# 具有遗传性疾病和性状的遗传位点分析

田一丁, 曹赟, 蔡涵

上海交通大学APEX实验室

{killa, , }@apex.sjtu.edu.cn

## 摘要

统计遗传学是一门随着计算机发展及机器学习技术发展而兴起的学科，主要研究领域是分析染色体、基因及位点等遗传信息载体对某些遗传性状或遗传疾病的影响。机器学习技术在此领域具有重要的作用，例如，利用机器学习技术，训练根据样本的基因信息，预测样本患病概率的模型，或者将基因信息视为特征，通过特征选择方法判断某些疾病的致病基因，达到控制疾病的目的。本文提出了一种利用统计学及机器学习方法，筛选对遗传性状和疾病有影响的位点和基因的方法，并对此方法进行了评估。

**关键词：**遗传位点，遗传疾病，统计遗传学，机器学习，特征选择

## 背景介绍

我们已经知道人类的遗传性状由染色体上的DNA分子携带，DNA分子是一种长链双螺旋结构，其包含许许多多的碱基对，这些碱基对由A、T、C、G四种碱基构成。人类在长期繁衍中会发生一些变异或者突变，造成性状的多样性。DNA上有一些碱基对容易发生突变，这些碱基对被称为位点。

即使人类早在一百多年前就开始对遗传学进行探索和研究，但是至今我们对DNA分子，基因这些东西所知仍然甚少。现如今，随着计算机技术的不断发展，计算速度的不断上升以及统计学习的发展，生物遗传学进入了一个新的发展领域——统计遗传学。从大量的病例样本中学习哪些是引起某些疾病的致病基因是统计遗传学中一个重要的课题。

位点在DNA中占到大约1/1000，出现的频率比较高，每个位点变异的方式也有多种，所以近年来被研究人员视为重要的研究对象。大量的研究也表明位点或者包含很多位点的基因与一些遗传病的有无十分相关。位点通常由两个碱基表示，例如TT、CC等，一个给定的位点最多可能有三种编码。

## 题目介绍及分析

本题目是在遗传统计学的背景下衍生的题目，共有四个小题目，已知某种遗传病，给定一些患病样本和健康样本，以及这些样本某些位点的碱基编码，要求分析位点们与遗传病的关系。

第一题要求我们将位点的碱基编码转化为可计算的数值编码；第二题要求我们分析某个或某些位点与这种遗传病相关；第三题将所有位点进行划分，分为不同的一些基因，要求我们分析遗传病与基因的关系；第四题给出十种症状，要求我们分析这些症状与位点之间的关联。

题目中给出的数据包含1000个样本，其中包含500个健康样本和500个患病样本。每个样本给出其9445个位点的编码信息，以及十种性状信息。同时给出300个基因以及这些基因包含的位点。数据中的每个位点都有一个唯一的名称。

我们经过分析后发现，由于样本的患病与否可以用0或1来表示，这十分符合机器学习中二分类问题的输出，同时1000个样本可以作为训练数据和检验数据，故这个问题可以被看作一个机器学习问题，并且是一个监督学习问题（**Supervised Learning**）。每个样本的位点们可以看作这个样本的特征（**Feature**），分析哪些位点与疾病相关即分析哪些特征对最终结果的输出会产生影响，所以实际上这是一个特征选择（**Feature Selection**）问题。直观的选择思路便是选择那些使用后能够明显提升训练效果的特征。

## 解题思路

发现了问题的本质之后我们就可以着手解决问题了。在这之前需要确定使用的编程语言以及工具，综合熟练程度、运行速度、现有工具丰富度等方面因素之后，我们决定使用Python作为我们的编程语言，并使用其**scikit-learn**工具包进行机器学习。

#### 问题一：碱基编码

目前生物学中以及本题给出的数据中都是使用A、T、G、C四个字母对碱基进行编码，这种编码方式在碱基表示及描述中起到了很好的作用，但是文字编码并不能用于计算，所以我们需要一种利于计算的数值编码方式来对碱基进行表述。由于每个位点的碱基对有且只有三种可能的值，所以一种比较直观的编码方法是使用1、2、3，三个数字来表示三种可能的值，但是这种编码方式存在一个致命的问题：这种编码下三种碱基对的地位并不是相同的，比如2所对应的碱基对在数值上是1所对应的碱基对的两倍，这是不科学的。另外这种编码方式导致三种可能的碱基对在它们的空间（一维空间）上两两之间的距离不相等，然而作为特征，这三种碱基对之间应该没有距离差。

在简单对比了几种编码方式之后，我们最终选择了使用三维one-hot向量表示碱基对。one-hot向量是机器学习中一种经典的数据表示方式，向量中的每一维只有0和1两种取值，一个向量中只能有一维的值为1，其他值均为零。在这里，向量中的一维代表碱基对的一个可能取值，如果向量中某一位为1，表示碱基对的取值是这个1所对应的取值。比如某个位点的碱基对可能有三种取值：TT、TC、CC，则构建一个三维one-hot向量，如果碱基对的取值是TT，则其对应的one-hot为 (1, 0, 0)；TC对应 (0, 1, 0)，依次类推。这样三种碱基对应的值没有数值上的差别，并且在三维空间中它们两两之间的距离相等。

#### 问题二/三：分析遗传病与位点及基因的关系

对于问题二与问题三，由于其输出（是否患病）只有两种取值：1或者0，故我们将其视作一个分类问题。常用的监督学习的分类器有逻辑回归（**Logistic Regression**）、支持向量机（**Support Vector Machine**）、决策树分类器（**Decision Tree Classifier**）、随机森林分类器（**Random Forest Classifier**）、梯度推进分类器（**Gradient Boosting Classifier**）等。

逻辑回归分类器是一种比较简单的线性分类器，它为每一个输入训练一个权重，加权求和之后使用一个**sigmoid**函数进行分类，**sigmoid**函数一般使用 或者tanh函数。

支持向量机分类器是一种多种维度适用的分类器模型，这得益于支持向量机的核函数思想，它的原理是将数据映射到一个高维空间中，并使用一个高维空间中的超平面来分类。在迭代过程中支持向量机不断寻找与两边数据点的间隔最大的超平面。

决策树分类器是一种树模型分类器，其中树的每个决策节点由一些简单的决策规则构成，可以依靠计算条件概率的方法来构建。

随机森林分类器是一个决策树的集合，其中每个决策树使用从所有特征中随机挑选的特征来对决策树进行训练，分类时使用决策森林分类结果中的众数作为最终的分类结果。

梯度推进分类器实际上是一种框架，它会在多次迭代中训练许多效果较差的分类器，每次迭代都会给上次迭代结果较差的训练样本加上更高的权重，最后将这些分类器组合（加权求和或投票）后成为一个效果较好的分类器，在训练新的分类器时，以残差减少的梯度方向进行。

在实验过程中，我们分别尝试了这几种分类器模型并比较了它们的结果（见第四章），由于随机森林分类器训练的效率太低，而决策树分类器的效果欠佳，故最终放弃了这两种模型。

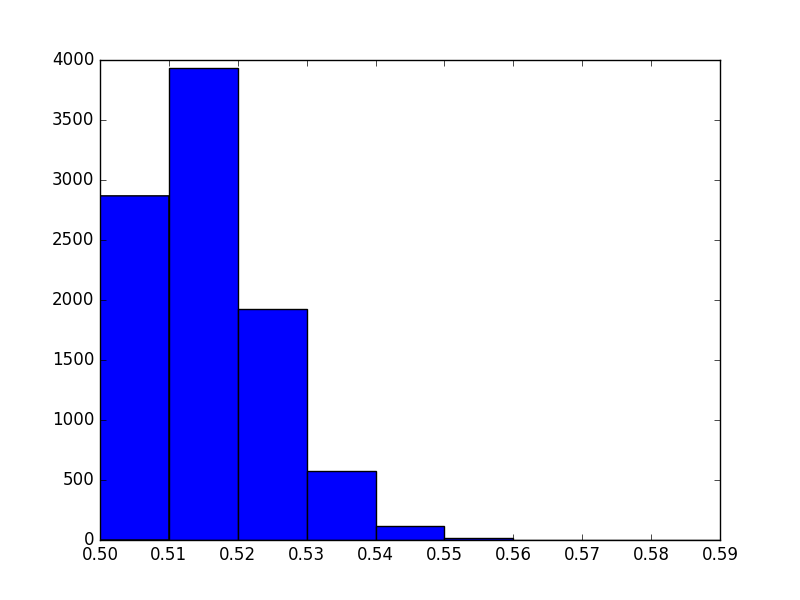
我们首先将所有位点都作为特征进行训练，在测试集上进行测试后发现准确度很低，因此如果在所有位点的基础上进行特征选择和排除，一来由于特征数量过于庞大，二来由于准确度很低，移除一些特征可能会导致结果变化并不明显，故考虑在初期先对特征进行筛选。我们针对每个特征单独训练了逻辑回归分类器，并对它们的测试结果进行了排序，由此选取单独条件下对结果影响最大的30个位点作为初步筛选特征的结果。

在进一步的特征筛选中，我们使用了递归特征消除（**Recursive Feature Elimination**）的方法，这是一种特征消除的框架，其思想为在迭代过程中不断剔除不重要的特征（具体表现为特征对应的权重较小）。具体方法是在每一轮迭代过后，排除对应权重最小的一个或几个特征，然后重新进行训练进入下一次迭代，直到特征数量达到要求的值为止。如果不确定最优的特征数量，可以结合交叉验证寻找最优的特征数量。递归特征消除这种方法是一种从模型顶层消除特征的方法，其弱点在于它的效果对模型的稳定性依赖较强，所在模型设计的过程中一定要进行正则化操作，防止过拟合。

使用递归特征消除的方法，我们成功筛选出了16个与遗传病相关的位点和【多少】个基因。（见第四章）

#### 问题四：分析不同性状与位点关系

## 实验结果及分析



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Score 30 | Score 29 | Score 28 | Score 27 | Score 26 | Score 25 | Score 24 |
| Feature 0 | 0.66 | 0.649 | 0.663 | 0.665 | 0.674 | 0.675 | 0.67 |
| Feature 1 | 0.66 | 0.651 | 0.644 | 0.654 | 0.653 | 0.656 | 0.661 |
| Feature 2 | 0.66 | 0.656 | 0.665 | 0.668 | 0.675 | **0.682** |  |
| Feature 3 | 0.66 | 0.662 | 0.663 | 0.67 | 0.67 | 0.671 | 0.679 |
| Feature 4 | 0.66 | 0.664 | 0.673 | 0.673 | 0.668 | 0.668 | 0.664 |
| Feature 5 | 0.66 | 0.654 | 0.651 | 0.658 | 0.651 | 0.655 | 0.643 |
| Feature 6 | 0.66 | 0.652 | 0.655 | 0.653 | 0.656 | 0.662 | 0.661 |
| Feature 7 | 0.66 | 0.662 | 0.663 | 0.666 | 0.67 | 0.671 | 0.668 |
| Feature 8 | 0.66 | 0.661 | 0.656 | 0.659 | 0.661 | 0.657 | 0.666 |
| Feature 9 | 0.66 | 0.655 | 0.659 | 0.661 | 0.667 | 0.673 | 0.674 |
| Feature 10 | 0.66 | 0.646 | 0.665 | 0.666 | 0.665 | 0.665 | 0.672 |
| Feature 11 | 0.66 | 0.665 | 0.668 | **0.678** |  |  |  |
| Feature 12 | 0.66 | 0.656 | 0.664 | 0.668 | 0.667 | 0.671 | 0.666 |
| Feature 13 | 0.66 | 0.656 | 0.665 | 0.664 | 0.663 | 0.664 | 0.668 |
| Feature 14 | 0.66 | 0.662 | 0.662 | 0.663 | 0.666 | 0.66 | 0.663 |
| Feature 15 | 0.66 | 0.649 | 0.655 | 0.661 | 0.668 | 0.663 | 0.661 |
| Feature 16 | 0.66 | 0.646 | 0.656 | 0.661 | 0.661 | 0.657 | 0.662 |
| Feature 17 | 0.66 | 0.661 | **0.675** |  |  |  |  |
| Feature 18 | 0.66 | 0.658 | 0.668 | 0.675 | 0.677 | 0.673 | 0.676 |
| Feature 19 | 0.66 | 0.663 | 0.666 | 0.676 | 0.665 | 0.662 | 0.665 |
| Feature 20 | 0.66 | 0.658 | 0.654 | 0.657 | 0.662 | 0.658 | 0.65 |
| Feature 21 | 0.66 | 0.657 | 0.662 | 0.67 | 0.66 | 0.659 | 0.656 |
| Feature 22 | 0.66 | 0.655 | 0.664 | 0.668 | **0.679** |  |  |
| Feature 23 | 0.66 | 0.657 | 0.67 | 0.667 | 0.675 | 0.673 | **0.681** |
| Feature 24 | 0.66 | 0.664 | 0.669 | 0.673 | 0.662 | 0.668 | 0.665 |
| Feature 25 | 0.66 | 0.656 | 0.665 | 0.674 | 0.669 | 0.664 | 0.679 |
| Feature 26 | 0.66 | 0.661 | 0.664 | 0.669 | 0.67 | 0.665 | 0.665 |
| Feature 27 | 0.66 | 0.665 | 0.669 | 0.671 | 0.67 | 0.671 | 0.674 |
| Feature 28 | 0.66 | **0.668** |  |  |  |  |  |
| Feature 29 | 0.66 | 0.657 | 0.671 | 0.675 | 0.665 | 0.673 | 0.669 |

表 LR-l1模型递归特征消除迭代准确率

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Score 30 | Score 29 | Score 28 | Score 27 | Score 26 | Score 25 |
| Feature 0 | 0.709 | 0.707 | 0.68 | 0.687 | 0.685 | 0.687 |
| Feature 1 | 0.709 | 0.697 | 0.694 | 0.696 | 0.699 | 0.692 |
| Feature 2 | 0.709 | 0.718 | 0.715 | 0.718 | 0.725 | 0.722 |
| Feature 3 | 0.709 | 0.706 | 0.704 | 0.709 | 0.708 | 0.707 |
| Feature 4 | 0.709 | 0.715 | 0.716 | 0.721 | 0.724 | 0.724 |
| Feature 5 | 0.709 | 0.706 | 0.721 | 0.727 | **0.733** |  |
| Feature 6 | 0.709 | 0.708 | 0.717 | 0.713 | 0.717 | 0.721 |
| Feature 7 | 0.709 | 0.713 | 0.709 | 0.721 | 0.731 | 0.71 |
| Feature 8 | 0.709 | 0.713 | 0.703 | 0.706 | 0.709 | 0.71 |
| Feature 9 | 0.709 | 0.716 | 0.723 | 0.714 | 0.716 | 0.718 |
| Feature 10 | 0.709 | 0.706 | 0.714 | 0.712 | 0.719 | 0.724 |
| Feature 11 | 0.709 | 0.715 | 0.723 | 0.72 | 0.722 | 0.726 |
| Feature 12 | 0.709 | 0.702 | 0.716 | 0.725 | 0.726 | 0.718 |
| Feature 13 | 0.709 | 0.713 | 0.707 | 0.704 | 0.713 | 0.713 |
| Feature 14 | 0.709 | 0.7 | 0.721 | 0.728 | 0.728 | **0.73** |
| Feature 15 | 0.709 | 0.711 | 0.708 | 0.711 | 0.716 | 0.709 |
| Feature 16 | 0.709 | 0.695 | 0.701 | 0.706 | 0.71 | 0.697 |
| Feature 17 | 0.709 | **0.721** |  |  |  |  |
| Feature 18 | 0.709 | 0.713 | 0.713 | 0.72 | 0.727 | 0.725 |
| Feature 19 | 0.709 | 0.704 | 0.721 | 0.713 | 0.716 | 0.715 |
| Feature 20 | 0.709 | 0.708 | 0.718 | 0.725 | 0.728 | 0.724 |
| Feature 21 | 0.709 | 0.695 | 0.705 | 0.704 | 0.706 | 0.711 |
| Feature 22 | 0.709 | 0.714 | 0.723 | **0.729** |  |  |
| Feature 23 | 0.709 | 0.707 | 0.702 | 0.711 | 0.712 | 0.71 |
| Feature 24 | 0.709 | 0.711 | **0.726** |  |  |  |
| Feature 25 | 0.709 | 0.706 | 0.709 | 0.708 | 0.709 | 0.71 |
| Feature 26 | 0.709 | 0.698 | 0.707 | 0.706 | 0.709 | 0.715 |
| Feature 27 | 0.709 | 0.703 | 0.714 | 0.722 | 0.73 | 0.726 |
| Feature 28 | 0.709 | 0.712 | 0.709 | 0.715 | 0.723 | 0.715 |
| Feature 29 | 0.709 | 0.716 | 0.701 | 0.711 | 0.711 | 0.712 |

表 LR-l2模型递归特征消除迭代准确率

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Score 30 | Score 29 | Score 28 | Score 27 | Score 26 |
| Feature 0 | 0.662 | 0.65 | 0.66 | 0.66 | 0.662 |
| Feature 1 | 0.662 | 0.644 | 0.65 | 0.641 | 0.633 |
| Feature 2 | 0.662 | 0.665 | 0.683 | 0.683 | 0.683 |
| Feature 3 | 0.662 | 0.666 | 0.665 | 0.67 | 0.663 |
| Feature 4 | 0.662 | 0.667 | 0.672 | 0.668 | 0.682 |
| Feature 5 | 0.662 | 0.658 | 0.674 | 0.669 | 0.68 |
| Feature 6 | 0.662 | 0.672 | 0.688 | 0.683 | 0.678 |
| Feature 7 | 0.662 | 0.654 | 0.643 | 0.641 | 0.655 |
| Feature 8 | 0.662 | 0.665 | 0.679 | 0.667 | 0.675 |
| Feature 9 | 0.662 | 0.668 | 0.678 | 0.689 | 0.68 |
| Feature 10 | 0.662 | 0.677 | 0.689 | 0.673 | 0.683 |
| Feature 11 | 0.662 | 0.665 | 0.674 | 0.673 | 0.682 |
| Feature 12 | 0.662 | 0.675 | 0.672 | 0.673 | 0.68 |
| Feature 13 | 0.662 | 0.657 | 0.683 | 0.672 | 0.668 |
| Feature 14 | 0.662 | **0.684** |  |  |  |
| Feature 15 | 0.662 | 0.658 | 0.683 | 0.682 | 0.684 |
| Feature 16 | 0.662 | 0.67 | 0.684 | 0.68 | 0.66 |
| Feature 17 | 0.662 | 0.66 | 0.672 | 0.674 | 0.682 |
| Feature 18 | 0.662 | 0.663 | 0.681 | 0.685 | 0.68 |
| Feature 19 | 0.662 | 0.675 | 0.677 | 0.677 | 0.673 |
| Feature 20 | 0.662 | 0.669 | 0.671 | 0.677 | **0.687** |
| Feature 21 | 0.662 | 0.677 | 0.671 | 0.674 | 0.681 |
| Feature 22 | 0.662 | 0.652 | 0.658 | 0.657 | 0.643 |
| Feature 23 | 0.662 | 0.664 | 0.68 | **0.691** |  |
| Feature 24 | 0.662 | 0.669 | **0.689** |  |  |
| Feature 25 | 0.662 | 0.654 | 0.671 | 0.681 | 0.676 |
| Feature 26 | 0.662 | 0.653 | 0.684 | 0.669 | 0.685 |
| Feature 27 | 0.662 | 0.673 | 0.674 | 0.669 | 0.666 |
| Feature 28 | 0.662 | 0.66 | 0.68 | 0.686 | 0.664 |
| Feature 29 | 0.662 | 0.673 | 0.685 | 0.677 | 0.675 |

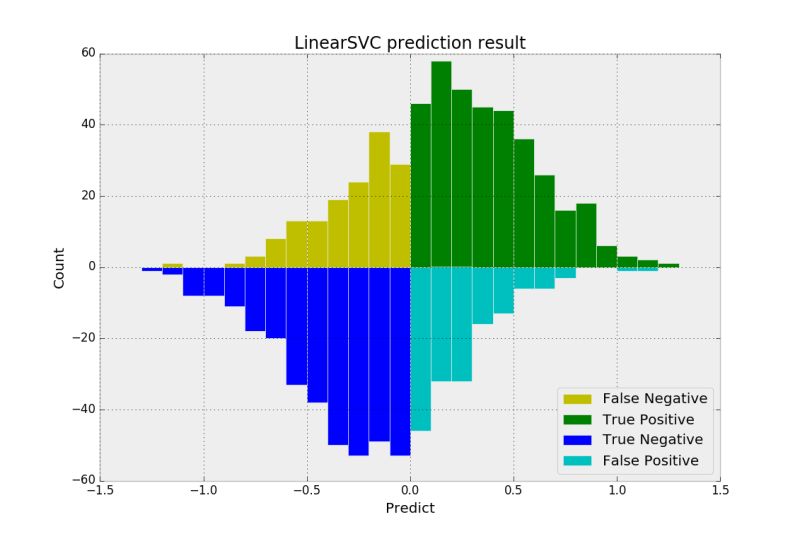
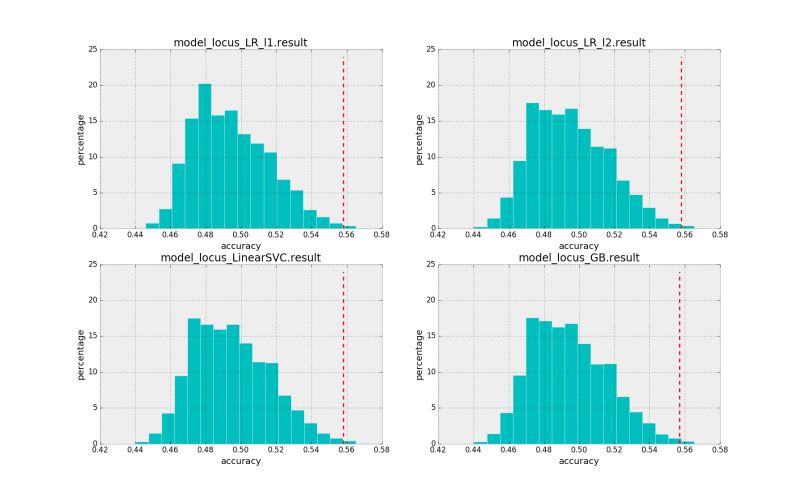
表 Gradiant Boosting模型递归特征消除迭代准确率

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Score 30 | Score 29 | Score 28 | Score 27 | Score 26 | Score 25 |
| Feature 0 | 0.726 | 0.708 | 0.683 | 0.701 | 0.697 | 0.699 |
| Feature 1 | 0.726 | 0.693 | 0.689 | 0.698 | 0.709 | 0.704 |
| Feature 2 | 0.726 | 0.726 | 0.727 | 0.728 | 0.727 | 0.729 |
| Feature 3 | 0.726 | 0.708 | 0.713 | 0.72 | 0.715 | 0.712 |
| Feature 4 | 0.726 | 0.71 | 0.709 | 0.716 | 0.715 | 0.719 |
| Feature 5 | 0.726 | **0.729** |  |  |  |  |
| Feature 6 | 0.726 | 0.717 | 0.718 | 0.73 | 0.721 | **0.73** |
| Feature 7 | 0.726 | 0.72 | 0.713 | 0.718 | 0.714 | 0.714 |
| Feature 8 | 0.726 | 0.716 | 0.707 | 0.717 | 0.72 | 0.715 |
| Feature 9 | 0.726 | 0.723 | 0.72 | 0.723 | 0.726 | 0.725 |
| Feature 10 | 0.726 | 0.724 | **0.73** |  |  |  |
| Feature 11 | 0.726 | 0.718 | 0.714 | 0.721 | 0.721 | 0.715 |
| Feature 12 | 0.726 | 0.723 | 0.718 | 0.719 | 0.732 | 0.728 |
| Feature 13 | 0.726 | 0.719 | 0.724 | 0.729 | 0.723 | 0.725 |
| Feature 14 | 0.726 | 0.713 | 0.71 | 0.72 | 0.719 | 0.719 |
| Feature 15 | 0.726 | 0.708 | 0.701 | 0.709 | 0.713 | 0.709 |
| Feature 16 | 0.726 | 0.714 | 0.717 | 0.716 | 0.727 | 0.724 |
| Feature 17 | 0.726 | 0.71 | 0.718 | 0.718 | 0.722 | 0.72 |
| Feature 18 | 0.726 | 0.715 | 0.726 | 0.722 | 0.732 | 0.723 |
| Feature 19 | 0.726 | 0.717 | 0.71 | 0.714 | 0.716 | 0.713 |
| Feature 20 | 0.726 | 0.728 | 0.727 | 0.727 | **0.734** |  |
| Feature 21 | 0.726 | 0.711 | 0.707 | 0.71 | 0.726 | 0.72 |
| Feature 22 | 0.726 | 0.721 | 0.73 | 0.725 | 0.727 | 0.725 |
| Feature 23 | 0.726 | 0.707 | 0.709 | 0.717 | 0.713 | 0.714 |
| Feature 24 | 0.726 | 0.706 | 0.709 | 0.717 | 0.718 | 0.714 |
| Feature 25 | 0.726 | 0.722 | 0.72 | 0.726 | 0.728 | 0.728 |
| Feature 26 | 0.726 | 0.724 | 0.725 | **0.733** |  |  |
| Feature 27 | 0.726 | 0.716 | 0.71 | 0.714 | 0.722 | 0.716 |
| Feature 28 | 0.726 | 0.724 | 0.723 | 0.722 | 0.731 | 0.73 |
| Feature 29 | 0.726 | 0.722 | 0.727 | 0.726 | 0.723 | 0.721 |

表 Linear SVC模型递归特征消除迭代准确率

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Label 1 | Label 2 | Label 3 | Label 4 | Label 5 | Label 6 | Label 7 | Label 8 | Label 9 | Label 10 |
| Label 1 | 0 | 429 | 435 | 427 | 421 | 420 | 435 | 428 | 436 | 432 |
| Label 2 | 0 | 0 | 428 | 422 | 420 | 423 | 429 | 433 | 434 | 437 |
| Label 3 | 0 | 0 | 0 | 425 | 431 | 423 | 431 | 432 | 428 | 437 |
| Label 4 | 0 | 0 | 0 | 0 | 421 | 419 | 430 | 433 | 430 | 428 |
| Label 5 | 0 | 0 | 0 | 0 | 0 | 437 | 420 | 427 | 417 | 423 |
| Label 6 | 0 | 0 | 0 | 0 | 0 | 0 | 421 | 427 | 419 | 417 |
| Label 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 434 | 430 | 440 |
| Label 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 432 | 440 |
| Label 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 436 |
| Label 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

表 co-positive count table



## 总结

本文分析了遗传性疾病和性状的遗传位点分析问题的本质，提出此问题可映射为机器学习问题中的特征选择问题。由此对问题建模，以位点（或者基因）作为机器学习问题中的特征，样本的性状或获病与否作为分类输出，尝试并比较了多种分类模型，利用机器学习的相关方法对遗传病及性状相关的位点和基因进行了筛选和分析，提出了一种分析人类DNA分子结构和表现性状对应关系的方法，并达到了较好的效果。

## 不足与展望

由于时间紧迫，很多想法都来不及去尝试。首先位点的编码方式使得训练数据的维度大幅增加，增大了计算量，可能有更好的方法在不增加维度的情况下对位点碱基对进行编码。另外我们的方法现阶段主要还是使用了逻辑回归这种简单的模型，如果使用更加复杂一点的模型或者更加个性化的模型可能会达到更好的效果。其次，初步筛选特征的过程模型比较简单，粒度较粗糙，在后期研究中如果使用更加细粒度的筛选方法可能会达到更好的效果。

## 鸣谢

感谢全国研究生数学建模竞赛组委会举办这次比赛，为我们提供一个检验自己知识水平和展示自己的平台，感谢所有工作人员的辛勤付出，感谢队友们的努力奋斗

## 参考文献