 A 可能相关的位点，采用列联表分析方法，通过卡方检验和 Fishers 精确检验，计算每个 SNP 等位基因与疾病 A 的统计量，统计检验显著的位点即为与疾病 A 相关联的致病位点。为进一步筛选出与疾病 A 关联性较强的位点，我们引入了在信用评分、营销响应预测中常用的变量选择方法——信息值 IV，通过计算每个位点关于疾病 A 的 IV 值，IV 值越大则影响度越高。综合对比两种方法所得的致病位点，并通过具体分析排除信息值为无穷大的特殊位点，最终，我们认为，与疾病 A 最有可能关联的致病位点是 rs2273298。

对于问题 3，考虑基因与疾病之间的关联性，实际上是个分类预测问题，即当个体拥有某基因时，判断其进入健康组或患病组的概率，因此，我们采用决策树的方法求解。考虑到若直接对 300 个基因分别做决策树，判断基因与疾病的关联性，计算量过大，故我们采取降维的思想。由于 300 个基因中位点对应唯一基因，因此我们先以位点为研究对象，基于问题 2 的统计检验结果和相关挑选规则，选出满足筛选条件的 10 个位点，也即 10 个基因。在此基础上，我们对每个基因做决策树，以混淆矩阵和 ROC 曲线作为评价标准，比较各决策树的优劣程度。最终发现 gene102 和 gene55 的准确率最高，分别为 63.8%和 61.6%。因此我们认为这两个基因与疾病 A 的关联性最强。

对于问题 4，由于观测样本包含 10 个性状的不同观测的组合，其理论的性状类别有 2^{10} 即 1024 种，即使考虑 1000 个样本的实际性状表现不超过 1000，显然维度过高，因此先对样本的性状表现做聚类分析，提出出有代表性的综合性状。我们考虑基于类平均法的距离公式，对样本进行系统聚类，以 R2 和偏 R2 为参考指标，最终确认了 7,16,20,50,100 等不同的聚类数。在确认聚类数后，利用 k-均值聚类法，计算出基于

每一个聚类数所得的综合性指标，类似于问题 2，再将综合指标关于位点做统计检验，找出与综合指标有显著关联的 10 个位点。进而，为了考察所选位点的正确性，我们进一步对所选的 10 个位点和原始 1000 个样本的 10 个性状做关联性分析，统计结果显示所挑选的 10 个位点中的 8 个位点，都与题目所给的 10 个初始性状中的一个或多个性状高度相关。因此，我们认为与 10 个关联性状所有表现出的综合性状相关的位点有 rs12746773，rs4584380，rs11249201，rs12139270，rs2075972，rs1985278，rs6603797，rs10917268。

关键词：位点识别　全基因组关联分析　IV 值　决策树　聚类分析

目录

一. 问题重述

1.1. 问题背景

尽管随着医学技术的发展，一些过去严重威胁人类健康的疾病得以控制，但遗传病却依然是一个棘手的难题，发展到如今已成为比较突出的问题。据统计，现今已有 6000 种左右的遗传病被人们所认识，患者约占总人数的 10%。此外，遗传病种类仍以每年新增加 100 种的速度上升，因此遗传病可以说是一种多发病、常见病，日益成为威胁人类健康的重要疾病之一。

各种遗传病的发生，都源于基因突变。基因突变是指一些特定位置的单个核苷酸发生变异，我们称之为位点(SNPs)，使得基因在分子结构上发生改变，这种 DNA 分子结构的改变导致基因功能的异常，从而导致遗传病。根据基因突变的个数，又可将遗传病分为单基因遗传病和多基因遗传病，单基因遗传病指由单个致病基因引起的遗传病，多基因遗传病则是由多个基因的累加效应引起的。

另一方面，人体的许多遗传疾病和性状是有关联的。遗传性状是指由基因控制的生物性状，是可以通过 DNA 的形式由亲代传给子代的一切形态特征、生理特性、行为本能等。遗传病与性状相关联的原因在于，遗传病源于细胞内遗传物质（基因）的改变，基因控制着细胞中的蛋白质合成，从而控制着生物的各种遗传性状。因此，科研人员往往把相关的性状或疾病放在一起研究。

为了解遗传病的作用机理，我们需要识别致病 SNP 位点，即定位与疾病或性状相关联的位点在染色体或基因中的位置，找到致病基因或易感基因，从而帮助研究人员了解疾病，对致病位点加以干预，防止一些遗传病的发生，为人类的健康做出贡献。因此，如何准确、有效的定位 SPN，对于解决遗传疾病有重要意义。

1.2. 问题提出

SNP 数量多，分布广泛，在整个人类基因组中约有 1000 万左右的 SNPs，只有探索影响疾病发生、发展的遗传因素，有效准确的识别致病 SNP 位点，才能对遗传病进行防范治理，否则容易产生新的基因突变或疾病。因此，本文量力而行，尝试用多种方法寻求最准确的致病 SNP 位点。本题要求根据患有遗传疾病 A 的不同样本信息（单个位点/多个位点、疾病/性状），找出该疾病最有可能的致病位点或基因。为此，需要做好以下几项工作：

第一，对位点做好数值编码工作。在人类基因组中 SNP 一般为 2 个等位基因，如 AG，即所谓的二态性，将碱基(A,T,C,G)字符型的编码方式转化成数值编码方式，既方便计算机存储，也便于进行后续数据分析。

第二，判断位点与疾病的关联性，识别致病位点。基于上述的位点数值编码，根据 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，找出该疾病最有可能的致病位点。

第三，判断单体型和疾病的关联性，识别单体型中的致病基因。根据 1000 个样本在 300 个基因上位点的编码信息和样本患有遗传疾病 A 的信息，找出该疾病最有可能的致病基因。

第四，基于性状识别单基因遗传疾病中的致病位点。由于遗传疾病和性状相互关联，

基于该遗传病的 10 个性状，利用 1000 个样本的相关联性状的信息及其 9445 个位点的编码信息，可以找出与 10 个性状有关联的位点。

二．问题分析

致病位点的定位问题一直以来是生物学界关注的重点。同时，遗传病问题也与我们的生活息息相关，所以，准确的识别位点也具有很强的现实意义。对疾病 A 患者的信息及其相关性状的数据，位点的编码数据和包含这些位点的基因数据进行统计分析和建模对于确定致病位点有重要的参考的意义。

问题 1 要求对每个位点进行编码，考虑到每个位点由两个字符型碱基构成，而字符型数据占用内存空间，影响程序运行速度，因此需将字符型编码转换为数值型编码方式。

问题 2 给出 1000 个样本在染色体片段上的 9445 个位点信息和样本患病信息，要求找出该疾病最有可能的致病位点。在进行数据建模前，考虑到对 SNP 位点的质量控制，我们对初始的 9445 个单核苷酸多态性特征进行预处理，去掉最小等位基因频率小于 0.01 和不符合哈代温伯格定律的 SNP。在此基础上，我们首先可以对数据进行统计分析，即全基因组关联性分析，比较患病组和对照组中每个 SNP 等位基因频率差别，通过卡方检验和 Fishers 精确检验，发现易感基因位点。但是上述过程是从“变量之间是否独立”角度筛选自变量，对满足原假设的自变量无法进一步筛选，即所得易感基因较多同时无法在结果中比较。因此，我们从“变量的预测能力”角度出发，对每个位点赋予一个唯一的证据权重（WOE）值，并以此为基础计算信息值（IV 值），量化预测能力的大小，选择预测能力较大的位点。由于变量的预测能力也通过变量之间的相关性体现，因此我们可以预测两种方法所得结果可能存在一定的一致性。

问题 3 要求根据 1000 个样本在 300 个基因上位点的编码信息，判断基因和疾病的关联性，这等价于判断已知基因进入健康组或患病组的概率，因此该问题实际上是分类预测问题。同时考虑到基因中位点间的相关性，我们使用决策树的方法解决该问题。若直接对 300 个基因分别做决策树，评估分类模型的正确性，模型计算时间冗长，效率过低。因此，我们可以基于问题 2 的结果，根据一定的筛选标准，从 9000 多位点中选出通过检验的 10 个位点，再对应 10 个位点找出其所在的基因，作为备选基因。对上述备选基因做决策树，挑选出分类准确率最高的决策树，对应基因即为与疾病关联度最高的基因。

问题 4 要求基于 10 个相关联性状和位点信息，寻求与性状相关的位点。问题 4 实际上是问题 2 的延伸，在问题 2 中，对于是否患病这个因变量，我们最少只需一个位点，因为一个位点上有三个基因型，因此可以识别是否患病。当延伸至 10 个性状时，至多有 1000 个表现型($1000<2^{10}$)，此时需要至少 7 个位点才能共同识别($3^7>1000$)。与问题 3 相同，若直接考虑位点与表现型之间的相关性，计算量过大，我们先采取了降维的策略。将 1000 个样本的性状进行系统聚类，确定几个可能比较好的聚类数，得到更少的综合性状。在各聚类数下，分别利用检验统计量，得出最显著的位点。为了进一步验证所得位点与原始性状之间的关联性，我们可以再次利用检验统计量，观察位点能否识别初始性状。

三．基本假设

全基因组关联分析满足普通疾病一普通变量(common disease common variant，CDCV)的前提假设，即主张常见复杂疾病的基因风险部分大部分是由少数几个频率较高但是具有小到适中效应的位点所承担。

四．模型建立与求解

4.1 问题 1 的分析

对 DNA 序列进行数字编码具有十分重要的意义。一方面数字化是当前信息革命的主要趋势；另一方面，对 DNA 序列进行数字编码更简单明了，更便于存储和查询以及数值运算、数据处理。

现行对碱基的数值型编码，主要是应用 0(00),1(01),2(10),3(11)4 个数字对 DNA 的 4 种碱基(C,T,A,G)进行二进制数字编码。同时考虑到碱基互补原则和分子量顺序编码原则，最终编码确定为 0123(CTAG)。但是在本问题中，我们都以位点为最小单位作为分析的对象，每个位点有 2 个等位基因，且不同的位点之间相互独立，因此，我们只需对每个位点上的 3 个基因型分别编码，不影响数据的分析。以位点 rs3094315 为例，该位点有 TT、TC、CC 三种碱基对，相应的数值编码如下表所示。

基因型	TT	TC	CC
数值编码	0	1	2

表 1：碱基对编码数值举例

4.2 问题 2 的模型与求解

对于问题 2，我们的建模思路分三部。第一步是对数据的预处理，在高通量数据分析之前，对于不满足遗传学中最小等位频率小于 0.01 和哈代温伯格定律的 SNP 位点，从数据集中剔除，减少后续计算量。第二步是通过统计分析和检验，以位点为自变量，是否患病为因变量，利用卡方检验和 Fisher 精确检验，分析出与疾病相关的位点。由于上述关联分析主要从自变量和因变量之间独立性的角度出发，我们第三步主要考察自变量对因变量的预测能力，计算信息值（information value, IV），由于不同的 IV 值代表不同的预测能力，IV 值取值越大，预测能力也就越强，通过比较 IV 值的大小，并结合全基因组关联分析的结果，便可求得疾病 A 最有可能的致病位点。

4.2.1 数据预处理

在遗传统计学中，对基因组范围内 SNP 分析前，需涉及到对其 2 个方面的处理：一

是最小等位频率>0.01；二是 H-W 检验 p 值<0.01.

（1）最小等位基因频率[1]：MAF(minor allele frequency)

最小等位基因频率通常是指在给定人群中的不常见的等位基因发生频率，例如 TT，TC，CC 三个基因型，在人群中 C 的频率=0.28，T 的频率=0.72，则等位基因 C 的频率为最小等位基因频率，MAF=0.28。

在关联研究中，较小的 MAF 将会使统计效能降低，从而造成假阴性的结果。通常情况下要求 MAF>0.01 或 0.05。

（2）Hardy-Weinberg 平衡定律

在理想状态下，各等位基因的频率和等位基因的基因型频率在遗传中是稳定不变的，即保持着基因平衡。

设某一基因上有一对等为基因 T，C，三种可能的基因型分别是 TT，TC，CC，对应的频数为 N1，N2，N3（N=N1+N2+N3），f_{TT}，f_{TC}，f_{CC}分别相应基因型的频率。群体中等位基因 T 的频率 p 为：$p= f_{TT}+1/2*f_{TC}$，等位基因 C 的频率 q 为：$q=1/2*f_{TC}+f_{CC}$，$q=1-p$。三种基因型的平衡频率符合：AA：Aa：aa=p2：2pq：q2。假设检验中，把前面计算得到的等位基因频率带入基因型的平衡频率，再乘以总人数，就可以得到满足 Hardy-Weinberg 平衡的基因型期望频数。再将之与各基因型实际人数比较，进行 x^2 检验，即：

$$\chi^2 = (N_1 - Np^2)/Np^2 + (N_2 - 2Npq)/2Npq + (N_3 - Nq^2)/Nq^2 \sim \chi^2(1)$$

通常，H-W 检验的 p 值大于 0.01，即可认为该位点符合 Hardy-Weinberg 平衡定律。

经过上述两步的数据预处理后，我们从中剔除 rs6681946、rs645520 等 97 个不满足上述条件的位点，具体 MAF、HW 值及其 P 值见附录表 A“位点筛选排除结果”。

位点	MIF	HW值	HW-P值
rs6681946	0.37	16.31	5.4E-05
rs645520	0.15	13.70	2.1E-04
rs9804128	0.24	13.49	2.4E-04
rs271344	0.47	13.00	3.1E-04
rs868688	0.21	12.70	3.6E-04
…	…	…	…

表 2：位点筛选排除结果举例

[1]:http://wenku.baidu.com/link?url=3SIFJ_DESF_-2TTkBYqbRwYQ1KMcuWaInbpDTk63XLtSTQyM1MOv6waXmMnjUC52kOBPK_lJ8LHljb-lUcRyTzB3cT08iQrzzYmE_RGCXuG

4.2.2 全基因组关联分析

全基因组关联分析的原理是提取生病组和正常组个体的基因组 DNA，利用基因芯片做全基因组 SNP 分析，以百万计的 SNP 为标记，用统计学的方法进行病例一对照总体关联分析，通过比较找出两组个体之间每个遗传变异及其频率的差异，统计分析每个变异与目标性状之间的关联性大小，选出有显著不同的 SNP 位点，从而将疾病的病因定位于那些 SNP 位点上，而具有那些 SNP 的基因变成那些疾病的“易感基因”。理论上，检测到的多态位点越多，识别致病基因的可能性就越大。

在全基因组关联分析的整个过程中，我们已经得到 9348 个位点，因此只需直接对 1000 个样本的位点进行统计检验即可。对每个位点，我们提出原假设和备择假设：

- H_0：位点上三种不同的基因型与患病独立；
- H_1：位点上三种不同的基因型与患病不独立

对此假设，构造 χ^2 统计量，进行 χ^2 检验。以位点 rs3094315 为例，χ^2 检验过程如下。

在 1000 个样本中，位点 rs3094315 共有 3 个基因型：CC,TC 和 TT，其观测值和期望值如下表所示：

观测值：

基因型	CC	TC	TT	合计
健康者	26	147	327	500
患者	17	146	337	500
合计	43	293	664	1000

期望值：

基因型	CC	TC	TT	合计
健康者	21.5	146.5	332	500
患者	21.5	146.5	332	500
合计	43	293	664	1000

$$\chi^2 = \frac{(26-21.5)^2}{21.5} + \frac{(147-146.5)^2}{146.5} + \frac{(327-332)^2}{332}$$
$$+ \frac{(17-21.5)^2}{21.5} + \frac{(146-146.5)^2}{146.5} + \frac{(337-332)^2}{332} = 2.0377$$

当样本较大时，χ^2 统计量近似服从自由度为 2 的 χ^2 分布。上述卡方值相应概率 p=0.3610>0.01，故无法拒绝原假设。

针对满足 Hardy-Weinberg 平衡定律及最小等位基因频率准则的 9348 个位点，我们

采用和上述相同方法，利用 SAS 软件，对所有位点分别与疾病 A 的观测结果两两进行关联分析，得出所有位点的 χ^2 统计量和相应的 p 值，在 $\alpha=0.01$ 的显著性水平下，最终得到 73 个与患病相关的位点，由于 χ^2 统计量是一个近似的统计量，因此我们还提供了 Fisher 精确检验的结果，所得结果与 χ^2 检验高度一致，具体结果如下所示。

基因	卡方检验	卡方 - P值	FISHR精确检验	FISHER-P值	基因	卡方检验	卡方 - P值	FISHR精确检验	FISHER-P值
rs2273298	28.84	5.5E-07	2.5E-09	5.1E-07	rs12141588	10.52	5.2E-03	4.8E-05	4.6E-03
rs2250358	19.10	7.1E-05	3.2E-07	5.0E-05	rs1188347	10.37	5.6E-03	4.2E-05	5.5E-03
rs7543405	16.91	2.1E-04	7.8E-07	2.2E-04	rs364642	10.34	5.7E-03	2.1E-05	5.9E-03
rs932372	16.65	2.4E-04	3.9E-06	1.4E-04	rs12139487	10.27	5.9E-03	2.3E-05	5.8E-03
rs12036216	15.54	4.2E-04	5.9E-06	4.1E-04	rs4920653	10.19	6.1E-03	4.6E-05	6.3E-03
rs9426306	15.50	4.3E-04	1.5E-06	4.3E-04	rs4243820	10.14	6.3E-03	4.4E-05	5.3E-03
rs12145450	15.03	5.5E-04	2.0E-06	5.3E-04	rs2788891	10.05	6.6E-03	2.4E-05	6.7E-03
rs7368252	14.56	6.9E-04	3.9E-06	7.0E-04	rs2745260	10.03	6.6E-03	7.0E-05	6.3E-03
rs4391636	14.36	7.6E-04	3.6E-06	7.4E-04	rs11121557	10.03	6.7E-03	2.4E-05	6.8E-03
rs5746051	13.42	1.2E-03	4.7E-06	1.2E-03	rs12036552	9.98	6.8E-03	4.3E-05	6.7E-03
rs7522344	13.35	1.3E-03	4.7E-06	1.3E-03	rs2095518	9.94	6.9E-03	2.5E-05	6.9E-03
rs2999878	13.29	1.3E-03	1.2E-05	1.0E-03	rs848214	9.88	7.1E-03	5.2E-05	6.6E-03
rs2807345	13.20	1.4E-03	7.8E-06	1.4E-03	rs2182703	9.87	7.2E-03	2.6E-05	7.1E-03
rs4646092	12.93	1.6E-03	7.1E-06	1.5E-03	rs1891419	9.84	7.3E-03	4.7E-05	7.0E-03
rs880801	12.88	1.6E-03	1.3E-05	1.5E-03	rs1148455	9.70	7.8E-03	2.8E-05	7.8E-03
rs1883567	12.88	1.6E-03	5.7E-06	1.6E-03	rs6699113	9.67	7.9E-03	5.0E-05	7.6E-03
rs15045	12.85	1.6E-03	6.0E-06	1.7E-03	rs7555715	9.66	8.0E-03	3.4E-05	8.4E-03
rs3013045	12.68	1.8E-03	9.6E-06	1.7E-03	rs3818033	9.64	8.1E-03	5.1E-05	7.8E-03
rs11580218	12.58	1.9E-03	1.1E-05	1.9E-03	rs10779765	9.60	8.2E-03	5.3E-05	8.7E-03
rs12133956	12.52	1.9E-03	9.9E-06	1.8E-03	rs556596	9.56	8.4E-03	4.1E-05	8.0E-03
rs2143810	12.25	2.2E-03	8.5E-06	2.2E-03	rs2244300	9.50	8.6E-03	4.4E-05	8.8E-03
rs10754873	12.13	2.3E-03	1.2E-05	2.1E-03	rs2473246	9.50	8.7E-03	5.3E-05	8.5E-03
rs4912019	12.04	2.4E-03	1.6E-05	2.0E-03	rs2097518	9.49	8.7E-03	7.7E-05	7.5E-03
rs1541318	11.92	2.6E-03	1.1E-05	2.5E-03	rs12028945	9.48	8.8E-03	1.1E-04	9.5E-03
rs11247865	11.85	2.7E-03	3.2E-05	2.3E-03	rs1201394	9.45	8.9E-03	4.3E-05	8.8E-03
rs12752833	11.49	3.2E-03	1.7E-05	3.2E-03	rs6657574	9.42	9.0E-03	9.4E-05	9.0E-03
rs3765695	11.45	3.3E-03	2.1E-05	2.9E-03	rs7550997	9.42	9.0E-03	1.2E-04	8.3E-03
rs6658098	11.30	3.5E-03	1.4E-05	3.6E-03	rs495223	9.42	9.0E-03	6.1E-05	8.5E-03
rs10779763	11.12	3.9E-03	2.0E-05	3.6E-03	rs10915577	9.38	9.2E-03	1.0E-04	9.3E-03
rs11573253	10.90	4.3E-03	1.6E-05	4.3E-03	rs946758	9.37	9.2E-03	4.0E-05	9.1E-03
rs707472	10.88	4.3E-03	1.6E-05	4.3E-03	rs6683624	9.37	9.2E-03	3.8E-05	9.1E-03
rs9659647	10.84	4.4E-03	2.2E-05	4.3E-03	rs2038095	9.34	9.4E-03	3.5E-05	9.3E-03
rs1888759	10.77	4.6E-03	2.3E-05	4.2E-03	rs10864304	9.31	9.5E-03	4.2E-05	9.4E-03
rs2301461	10.69	4.8E-03	3.3E-05	4.5E-03	rs829404	9.30	9.5E-03	3.7E-05	9.6E-03
rs590368	10.63	4.9E-03	2.3E-05	4.8E-03	rs2651935	9.24	9.8E-03	3.5E-05	9.9E-03
rs731024	10.58	5.1E-03	5.3E-05	4.4E-03	rs4310409	9.22	1.0E-02	1.2E-04	8.8E-03
rs1138333	10.55	5.1E-03	5.7E-05	5.5E-03					

表 4：位点相关性检验结果（$\alpha=0.01$）

上表显示统计检验最显著的 7 个位点： rs2273298，rs2250358，rs7543405，rs932372，rs12036216，rs9426306，rs12145450 的信息，可以看出，在显著性平$\alpha = 0.01$的情况下，没有理由拒绝原假设，即可以认为以上 7 个位点与患病相关。但是统计分析只是从自变量与因变量相关性的角度出发，选择与患病相关的位点，无法提供位点与疾病之间联系的强度，因此我们需考虑位点所提供的信息值。

4.2.3 基于信息值 IV 的变量选择

挑选解释变量的过程是个比较复杂的过程，需要考虑的因素很多，比如：变量的预测能力，变量之间的相关性，变量的可解释性等等。上一节我们主要考虑变量之间的相关性，但是，变量选择最主要和最直接的衡量标准是变量的预测能力。在方法的选择上，也是种类多样，比如主成分分析、信息熵、基尼系数法、遗传算法等，针对本问题，我们采用信息值的方法来进行变量选择，不仅仅是因为 IV 值是根据 WOE 值衍生而来，使同一变量的不同组间产生很大的差别，而且通过 IV 值能够判断变量对于类别标签影响的重要程度，即预测能力。

本节将先介绍证据权重值(WOE)的基本概念和计算方法，基于此定义信息值，并选择信息值最大，预测能力最强的位点，将所得结果与统计检验的结果分析对比，选出最有可能的致病位点。

（1）证据权重转换(weight of evidence)

证据权重转换的目的就是为了对每个位点下三个不同的基因型进行重新编码，给每个位点下的每个基因型赋予一个唯一的 WOE 值。WOE 转换的结果就是同一变量的不同组间将会产生很大的差别，可以用来衡量变量某个属性的风险。

首先，我们仍以位点 rs3094315 为例，样本总数 1000，其中基因型为 CC 的样本有 43 个，基因型为 TC 的样本有 293 个，基因型为 TT 的样本有 664 个，健康者和患者分布如下表：

基因型	健康者	患者	健康者的分布	患者的分布	WOE
CC	26	17	5.2	3.4	-0.4249
TC	147	146	29.4	29.2	-0.0068
TT	327	337	65.4	67.4	0.03012
合计	500	500			

表 5：WOE 值计算举例

其中两个分布列分别由每一类别中的频数除以健康者或患者的总数而得到。例如基因型 CC 健康者的分布为 26/500=5.2。而每个类别的 WOE 定义如下：

$$WOE = \ln(\frac{bad\ \ distribution}{good\ \ distribution})$$

$$good\ \ distribution = \frac{number\ \ of\ \ good}{total\ \ \ number\ \ of\ \ good}$$

$$bad\ \ distribution = \frac{number\ \ of\ \ bad}{total\ \ \ number\ \ of\ \ bad}$$

对数部分的分子和分母顺序可以调换，不同的分子分母，WOE 值表示的意思不同但是在同一模型中要保持一致。我们之后 WOE 的计算均采用上述的定义。且 WOE 的值越高，代表着该分组中样本是患者的风险越高。

（2）信息值(information value, IV)

信息价值 IV 是用来衡量位点对健康者和患者区分能力的一个指标，其计算公式如下：

$$IV_i = \sum_{j=1}^{3}(bad\ \ \ distribution_j - good\ \ \ distribution_j) * WOE_j$$

其中 i 表示位点个数，j=1,2,3 表示对每个位点 i 有 3 个基因型分组，上式所得即为第 i 个位点对于的信息值。IV 的值越大，表示是否患病在该位点上的分布差异越大，也即该位点的区分能力越好，且 IV 的取值范围为[0,+∞).

根据上述的方法，我们计算出所有位点的 IV 值。为了充分说明 IV 值的有效性，我们从 IV=0,IV∈(0,+∞)以及 IV=+∞三个方面考察位点包含的信息值。

当 IV=0 时，其频率分布表如图示。可以看出，对于位点 rs910660，无论基因型取值为何，对表现型没有影响，即以 50%的概率进入健康组或患病组，说明该位点没有提供任何信息，与 IV=0 结果一致。

频数 百分比 行百分比 列百分比	表－phenotype＊rs910660			
	rs910660			
phenotype	CC	CT	TT	合计
0	276 27.60 55.20 50.00	181 18.10 36.20 50.00	43 4.30 8.60 50.00	500 50.00
1	276 27.60 55.20 50.00	181 18.10 36.20 50.00	43 4.30 8.60 50.00	500 50.00
合计	552 55.20	362 36.20	86 8.60	1000 100.00

表 6：位点 rs910660 与疾病 A 频数统计

对于 IV=+∞，从如下的频率分布表可以看出，对于位点 rs12742921，只要该位点的基因型取值为 CC，我们可以判断该样本患病，即位点可以给出确定的信息，因此对应信息值为无穷。

频数 百分比 行百分比 列百分比	表－phenotype＊rs12742921			
	rs12742921			
phenotype	CC	TC	TT	合计
0	0 0.00 0.00 0.00	88 8.80 17.60 49.16	412 41.20 82.40 50.49	500 50.00
1	5 0.50 1.00 100.00	91 9.10 18.20 50.84	404 40.40 80.80 49.51	500 50.00
合计	5 0.50	179 17.90	816 81.60	1000 100.00

表 7：位点 rs12742921 与疾病 A 频数统计

若 IV 取有限值，如位点 rs3094315，若该位点的基因型为 CC，则我们有 60%的概率判断该样本进入健康组，若该位点的基因型为 TC 或是 TT，则样本以几乎相同的概率进入健康者或疾病组，也即该位点提供有限信息。

频数 期望值 偏差 百分比 行百分比 列百分比	表－phenotype＊rs3094315			
	rs3094315			
phenotype	CC	TC	TT	合计
0	26 21.5 4.5 2.60 5.20 60.47	147 146.5 0.5 14.70 29.40 50.17	327 332 -5 32.70 65.40 49.25	500 50.00
1	17 21.5 -4.5 1.70 3.40 39.53	146 146.5 -0.5 14.60 29.20 49.83	337 332 5 33.70 67.40 50.75	500 50.00
合计	43 4.30	293 29.30	664 66.40	1000 100.00

表 8：位点 rs:3094315 与疾病 A 频数统计

通过上述分析，可以看出 IV 值的有效性。基于此，我们挑选出 IV 值最高的 7 个位点，如下表所示：

位点	IV值
rs17262293	inf
rs12742921	inf
rs2273298	11.68
rs2250358	8.58
rs7543405	6.80
rs932372	6.77
rs12036216	6.44

表 9：IV 值筛选结果

将 IV 值筛选出的位点与统计检验筛选的位点对比，可以发现，信息值挑选出两个信息量为无穷的位点:rs12742921,rs17262293。为了考察两个位点特殊性的原因所在，我们可以分析这两个位点的频率分布图：

频数 百分比 行百分比 列百分比	表 - phenotype * rs12742921			
	rs12742921			
phenotype	CC	TC	TT	合计
0	0 0.00 0.00 0.00	88 8.80 17.60 49.16	412 41.20 82.40 50.49	500 50.00
1	5 0.50 1.00 100.00	91 9.10 18.20 50.84	404 40.40 80.80 49.51	500 50.00
合计	5 0.50	179 17.90	816 81.60	1000 100.00

表 10:特殊位点 rs12742921 统计结果

频数 百分比 行百分比 列百分比	表－phenotype＊rs17262293			
	rs17262293			
phenotype	AA	GA	GG	合计
0	0 0.00 0.00 0.00	111 11.10 22.20 52.86	389 38.90 77.80 49.62	500 50.00
1	6 0.60 1.20 100.00	99 9.90 19.80 47.14	395 39.50 79.00 50.38	500 50.00
合计	6 0.60	210 21.00	784 78.40	1000 100.00

表 11:特殊位点 rs17262293 统计结果

从位点 rs17262293 与疾病 A 的频数统计表可知，只要该位点上的基因型为 AA，则可以断定样本一定患病，同理，对于位点 rs12742921，只要该位点上的基因型为 CC，也可断定样本患病。但是上述推断存在问题，即两个位点在相应基因型上的样本数很少，分别是 5 和 6，因此可能存在因为样本量太少而造成误断。

利用 χ^2 检验无法扑捉到这两个位点的原因在于，进行 χ^2 检验时，需要满足 χ^2 分布的期望值准则，即当交叉分类为两类是，要求样本量应该足够大，且每一交叉类别的期望频数不能偏小，否则进行 χ^2 检验可能会得出错误的结论。由于上述两个位点在对应基因型上频数偏小，因此导致 χ^2 检验出现错误。

分析变量： IV								
N	最大值	均值	最小值	众数	方差	标准差	偏度	峰度
9445	11.6814819	0.7890613	0	0	0.6488917	0.8055381	2.2509535	9.5553965

表 12:IV 值全集统计信息

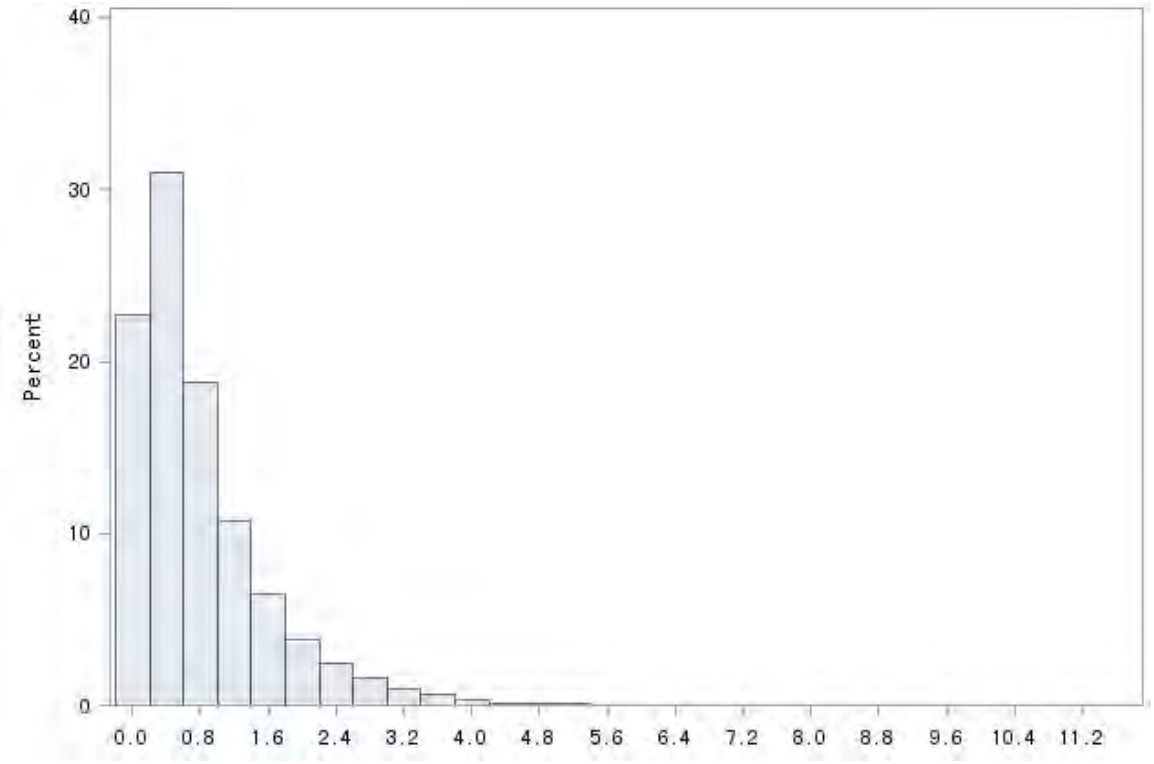


表 13:IV 值频数统计图

综合统计检验和信息值的结果，我们最终认为位点 rs2273298 是疾病 A 最有可能的致病位点。因为该位点所包含的信息量最大，且统计显著。

频数 百分比 行百分比 列百分比	表－phenotype＊rs2273298			
	rs2273298			
phenotype	AA	AG	GG	合计
0	305 30.50 61.00 57.87	161 16.10 32.20 42.48	34 3.40 6.80 36.17	500 50.00
1	222 22.20 44.40 42.13	218 21.80 43.60 57.52	60 6.00 12.00 63.83	500 50.00
合计	527 52.70	379 37.90	94 9.40	1000 100.00

表 14:位点 rs2273298 的统计表

4.3 问题 3 的建模与求解

为了从 300 个基因中挑选出于疾病关联度最大的基因，我们选取了决策树的方法。由于直接对所有基因做决策树，耗时巨大，因此我们采用降维的策略，基于一定的筛选

标准，选出备择基因集。在此基础上，利用决策树进一步挑选。

4.3.1 备选基因的筛选

由于基因与疾病的相关性通过基因的最小单位——位点起作用，因此在筛选备选基因时，可以从位点与疾病的相关性角度出发。基于问题 2 的求解，利用 IV 值和 χ^2 统计量，我们从中挑选出 10 个 IV 值最高的位点，对应得到 10 个基因，如表 15 所示。基因筛选时需要考虑的另一个因素是每个基因出现的频数，频数越高，即说明基因中对疾病有影响的位点越多，我们可以初步判断这些基因与疾病存在相关性。基于此标准，我们得到表 16 所示的 76 个基因。综合考虑上述两个标准，我们最终得出表 17 所示的 7 个备选基因：

排名	基因	位点	IV值	卡方统计量	卡方-P值
1	gene_102	rs2273298	11.68	28.84	5.5E-07
2	gene_62	rs2250358	8.58	19.10	7.1E-05
3	gene_265	rs7543405	6.80	16.91	2.1E-04
4	gene_245	rs932372	6.77	16.65	2.4E-04
5	gene_3	rs12036216	6.44	15.54	4.2E-04
6	gene_274	rs9426306	6.30	15.50	4.3E-04
7	gene_125	rs2999878	6.12	13.29	1.3E-03
8	gene_298	rs12145450	6.07	15.03	5.5E-04
9	gene_55	rs7368252	5.86	14.56	6.9E-04
10	gene_35	rs4391636	5.79	14.36	7.6E-04

表 15: 基于 IV 值筛选基因

频数	基因	基因相应位点	频数	基因	基因相应位点
5	gene_265	rs7543405,rs4243820, rs495223，rs556596，rs4466676	1	gene_183	rs1138333
3	gene_293	rs1188347，rs7555715，rs1201394	1	gene_185	rs4912048
3	gene_55	rs7368252，rs7522344，rs12036552	1	gene_192	rs1883567
2	gene_113	rs10779763，rs2745260	1	gene_193	rs4912019
2	gene_114	rs12139487，rs3117067	1	gene_196	rs10916703
2	gene_121	rs5746051，rs590368	1	gene_197	rs11573253
2	gene_125	rs2999878，rs3013045	1	gene_199	rs6657574
2	gene_144	rs10754873，rs16851049	1	gene_204	rs10916846
2	gene_162	rs2143810，rs3818033	1	gene_205	rs1888759
2	gene_169	rs16830759，rs648305	1	gene_208	rs1256341
2	gene_191	rs7535952，rs6702295	1	gene_21	rs3795263
2	gene_217	rs2807345，rs2473246	1	gene_210	rs829404
2	gene_22	rs2651935，rs10492941	1	gene_216	rs909813
2	gene_245	rs932372，rs6699113	1	gene_218	rs11580218
2	gene_254	rs11247865，rs7550997	1	gene_222	rs2182703
2	gene_274	rs9426306，rs2788891	1	gene_224	rs7543064
2	gene_30	rs3765695，rs10910024	1	gene_235	rs12028945
2	gene_43	rs6658098，rs10915577	1	gene_236	rs12045815
2	gene_78	rs2301461，rs2097518	1	gene_249	rs15045
1	gene_100	rs12032209	1	gene_26	rs946758
1	gene_102	rs2273298	1	gene_260	rs12752833
1	gene_103	rs4846212	1	gene_268	rs9659647
1	gene_104	rs11121557	1	gene_28	rs880801
1	gene_106	rs2480773	1	gene_280	rs9286945
1	gene_115	rs10779765	1	gene_281	rs11582200
1	gene_126	rs1148455	1	gene_286	rs2377060
1	gene_128	rs2038095	1	gene_294	rs1891419
1	gene_138	rs10927414	1	gene_298	rs12145450
1	gene_142	rs6429674	1	gene_3	rs12036216
1	gene_149	rs2473344	1	gene_35	rs4391636
1	gene_150	rs4646092	1	gene_42	rs1541318
1	gene_153	rs848214	1	gene_45	rs364642
1	gene_160	rs4920653	1	gene_49	rs6683624
1	gene_167	rs2244300	1	gene_62	rs2250358
1	gene_17	rs12133956	1	gene_67	rs731024
1	gene_172	rs2095518	1	gene_79	rs10864304
1	gene_175	rs4310409	1	gene_83	rs707472
1	gene_181	rs12141588	1	gene_85	rs7527904

表 16:基于基因频数筛选基因

基因	所含位点数	可能相关的位点
gene_3	44	rs12036216
gene_62	34	rs2250358
gene_102	10	rs2273298
gene_245	14	rs932372，rs6699113
gene_55	60	rs7368252，rs7522344，rs12036552
gene_293	60	rs1188347，rs7555715，rs1201394
gene_265	56	rs7543405,rs4243820, rs495223，rs556596，rs4466676

表 17:备选基因

4.3.2 利用决策树预测

决策树又称分类回归树(classification and regression tree，简称 CART)是 Breiman 等人于 1984 年提出的一种非参数方法，可以分为分类决策树和回归决策树两种。分类树通常用于解决因变量是分类变量（或称虚拟变量）的情况，而回归树则常用于因变量是连续变量的情况。本节我们关注的是分类树问题。

在构造决策树是，有一个重要的问题就是如何选择最佳分割，目前学术上有很多种度量方法可以用来确定最佳的分割方法，例如信息增益、增益率和基尼指数。这里我们选用基尼指数作为分裂准则，原因在于在对基因中某个位点进行分割时，位点存在三个基因型，即有三个划分，为了确定在节点上是二元划分，因此我们在节点处的测试可以看成是形如“TT ∈{TC,CC}？”的二元测试，因此选用基尼指数作为分割标准构造决策树。如下显示备选基因中 gene102 和 gene55 的决策树：

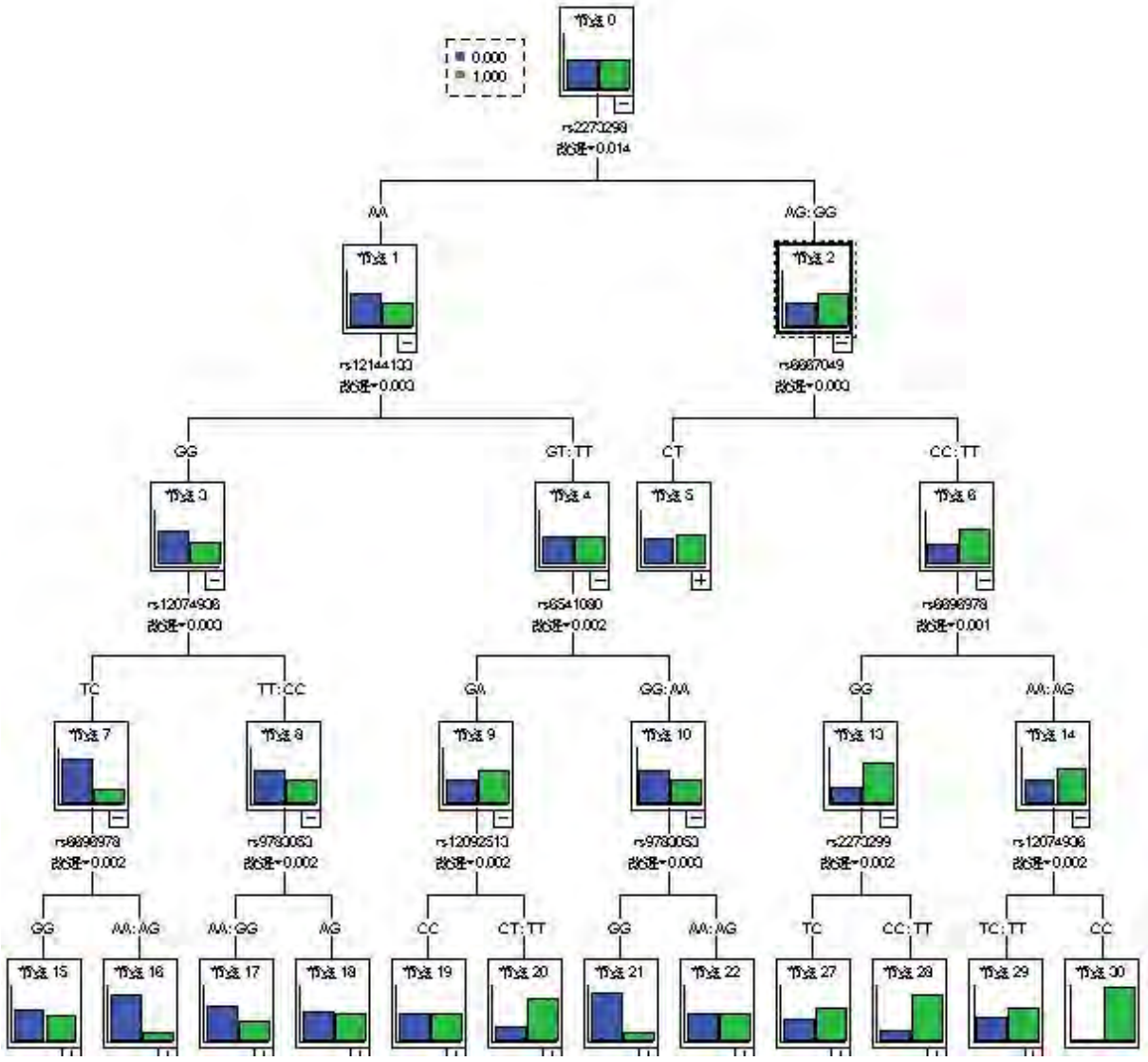


图 1：gene_102 作出的决策树

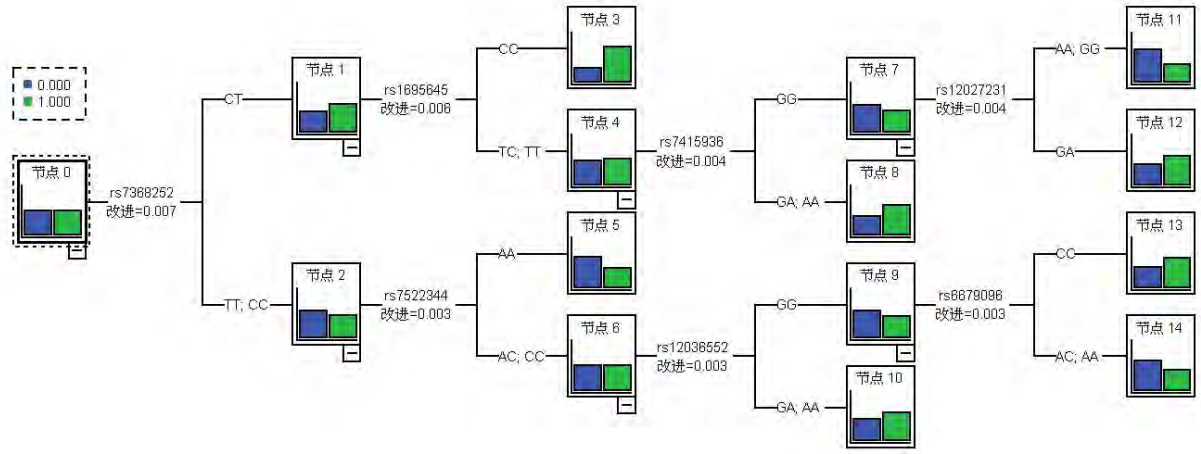


图 2：gene_55 作出的决策树

对于已经建立的分类模型，也即基于 10 个基因的决策树，我们需要对模型分类的准确率进行评估度量。分类准确率越高的决策树，说明其对应的基因越能反映健康状况，与疾病的关联性越强。在评估决策树预测的准确率时，我们主要从混淆矩阵和 ROC 曲线两个角度度量。

混淆矩阵是分析决策树识别不同类元组的一种有用工具。其表示如下：

		预测的类		
	列1	**Yes**	**No**	**合计**
实际的类	yes	TP	FN	P
	no	FP	TN	N
	合计	P’	N’	P+N

表 18:混淆矩阵示意图

其中 TP 和 TN 告诉我们决策树何时分类正确，而 FP 和 FN 告诉我们决策树何时分类错误。理想地，对于具有较高准确率的决策树，大部分元组应该被混淆矩阵中的对角线表示，而其他非对角线位置为 0 或接近 0. 即 FP 合 FN 接近 0.

接收者操作特征(receiver operating characteristic，ROC)曲线是反映决策树准确率的另一可视化工具。设 TP、FP、P 和 N 与上述混淆矩阵的定义相同，TPR 是该模型正确标记的正元组的比率，而 FPR 是该模型错误标记为正的负元组的比率，即 TPR=TP/P，TFP=FP/N，ROC 曲线显示了给定模型的真正例率(TPR)和假正例率(TFR)之间的权衡。ROC 曲线下方的面积是模型准确率的度量。ROC 曲线图还显示了一条对角线，代表随机猜测，模型的 ROC 曲线离对角线越近，模型的准确率越低。

仍以上述两个基因为例，对决策树的准确率进行评估，如下显示两个基因的混淆矩阵信息以及 ROC 曲线：

分类

已观测	已预测		
	0	1	正确百分比
0	307	193	61.4%
1	191	309	61.8%
总计百分比	49.8%	50.2%	61.6%

增长方法:CRT

因变量列表: health

表 19: Gene_55 混淆矩阵

分类

已观测	已预测		
	0	1	正确百分比
0	341	159	68.2%
1	203	297	59.4%
总计百分比	54.4%	45.6%	63.8%

增长方法:CRT

因变量列表: health

表 20: Gene102_混淆矩阵

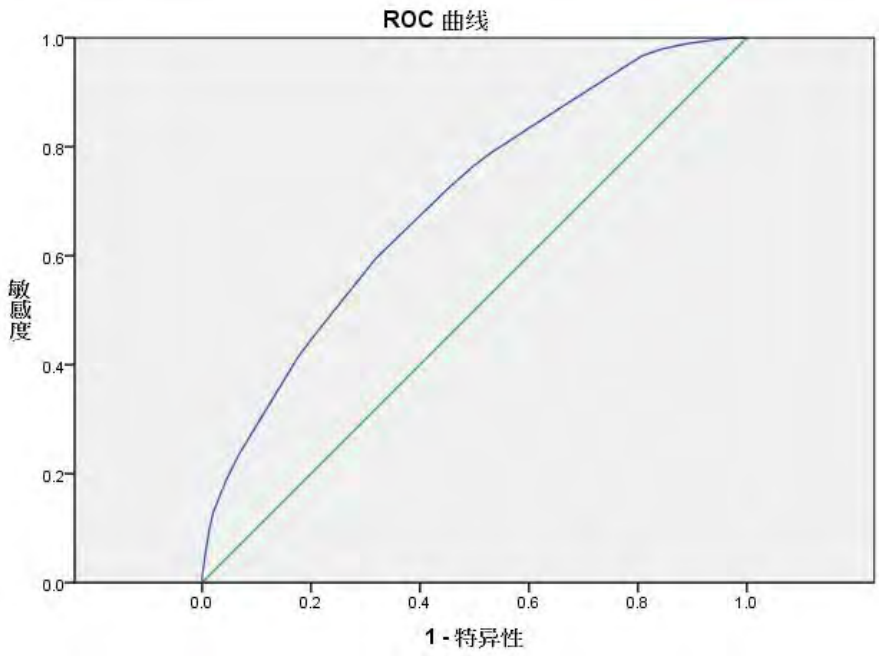


图 3: Gene_102 的 ROC 曲线图

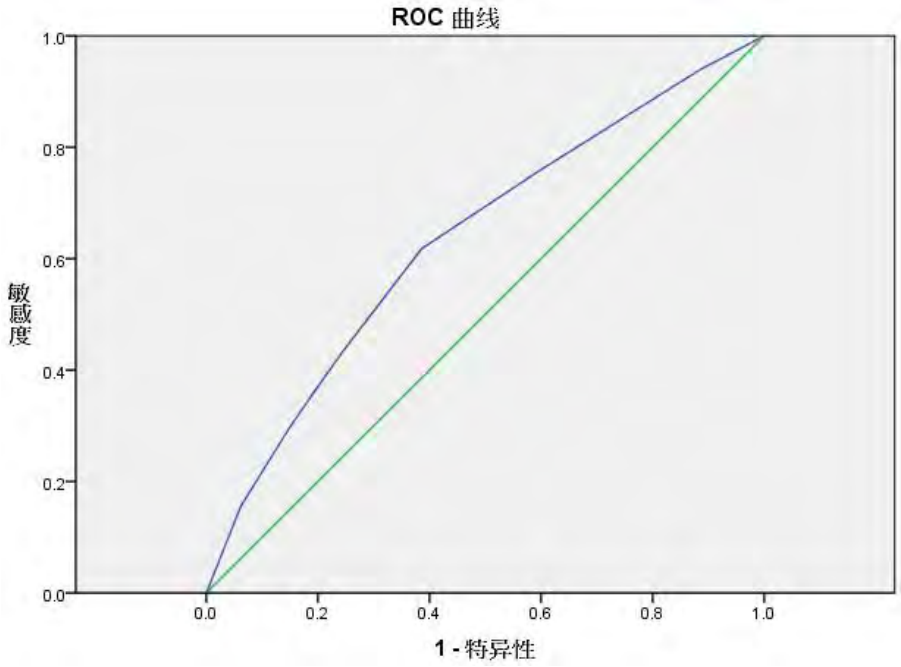


图 4: Gene_55 的 ROC 曲线图

通过比较所有备选基因决策树的准确率，最终发现 gene102 和 gene55 的准确率最高，分别为 63.8%和 61.6%。因此我们认为这两个基因与疾病 A 的关联性最强。

4.4 问题 4 的建模与求解

4.4.1 基于样本的聚类分析

对样本进行聚类分析，就是对数据集中的数据应用某种方法进行分组，将样本按相似程度或距离远近来划分类别，使得同一类中的样本之间的相似性比其他类的样本的相似性更强。其主要目的在于对本题中 1000 个样本的 10 个性状降维，发现样本的显性结构，找出与其相关的位点。

由于聚类数的设定对聚类结果会有一定的影响，我们首先通过系统聚类，确定几个较好的分类数。在进行聚类时，每个个体自成一类，然后根据距离最小原则，将最近的两类合成一类，并计算当前类与新类的距离，直至仅有一类。在分析中，我们选用类平均法衡量样本之间的距离。该方法用不同类的样本点两两之间的平均距离作为类间距离。如果有类 G_p 和 G_q，可以计算每类中每个样本点之间的平均距离，即：

$$D_{pq} = \frac{1}{N_p N_q} \sum_{i \in G_p} \sum_{j \in G_p} d(x_i, x_j)$$

若

$$d(x,y)=|x-y|^2$$

则新类 G_n 与其他的任意类 G_k 之间的距离系数由递推公式决定：

$$D_{kn}=\frac{N_pD_{kp}+N_qD_{kq}}{N_n}$$

即在并类的过程中，以类别样本点之间的平均距离作为依据并类，直至把所有样本归为一类。

在系统聚类的分类数划分上，我们主要参考参数 R2 和半偏 R2，R2 的值随着分类个数的减少而变小，每个样本各为一类时 R2=1，所有样本合并成一类时 R2=0。最合适分类的 R2 值不能太小，最好在 0.7 以上；其次，观察 R2 的变化，如果某次合并使得 R2 值减小很多，则这次合并前的分类数最佳。半偏 R2 = 上次合并后的 R2 值 - 这次合并后的 R2，显然半偏 R2 值较大的上一次合并为最佳合并。最终，我们确认 7,16,20,50,100 等不同的聚类数，各聚类数相应的统计量及指标如下

聚类数	立方聚类准则CCC	伪F统计量	近似期望总体R方
2	486.54	1990.17	0.08
3	361.97	1166.40	0.14
4	304.71	831.54	0.19
7	247.97	503.62	0.29
10	233.34	396.93	0.37
13	208.60	311.82	0.42
16	207.02	288.99	0.44
20	198.02	252.64	0.47
50	176.00	154.98	0.59
100	172.99	121.31	0.68

表 21:各聚类数的统计量

以聚类个数 7 为例，在确定聚类个数后，我们利用 K-均值聚类，将所有样本聚成 7 类综合性状。聚类结果如下。

类别	频数	均方根标准差	最近类别	质心间距离
1	422	0.20	6	1.39
2	414	0.20	7	1.40
3	51	0.45	7	1.05
4	27	0.44	5	1.19
5	43	0.42	4	1.19
6	20	0.42	3	1.16
7	23	0.42	3	1.05

表 22:聚类结果

Cluster	phenotype1	phenotype2	phenotype3	phenotype4	phenotype5	phenotype6	phenotype7	phenotype8	phenotype9	phenotype10
1	0.04	0.03	0.04	0.05	0.05	0.05	0.03	0.04	0.05	0.02
2	0.95	0.97	0.95	0.97	0.96	0.96	0.94	0.98	0.98	0.95
3	0.63	0.29	0.59	0.31	0.86	0.76	0.47	0.61	0.10	0.43
4	0.33	0.93	0.30	0.11	0.19	0.44	0.41	0.67	0.74	0.78
5	0.84	0.42	0.49	0.84	0.16	0.09	0.79	0.44	0.86	0.77
6	0.10	0.50	0.50	0.95	0.25	0.75	0.30	0.25	0.15	0.15
7	0.52	0.78	0.91	0.30	0.87	0.43	0.96	0.17	0.39	0.78

表 23:各类别信息

从各类别的频数统计结果可以发现，类别 1 有 422 个样本，类别 2 有 414 个样本。从各类别中心结果可以看出，聚类类别 Cluster1 的各个中心取值接近 0，反应在样本集中以十个性状都表现为 0 的样本位为代表。聚类类别 Cluster2 的各个中心取值接近于 1，反应在样本集中则以十个性状都表现为 1 的样本为代表。

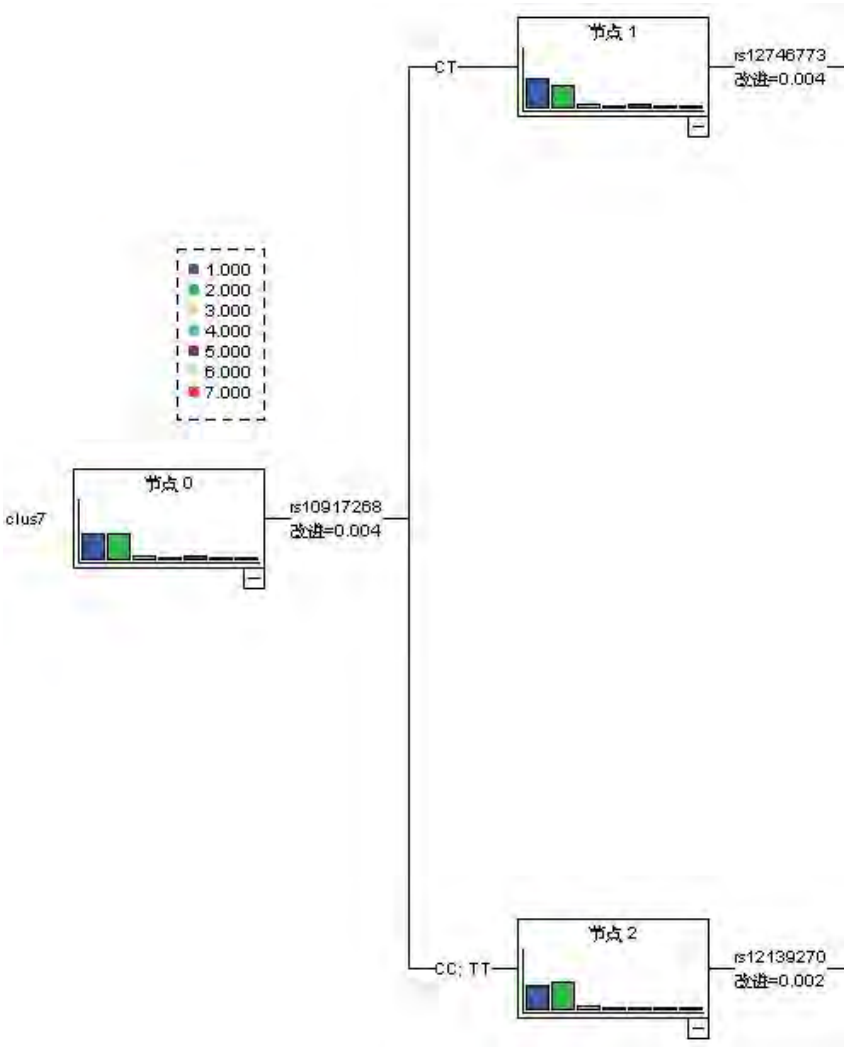
为了分析与 10 个性状，也即 7 类综合性状相关联的位点，我们对每个位点关于 7 类综合性状作 χ^2 统计检验，结果如表 24 所示。χ^2 统计量越大，即位点与综合性状的关联性越大，我们可以从中挑选出 χ^2 统计量较大的 10 个位点，作为致病位点的备选。

在选出最有可能的 10 个致病位点后，我们将数据还原成原始的 10 个初始性状，直接分析二者的相关性，验证上述结果的合理性。即对每个性状，利用 χ^2 统计检验，分别考察位点与该性状是否相关，得到不同置信水平下 χ^2 检验对应的 p 值，结果如下图。

rank_clus7	locus	卡方Clus7	卡方P值Clus7	P_phenotype1	P_phenotype2	P_phenotype3	P_phenotype4	P_phenotype5	P_phenotype6	P_phenotype7	P_phenotype8	P_phenotype9	P_phenotype10
1	rs12746773	42.34	2.92E-05	✓	✓	✓	✓	✗	✗	✓	!	✓	✓
2	rs4584380	39.50	8.7E-05	✗	✗	✗	!	✗	✓	✗	✗	✗	!
3	rs11249201	38.95	0.000107	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
4	rs12139270	37.91	0.000159	✓	!	!	✗	✗	✓	!	✗	!	✗
5	rs2075972	37.23	0.000205	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
6	rs1985278	36.42	0.000278	✗	✗	✗	✗	✗	!	✗	✗	✗	✗
7	rs6603797	36.03	0.00032	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
8	rs10917268	35.19	0.000436	✗	✓	✗	✓	✓	✓	✓	!	✗	✓
9	rs657816	35.18	0.000438	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
10	rs2821063	35.04	0.000461	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
			✓	<=1%		!	(1%, 5%]		✗	>5%			

表 24:检验统计量

由上图可知，所挑选的 10 个位点中的 8 个位点，都与题目所给的 10 个初始性状中的一个或多个性状高度相关。为考察所得的 10 个位点对综合性状的分类预测能力，建立如下的决策树：


节点 1
CT
rs12746773
改进=0.004
1.000
2.000
3.000
4.000
5.000
6.000
7.000
节点 0
clus7
rs10917268
改进=0.004
节点 2
CC; TT
rs12139270
改进=0.002

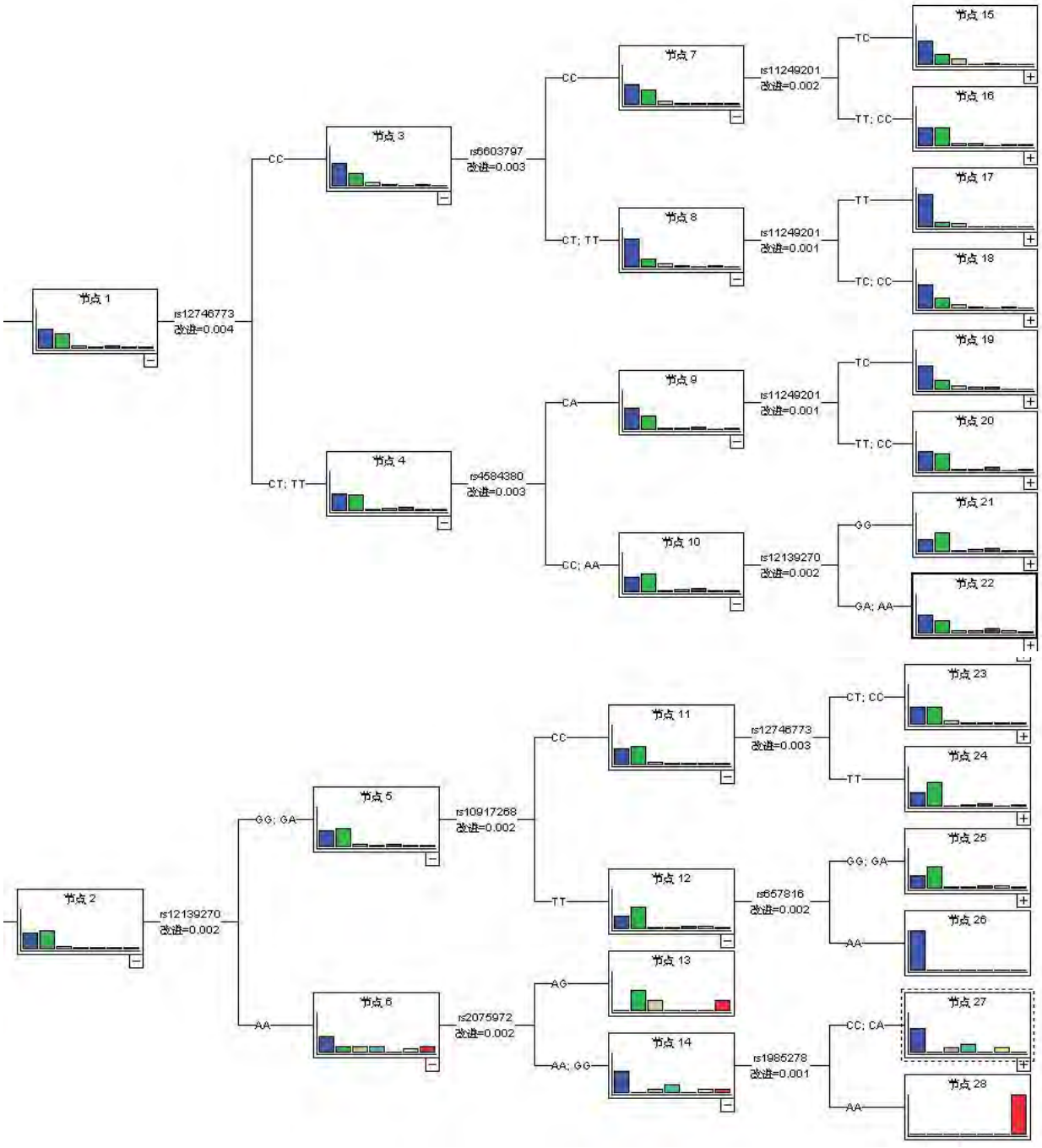


图 5:聚类后的决策树

根据混淆矩阵和 ROC 曲线，可以得到该决策树具有较好的预测能力，且由决策树可知位点 rs10917268 具有最多的信息量。

但聚类为 7 时，由于期望 R^2=0.29<0.5，即聚类后所包含的信息来那个过少，且 R^2 的值随着分类个数的减少而变小，因此我们又考虑了聚类个数为 50 的情况。和聚类个数为 7 相同的过程，利用 K-均值聚类法聚成 50 类后，挑选统计检验显著的前 10 个位点分别与性状做统计检验，χ^2 检验统计量对应 p 值如下表所示：

rank_clus50	locus	卡方Clus50	卡方P值Clus50	P值_phenotype1	P值_phenotype2	P值_phenotype3	P值_phenotype4	P值_phenotype5	P值_phenotype6	P值_phenotype7	P值_phenotype8	P值_phenotype9	P值_phenotype10
1	**rs7524800**	**205.80**	**1.2E-09**	✗	✗	✗	✗	!	!	✗	✗	✗	✗
2	rs4662057	205.09	1.5E-09	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
3	**rs6697749**	**198.77**	**7.8E-09**	✗	✗	✗	✓	✗	!	✗	✗	✗	✗
4	rs12085934	196.54	1.4E-08	✗	✗	✗	✗	✗	✗	✗	✗	✗	!
5	**rs883867**	**196.52**	**1.4E-08**	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
6	rs4908443	195.79	1.7E-08	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
7	**rs7532691**	**194.31**	**2.5E-08**	✗	✗	✗	✗	✗	✗	✗	!	✗	✗
8	rs12139270	189.05	9.5E-08	✓	!	!	✗	✗	✓	!	✗	!	✗
9	**rs2355459**	**186.29**	**1.9E-07**	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
10	rs4662072	184.33	3.1E-07	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
			✓	<=1%		!	(1%, 5%]		✗	>5%			

表 25:统计量

由上表可知，所挑选的 10 个位点中的 8 个位点，都与题目所给的 10 个初始性状中的一个或多个性状高度相关。

对于其他聚类类数下的统计分析见附录。

综合以上聚类数 7、和聚类数 50 统计检验的结果，我们认为与 10 个关联性状所有表现出的综合性状相关的位点有 rs12746773，rs4584380，rs11249201，rs12139270，rs2075972，rs1985278，rs6603797，rs10917268。

五．结论

六．对于问题一，由于每个位点之间相互独立，我们对每个位点上的 3 个基因型分别进行 0,1,2 型的数值编码，既便于存储和查询又方便数值运算、数据处理。

七．对于问题二，首先基于生物遗传学定律，对位点数据进行预处理，去除不符合要求的 97 个位点。然后对疾病和位点进行统计分析，通过 χ^2 检验和 Fisher 精确检验，二者得出一致的统计检验显著的位点，即与疾病存在相关的位点。但由于统计检验无法度量每个位点对疾病的预测能力，因此我们又基于 WOE 求出 IV 值，根据 IV 值越大，预测能力越强的原则筛选位点。最终，比较二者所筛选位点的结果，发现位点 rs2273298 是疾病 A 最有可能的致病位点。

八．对于问题三，我们首先从位点与疾病相关性和基因频率与疾病相关性两个角度筛选基因，得到 10 个备选基因，为了考察基因与疾病的关联性，我们利用决策树进行分类预测，通过考察混淆矩阵和 ROC 曲线，判断决策树分类的准确率。最终，通过比较 10 个备选基因对于的决策树准确率，我们挑选与疾病关联度较高的 gene102 和 gene55。

九．对于问题四，由于存在 10 个性状观察值下的 1000 个样本，且部分样本之间高度相关，因此首先利用系统聚类的方法，基于样本做聚类分析。根据 R^2 和偏 R^2 参数的选择，最终确认了 7,16,20,50,100 等不同的聚类数。在各聚类类数下，利用 χ^2 统计检验分别考察各位点与综合性状的相关性，得出显著相关的 10 个位点。进一步，为了检验上述 10 个位点的可靠性，我们分别分析 10 个位点与 10 个原始性状间的相关性，结果发现挑选的 10 个位点中的 8 个位点，都与题目所给的 10 个初始性状中的一个或多个性状高度相关。因此，我们认为与 10 个关联性状所有表现出的综合性状相关的位点有 rs12746773，rs4584380，rs11249201，rs12139270，rs2075972，rs1985278，rs6603797，rs10917268。

参考文献

[1].Peter Harrison .Machine Learning in Action [M].北京:人民邮电出版社,2013 .

[2].Kobert I. Kabacoff .R in Action Data Analysis and Graphics with R [M] .北京:人民邮电出版社,2013 .

[3].薛毅、陈立萍编著.R 统计建模与 R 软件[M].北京:清华大学出版社,2011 .

[4].张威等.GAGE 和 GSEA 在基因集研究中的有效性比较[J].现代生物医学进展,2013 .

[5].朱近等.数量性状的基因型值计算[J].生物数学学报,2007,22(1):178-186 .

[6].杨利英等.富集分析框架下的致病SNP位点识别[J].西安电子科技大学学报(自然科学版),2016(3):43-48 .

[7].黄金艳等.基于知识编码的剪切位点预测[J].同济大学学报(自然科学版),2017(11):1548-1561 .

[8].李书超、许进.DNA 序列高维空间数字编码运算法则的进一步结果[J].自然科学进展,2003(6):642-646 .

[9].陈惟昌等.DNA 序列高维空间数字编码的运算法则[J].生物物理学报,2001(3):542-549.

[10].陈惟昌等.遗传密码和 DNA 序列的高维空间数字编码[J].生物物理学报,2000(4):760-768 .

[11].http://blog.csdn.net/kevin7658/article/details/50780391 假阴性:被错误地标记为负元祖的正元组（237）

[12].阮敬、纪宏.实用 SAS 统计分析教程[M].北京:中国统计出版社,2013.

附录

附录 A——全基因组位点筛选排除结果

位点	MIF	HW值	HW-P值	位点	MIF	HW值	HW-P值	位点	MIF	HW值	HW-P值
rs6681946	0.37	16.31	5.4E-05	rs2039631	0.22	8.02	4.6E-03	rs11587046	0.29	7.12	7.6E-03
rs645520	0.15	13.70	2.1E-04	rs4573512	0.41	8.01	4.7E-03	rs4661496	0.18	7.11	7.7E-03
rs9804128	0.24	13.49	2.4E-04	rs10803378	0.13	7.97	4.8E-03	rs3829052	0.19	7.10	7.7E-03
rs271344	0.47	13.00	3.1E-04	rs9426750	0.27	7.93	4.9E-03	rs284235	0.43	7.08	7.8E-03
rs868688	0.21	12.70	3.6E-04	rs547976	0.29	7.89	5.0E-03	rs2999566	0.48	7.02	8.0E-03
rs12124726	0.45	11.97	5.4E-04	rs10799275	0.44	7.87	5.0E-03	rs2821023	0.15	7.02	8.1E-03
rs593993	0.26	11.72	6.2E-04	rs11574	0.16	7.85	5.1E-03	rs13306560	0.38	7.01	8.1E-03
rs10754853	0.44	11.70	6.3E-04	rs11583129	0.28	7.84	5.1E-03	rs11579365	0.39	7.00	8.2E-03
rs761429	0.35	10.99	9.1E-04	rs9438571	0.26	7.80	5.2E-03	rs2294641	0.33	6.99	8.2E-03
rs6702309	0.27	10.71	1.1E-03	rs930851	0.48	7.78	5.3E-03	rs707465	0.14	6.98	8.2E-03
rs10927852	0.20	10.65	1.1E-03	rs12032476	0.41	7.75	5.4E-03	rs11583665	0.48	6.94	8.4E-03
rs4436387	0.28	10.44	1.2E-03	rs9783019	0.11	7.74	5.4E-03	rs909823	0.14	6.93	8.5E-03
rs1076726	0.47	10.15	1.4E-03	rs2661868	0.18	7.66	5.6E-03	rs12033458	0.23	6.92	8.5E-03
rs10917452	0.40	10.10	1.5E-03	rs4654351	0.32	7.61	5.8E-03	rs1997927	0.29	6.92	8.5E-03
rs10493041	0.15	9.97	1.6E-03	rs7534558	0.38	7.61	5.8E-03	rs6429696	0.13	6.91	8.6E-03
rs213022	0.37	9.65	1.9E-03	rs2749152	0.14	7.59	5.9E-03	rs3766192	0.45	6.91	8.6E-03
rs696306	0.48	9.28	2.3E-03	rs6686733	0.17	7.59	5.9E-03	rs12138326	0.12	6.87	8.8E-03
rs4920530	0.20	9.16	2.5E-03	rs12072310	0.14	7.59	5.9E-03	rs3753270	0.15	6.87	8.8E-03
rs3000861	0.42	9.15	2.5E-03	rs950601	0.30	7.58	5.9E-03	rs11800014	0.28	6.86	8.8E-03
rs2050233	0.12	8.94	2.8E-03	rs17352351	0.32	7.57	5.9E-03	rs4515757	0.25	6.83	8.9E-03
rs171317	0.39	8.93	2.8E-03	rs12564136	0.19	7.51	6.2E-03	rs389709	0.23	6.80	9.1E-03
rs16862387	0.50	8.83	3.0E-03	rs753613	0.40	7.49	6.2E-03	rs4501814	0.30	6.77	9.2E-03
rs1815606	0.47	8.72	3.2E-03	rs11587848	0.14	7.49	6.2E-03	rs221035	0.10	6.75	9.4E-03
rs9308476	0.40	8.58	3.4E-03	rs4908591	0.11	7.49	6.2E-03	rs3003434	0.27	6.74	9.4E-03
rs9659997	0.11	8.54	3.5E-03	rs7524800	0.13	7.47	6.3E-03	rs10399665	0.38	6.72	9.5E-03
rs10915445	0.49	8.49	3.6E-03	rs1181878	0.30	7.46	6.3E-03	rs10779685	0.21	6.72	9.5E-03
rs4908547	0.18	8.49	3.6E-03	rs35675666	0.10	7.45	6.3E-03	rs1885865	0.27	6.72	9.6E-03
rs2743201	0.23	8.46	3.6E-03	rs2076977	0.18	7.33	6.8E-03	rs1295101	0.37	6.71	9.6E-03
rs2506088	0.25	8.44	3.7E-03	rs4846048	0.25	7.32	6.8E-03	rs11582960	0.27	6.71	9.6E-03
rs2746479	0.29	8.28	4.0E-03	rs642696	0.17	7.28	7.0E-03	rs7522460	0.48	6.68	9.8E-03
rs6671424	0.19	8.18	4.2E-03	rs2480063	0.23	7.28	7.0E-03	rs2495365	0.30	6.67	9.8E-03
rs17032942	0.32	8.10	4.4E-03	rs4846080	0.11	7.23	7.2E-03	rs2071986	0.42	6.66	9.9E-03
rs17376328	0.18	7.15	7.5E-03								

附录 B:基因决策树图

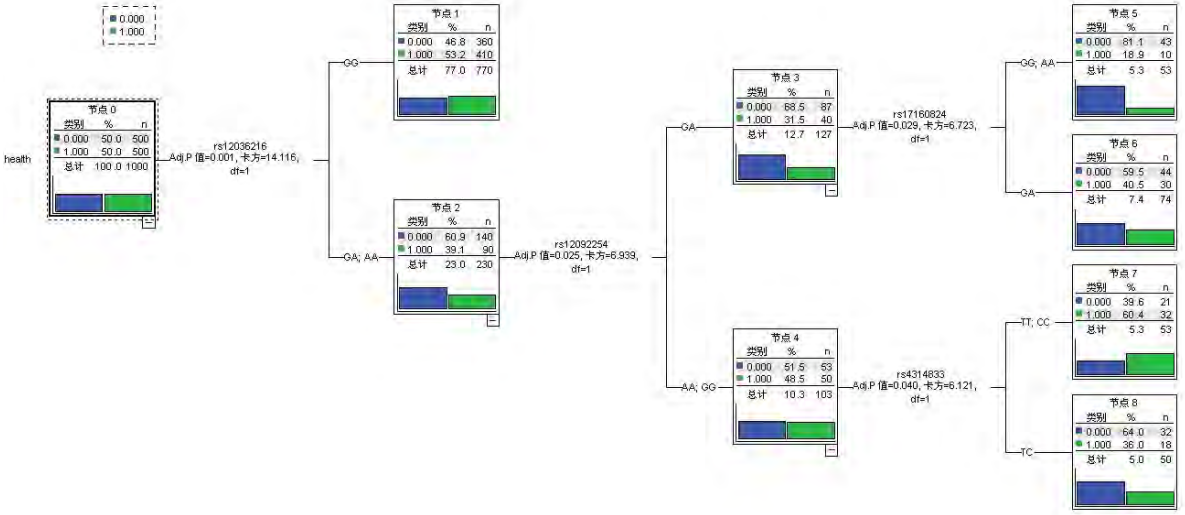


图:gene_3 的决策树图

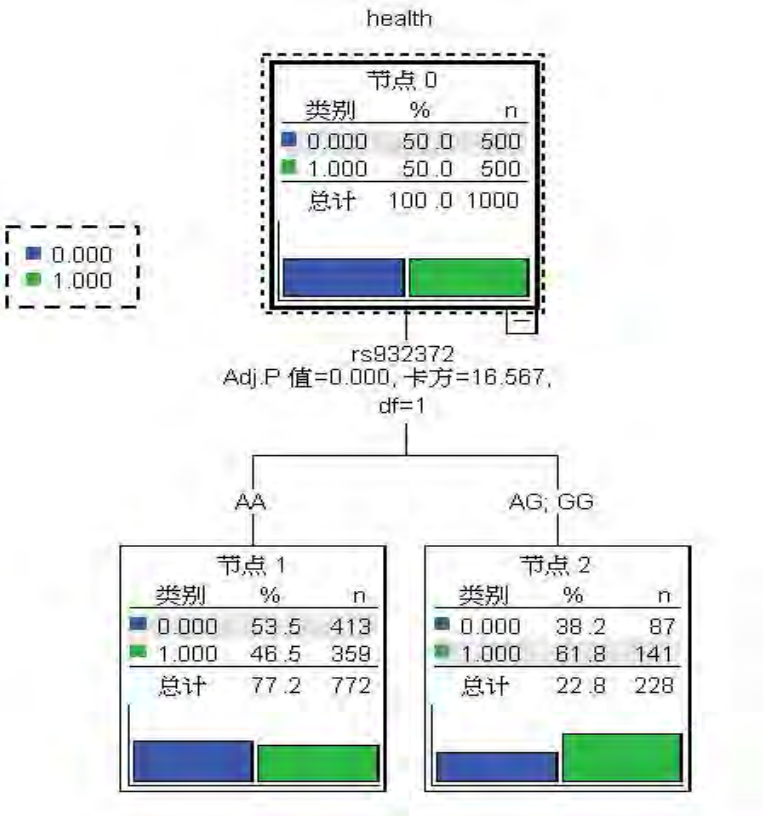


图:Gene_245 决策树示意图

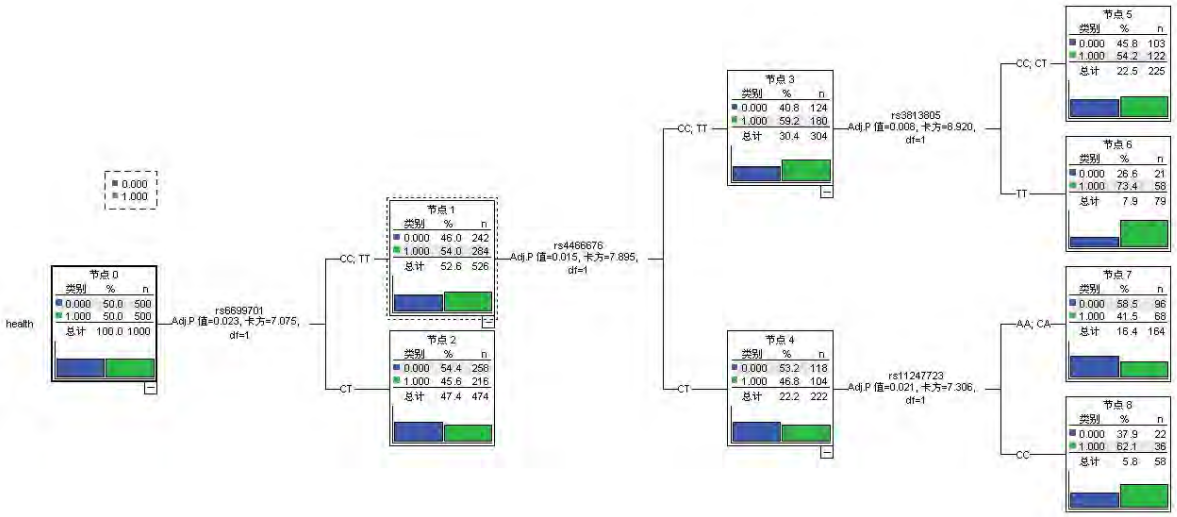


图:Gene_265 决策树示意图

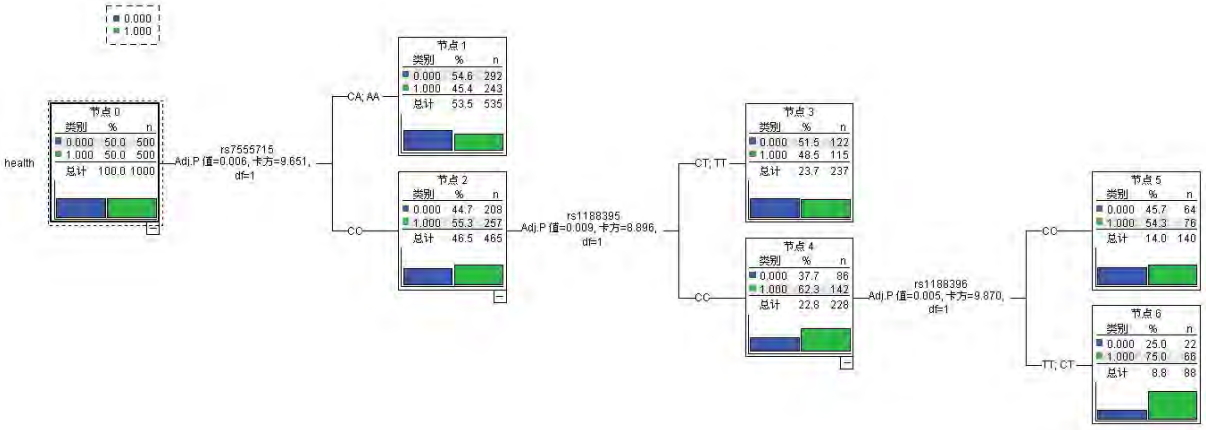


图:gene_293 决策树示意图

附录 C:SAS 代码

```
/*************************************************************************
********/
/*****************************      创建频数表
*****************************/
/*************************************************************************
********/
ods listing close;
ods results off;
ods output
   CrossTabFreqs=freq(rename=(&locus1.=value));
proc freq data=ans.model;
   tables &locus1.*phenotype/;
run;
ods output close;
ods results on;
ods listing;

data ans.freq;
   length locus $20;
   set freq;
   locus="&locus1.";
run;

%macro freq_analysis;
%do i=2 %to 9445;
ods listing close;
ods results off;
ods output
   CrossTabFreqs=freq(rename=(&&locus&i.=value));
proc freq data=ans.model;
   tables &&locus&i.*phenotype;
run;
ods output close;
ods results on;
ods listing;

data temp;
   length locus $20;
   set freq;
   locus="&&locus&i.";
run;
proc append base=ans.freq data=temp force;
```

```
run;
%end;
%mend;
%freq_analysis;

/*************************************************************************
********/
/*************************      WOE、IV 值分析
*************************/
/*************************************************************************
********/
/*WOE 计算*/
/*0:正常，1：患病*/
proc sql noprint;
create table ans.zero as
select
    locus,
    value,
    phenotype,
    Frequency as zero_num label='zero_num',
    ColPercent as zero_distribution label='zero_distribution'
    from ans.freq
    where phenotype=0 and _TYPE_='11';
create table ans.one as
select
    locus,
    value,
    phenotype,
    Frequency as one_num label='one_num',
    ColPercent as one_distribution label='one_distribution'
    from ans.freq
    where phenotype=1 and _TYPE_='11';
quit;

proc sql noprint;
create table ans.woe as
select
    a.locus,
    a.value,
    zero_num,
    one_num,
    zero_distribution,
    one_distribution,
```

```
    log(one_distribution/zero_distribution) as woe,
    one_distribution/zero_distribution as ratio,
    one_distribution-zero_distribution as diff,
    calculated diff2*calculated woe2 as iv_woe_diff
    from ans.zero a
    left join ans.one b
    on a.locus=b.locus and a.value=b.value;
quit;

/*IV 值计算*/
proc sql noprint;
create table ans.iv as
select
    locus,
    sum(iv_woe_diff) as iv_diff,
    from ans.woe
    group by locus;
quit;

proc means data=ans.iv n max p99 p95 mean p5 p1 min  mode var std skew kurt;
    title 'IV 值 描述性统计分析';
    var _numeric_;
run;

proc univariate data=ans.iv;
    title "IV 值 分布直方图";
    histogram iv_diff;
run;



/*************************************************************************
********/
/*********************        统计分析(卡方检验、费雪精确检验)
**************/
/*************************************************************************
*******/
%macro freq_statistical_test;
%do i=1 %to 9445;
ods listing close;
ods results off;
ods output
    ChiSq=ChiSq(drop=Table)
```

```
    FishersExact=FishersExact(drop=Table)
    Measures=Measures(drop=Table);
proc freq data=ans.model_updated;
    tables phenotype*&&locus&i./noprint chisq exact measures;
run;
ods output close;
ods results on;
ods listing;

data freq_chisq;
    length locus $20;
    set ChiSq;
    locus="&&locus&i.";
data freq_fishersexact;
    length locus $20;
    set FishersExact;
    locus="&&locus&i.";
data freq_measures;
    length locus $20;
    set Measures;
    locus="&&locus&i.";
run;

proc append base=ans.freq_chisq        data=freq_chisq        force;
proc append base=ans.freq_fishersexact data=freq_fishersexact force;
proc append base=ans.freq_measures     data=freq_measures     force;
run;
%end;
%mend;
%freq_statistical_test;


/*************************************************************************
********/
/*****************                          聚类分析
***************/
/*************************************************************************
*******/
/*类平均法进行系统聚类*/
ods listing close;
ods results off;
ods output
    ClusterHistory=ans.cluster_ClusterHistory
```

```
    AvDist=ans.cluster_AvDist
     RMSStd= ans.cluster_RMSStd
     EigenvalueTable=ans.cluster_EigenvalueTable;
    proc cluster data=ans.multi_phenos_id method=ave outtree=ans.cluster_tree;
        var phenotype1-phenotype10;
        id id;
    run;
ods output close;
ods results on;
ods listing;

/*指定聚类数进行快速聚类*/
%macro fastclus_analysis(i=);
ods listing close;
ods results off;
ods output
    InitialSeeds=clus.InitialSeeds_&i.
    Criterion=clus._Criterion_&i.
    ClusterSum=clus.ClusterSum_&i.
    VariableStat=clus.VariableStat_&i.
    PseudoFStat=clus.PseudoFStat_&i.
    ApproxExpOverAllRSq=clus.ApproxExpOverAllRSq_&i.
    CCC=clus.CCC_&i.
    ClusterCenters=clus.ClusterCenters_&i.
    ClusterDispersion=clus.ClusterDispersion&i.;
proc fastclus data=ans.multi_phenos_id
    maxclusters=&i. list out=clus.fastclus_&i.;
    var phenotype1-phenotype10;
    id id;
run;
ods listing;
ods results on;
%mend;

%fastclus_analysis(i=100);
%fastclus_analysis(i=50);
%fastclus_analysis(i=20);
%fastclus_analysis(i=16);
%fastclus_analysis(i=13);
%fastclus_analysis(i=10);
%fastclus_analysis(i=7);
%fastclus_analysis(i=4);
%fastclus_analysis(i=3);
```

```
%fastclus_analysis(i=2);

/*****************************************************************************
********/
/************************        基于单性状频数分析
************************/
/*****************************************************************************
********/

%macro freq_super_analysis;
%do k=10 %to 10;
ods listing close;
ods results off;
ods output
   ChiSq=ChiSq_phenotype&k.;
proc freq data=ans.model2;
   tables phenotype&k.*&locus1./chisq noprint;
run;
ods results on;
ods listing;

data ans.chisq_phenotype&k.;
   length locus $20;
   set ChiSq_phenotype&k.;
   locus="&locus1.";
run;

%do i=2 %to 9445;
ods listing close;
ods results off;
ods output
   ChiSq=ChiSq_phenotype&k.;
proc freq data=ans.model2;
   tables phenotype&k.*&&locus&i./chisq noprint;
run;
ods output close;
ods results on;
ods listing;

data temp2;
   length locus $20;
   set ChiSq_phenotype&k.;
   locus="&&locus&i.";
```

```
run;
proc append base=ans.chisq_phenotype&k. data=temp2 force;run;
%end;
%end;
%mend;

%freq_super_analysis;

%macro chisq_phenotype_locus;
%do i=1 %to 10;
proc sql noprint;
create table chisq_phenotype&i._locus as
select
    locus,
    Value as chisq_phenotype&i. label="chisq_phenotype&i.",
    Prob as P_phenotype&i. label="P_phenotype&i."
    from ans.chisq_phenotype&i.
    where Statistic='卡方';
quit;
%end;
%mend;
%chisq_phenotype_locus;


/*************************************************************************
********/
/**********************        基于聚类综合性状频数分析
**********************/
/*************************************************************************
********/
%macro freq_analysis;
%do i=1 %to 9445;
ods listing close;
ods results off;
ods output
    ChiSq=ChiSq_clus3;
proc freq data=ans.model2_clus;
    tables clus3*&&locus&i./chisq noprint;
run;
ods output close;
ods results on;
ods listing;
```

```
data temp2;
    length locus $20;
    set ChiSq_clus3;
    locus="&&locus&i.";
run;
proc append base=ans.chisq_clus3 data=temp2 force;run;
%end;
%mend;

%freq_analysis;
```

Rcode:

R code:

```
setwd("H:/2016 试题/B/B 题附件")
mydata=read.table("H:/2016 试题/B/B 题附件/genotype.dat",1)
health=read.table("H:/2016 试题/B/B 题附件/phenotype.txt")
colnames(health)="health"
mydata=as.matrix(mydata)
for (j in c(6070,8473)){
for (i in 1:1000){
  if (mydata[i,6070]=="II") mydata[i,6070]="TT"
  else if (mydata[i,6070]=="ID") mydata[i,6070]="TC"
  else if (mydata[i,6070]=="DD") mydata[i,6070]="CC"
  }
}
mydata=data.frame(mydata)
mydata_0=mydata[1:500,]
mydata_1=mydata[501:1000,]
Obs=matrix(nrow = 3,ncol = 9445)
for (i in 1:9445){
  Obs[,i]=t(levels(mydata[,i]))
  }
Fre_col_0=matrix(nrow=3,ncol=9445)
for (j in 1:3){
  for (h in 1:9445){
  Fre_col_0[j,h]=length(which(mydata_0[,h]==Obs[j,h]))/500
  }
}
Fre_col_1=matrix(nrow=3,ncol=9445)
for (j in 1:3){
  for (h in 1:9445){
    Fre_col_1[j,h]=length(which(mydata_1[,h]==Obs[j,h]))/500
  }
```

```
}
Fre_row_0=matrix(nrow=3,ncol=9445)
for (j in 1:3){
  for (h in 1:9445){
    Fre_row_0[j,h]=length(which(mydata_0[,h]==Obs[j,h]))/length(which(mydata[,h]==Obs[j,h]))
  }
}
Fre_row_1=matrix(nrow=3,ncol=9445)
for (j in 1:3){
  for (h in 1:9445){
    Fre_row_1[j,h]=length(which(mydata_1[,h]==Obs[j,h]))/length(which(mydata[,h]==Obs[j,h]))
  }
}
WOE=matrix(nrow = 3,ncol = 9445)
for (i in 1:3){
  for (j in 1:9445){
    WOE[i,j]=log(Fre_col_1[i,j]/Fre_col_0[i,j])
  }
}
IV=matrix(nrow = 1,ncol = 9445)
for (i in 1:9445){
  IV[,i]=WOE[,i]%*%(Fre_col_1[,i]/Fre_col_0[,i])
}
IV_1=which(IV>=10,arr.ind = T)
names(IV)=colnames(mydata)



library(rpart)
library(rpart.plot)
library(sandwich)
library(grid)
library(mvtnorm)
library(stats4)
library(modeltools)
library(zoo)
library(strucchange)
library(party)
library(RWeka)
library(partykit)
setwd("H:/2016 试题/B/B 题附件")
mydata=read.table("H:/2016 试题/B/B 题附件/genotype.dat",1)
health=read.table("H:/2016 试题/B/B 题附件/phenotype.txt")
mydata_names=colnames(mydata)
```

```
colnames(health)="health"
Obs=matrix(nrow = 3,ncol = 9445)
for (i in 1:9445){
   Obs[,i]=t(levels(mydata[,i]))
}
mydata=as.matrix(mydata)
for (i in 1:1000){
  for (j in 1:9445){
       if (mydata[i,j]==Obs[1,j])
          mydata[i,j]=0
       else if (mydata[i,j]==Obs[2,j])
          mydata[i,j]=1
       else if (mydata[i,j]==Obs[3,j])
          mydata[i,j]=2
       else
          mydata[i,j]=3
   }
}
health=as.matrix(health)
health=as.factor(health)
mydata=matrix(as.numeric(mydata),nrow = nrow(mydata))
mydata=as.data.frame(mydata)
names(mydata)=mydata_names
mydata=cbind(health,mydata)
ct=rpart.control(minsplit = 20,xval = 10,cp=0.02)
Fit=rpart(health~.,data=mydata,method="class",control=ct,parms = list(split="gini"))
plot(Fit,uniform = T,branch = 0,margin = 0.1,main="Classification Tree")
text(Fit,use.n = T,fancy = F,col="blue")
mydata_ctree=ctree(health~.,data = mydata)
plot(mydata_ctree, type="simple")
gene_1_data=mydata[,1:8]
gene_1_C45tree=J48(health~.,data =gene_1_data,control = Weka_control(R = T, M = 10))
plot(gene_1_C45tree)
gene_265_data=cbind(health,mydata[,8339:8394])
gene_265_C45tree=J48(health~.,data =gene_265_data,control = Weka_control(R = T, M = 15))
plot(gene_265_C45tree)
```