第 8 卷 第 5 期
2019 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 8 No. 5
Sep. 2019

引文格式：
胡奕绅, 朱木春, 殷鹏. 基于多步筛选法的心脑血管疾病全基因组关联研究［J］. 集成技术, 2019, 8(5): 72-85.
Hu YS, Zhu MC, Yin P. Genome-wide association study of cardiovascular and cerebrovascular diseases based on multi-step screening［J］. Journal of Integration Technology, 2019, 8(5): 72-85.

基于多步筛选法的心脑血管疾病全基因组关联研究

胡奕绅[1,2] 朱木春[1] 殷 鹏[1]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](深圳大学 深圳 518061)

摘 要 全基因组关联研究是研究复杂疾病和性状遗传效应的一种有效手段。现有关联分析主要用的是边缘统计检验的方法，但未考虑特征间相关性、阈值选取不稳定等问题。该文以心脑血管疾病为研究对象，提出了一种基于多步筛选法的全基因组关联分析新方法。该方法可以简要概括为以下两步：首先利用 Gini 指数做特征初始筛选，获得一个候选单核苷酸多态性子集，再用基于随机森林的递归聚类消除法从单核苷酸多态性子集中发现关联单核苷酸多态性。实验结果表明，多步筛选法比单步特征选择的效果更好，基于 Gini 指数的基于随机森林的递归聚类消除法筛选的单核苷酸多态性子集与疾病的关联性更高。

关键词 心脑血管疾病；特征选择；单核苷酸多态性；多步筛选
中图分类号 TG 156 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20190702002

Genome-Wide Association Study of Cardiovascular and Cerebrovascular Diseases Based on Multi-Step Screening

HU Yishen[1,2] ZHU Muchun[1] YIN Peng[1]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*Shenzhen University*, *Shenzhen* 518061, *China*)

Abstract Genome-wide association study (GWAS) is an effective method to study genetic variants associated with complex diseases or traits. Marginal statistical test is the common method of GWAS, however there following weakness such as lack of consideration of correlation between the features and unstable threshold selection. In this paper, we discuss a new method of GWAS based on multi-step tests model for cardio-cerebrovascular disease. The method can be divided into the following two steps: Gini index is used for first-

收稿日期：2019-07-02 **修回日期**：2019-08-12
基金项目：国家自然青年科学基金项目(11801542)；深圳市科创委学科布局项目(JCYJ20180703145002040)
作者简介：胡奕绅，硕士研究生，研究方向为基因组学、电子病历数据挖掘；朱木春，硕士，研究方向为多组学数据挖掘；殷鹏(通讯作者)，博士，副研究员，硕士研究生导师，研究方向为医疗健康大数据挖掘和人工智能等，E-mail：peng.yin@siat.ac.cn。

step feature selection to achieve a subset of single-nucleotide polymorphisms (SNPs), and then random forest recursive cluster elimination (RF-RCE) filters the associated SNPs subset from first-step candidate SNP set. Experiment results show that the multi-step feature selection is better than the single-step feature selection, and the selected SNPs are more suitable for cardio-cerebrovascular disease prediction.

Keywords　cardio-cerebrovascular disease; feature selection; single-nucleotide polymorphism; multi-step selection

1　引　言

全基因组关联研究(Genome-Wide Association Study，GWAS)是一种用于检测遗传变异与群体中常见疾病或性状关联性的研究方法，目的是深入了解疾病性状的形成机制和遗传结构特征。自从 2005 年 Science 期刊发表第一篇研究人类疾病的 GWAS 以来，目前已经发现有超过 50 万个与常见人类疾病或性状相关联的遗传变异。GWAS 在过去 10 年间彻底改变了复杂疾病遗传学领域，为人类复杂的性状和疾病提供了许多令人信服的研究[1]。

随着 GWAS 的不断发展，目前已有了相应的分析软件，如 Plink、GenABEL 等。根据这些软件的处理过程，可将 GWAS 分析过程分为数据清理和关联分析两步。其中，数据清理已经发展得相当成熟；关联分析主要用的是回归分析和卡方检验[2]。在关联分析中，用 *P* 值来表示每个单核苷酸多态性(Single-Nucleotide Polymorphism，SNP)位点与表型的关联强度。其中，*P* 值越小，表示关联性越强。*P* 值的阈值设定决定了 GWAS 找出关联位点的有效程度。近年来，基于回归分析和卡方检验的 GWAS 方法已经有了很多的应用。例如，Hadji-Turdeghal 等[3]通过回归分析发现了与晕厥相关联的变异位点 rs12465214；Nielsen 等[4]通过卡方检验发现了心血管疾病的易感基因 *ZNF529*；Matsukura 等[5]通过 *t* 检验和卡方检验发现了日本人周动脉疾病的敏感性位点；Klarin 等[6]通过逻辑回归分析在数亿个退伍军人的脱氧核糖核酸(DNA)中找到了几个与外周动脉疾病相关的新 SNP 位点，同时还找到了一种与静脉血栓形成相关的突变——the Factor V Leiden mutation(FVL)。然而，以上提到的这些方法都是只针对单个 SNP 分析的，忽略了多 SNP 间的相互作用。而疾病的发生不仅仅是某一个位点的单独作用，SNP 也不是独立的。因此，基于回归分析和卡方检验的 GWAS 方法会出现多假阳性和假阴性的问题，且在阈值的选取上没有稳定的标准。

机器学习作为数据挖掘的重要手段，提供了丰富的特征选择算法，可以弥补统计检验的不足，找到比统计检验关联性更强的位点。由于早期计算机计算能力的不足，惩罚回归分析(包括 Lasso 回归与岭回归等)、随机森林(Random Forest，RF)和梯度提升机等机器学习方法无法直接应用在数据量庞大的全基因组数据上，并且机器学习方法对生物解释性不如统计检验，因此，基于机器学习算法的 GWAS 方法应用还远少于传统的统计检验方法。但随着计算机性能的不断提高和生物解释性上的逐渐完善，基于机器学习的 GWAS 相关研究正在慢慢开展。例如，Sun 等[7]应用岭回归方法，分析具有已知类风湿性关节炎易感性位点的 1、6 和 9 号染色体区域的 SNP；Kim 等[8]将 RF 应用在模拟数据集上，找到一个跟心肌梗死相关的风险 SNP；Arshadi 等[9]使用梯度提升机的相对影响测量得到高排名

的 SNP 位点作为关联位点。因此，与统计检验相比，基于机器学习算法的 GWAS 方法的优势在于隐含了相互作用效应，无需设置特定的阈值来选取显著的 SNP 位点。

本文以心脑血管疾病为研究对象，提出一种基于多步筛选法发现关联 SNP 位点的方法。通过文献调研和比较分析，利用决策树中的 Gini 指数作为初始筛选来去除低效应的 SNP，减少计算复杂度，并应用遗传算法(Genetic Algorithm，GA)[10]和基于随机森林的递归聚类消除法(Random Forest Recursive Cluster Elimination，RF-RCE)[11]，使 SNP 的选取不受硬性阈值的限制，且同时又能考虑 SNP 位点之间的相互作用。其中，基于多步筛选法的优点在于：(1)减少了计算量，实现算法的可行性；(2)考虑特征的关联性，充分利用 SNP 位点之间相互作用的特性；(3)通过初始筛选剔除较多冗余位点，减少与不相关位点的联系。

2 材料与方法

2.1 数据

2.1.1 数据来源

本文所用数据集是心脑血管疾病患者的个人基因组数据，采集于科研项目合作的医院，共有 1 356 个样本，测得 596 848 个 SNP 位点基因型数据，包括 1 163 个正常人样本和 193 个患病样本。

2.1.2 数据预处理

本文所用的基因组数据如表 1 所示，其中行表示该样本的数据信息，包括家系 ID、个体 ID、父亲 ID、母亲 ID、性别、疾病状态和 SNP 基因型。在实际使用过程中，只使用疾病状态和 SNP 基因型两类数据，表示样本的特征和标签。

基因组数据预处理通过使用 Plink 软件(https://www.cog-genomics.org/plink/1.9/)，分以下 3 步来完成：(1)去除最小等位基因频率(低于 0.01 的 SNP；(2)去除不满足哈迪-温伯格定律(Hardy-Weinberg Equilibrium)的群体；(3)去除 SNP 缺失率高的样本，其中缺失率设为 0.01。通过使用 Python 脚本语言，将表 1 中的数据变成易于数据分析的格式，具体如表 2 所示。其中，表 1 中的 SNP 基因型数据由数字 0、1、2 代替。具体地，0 代表较大基因频率的纯合子，如 AA；1 代表杂合子，如 AG；2 代表较小基因频率的纯合子，如 GG。

2.2 方法

2.2.1 遗传算法

从众多 SNP 中选出与疾病相关的 SNP 子

表 1 基因组数据集的格式

Table 1 Format of genomic dataset

家系 ID	个体 ID	父亲 ID	母亲 ID	性别	疾病状态	SNP1	SNP2	SNP3	SNP4	SNP5	…
1	1	0	0	1	1	AG	AT	CG	CC	AT	…
1	2	1	3	1	2	AA	AA	CC	AA	TT	…
1	3	0	0	2	2	AA	AT	GG	AC	AA	…
2	4	0	0	1	1	GG	TT	GC	AC	AT	…
2	5	4	6	2	2	AA	TT	CC	AA	AA	…
2	6	0	0	2	1	AG	AT	CG	CC	TT	…
3	7	0	0	1	2	AA	AA	CC	CC	TT	…
3	8	7	9	2	2	GG	AT	GG	AA	AT	…

注：疾病状态列中的 1 代表没有患病，2 代表患病；SNP 列中的字母表示通过基因测序获得的 SNP 基因分型

表 2　预处理后的基因组数据集

Table 2　Pretreated genomic dataset

SNP1	SNP2	SNP3	SNP4	SNP5	…	疾病状态
1	1	1	2	1	…	1
0	2	0	0	0	…	2
0	1	2	1	2	…	2
2	0	1	1	1	…	1
0	0	0	0	2	…	2
1	1	1	2	0	…	1
0	2	0	2	0	…	2
2	1	2	0	1	…	2
0	0	1	1	0	…	1

集，是一个多目标组合优化问题，所以可以使用遗传算法来解决该问题[10]。遗传算法是一种以生物进化论为基础，根据适者生存和自然选择规律筛选出最优个体构成种群的全局搜索最优算法[12]。该算法包括基因编码、种群初始化、适应度选择、交叉、变异和种群进化结束 6 个过程。

(1)基因编码

用二进制编码代表种群中的一个个体。其中，1 表示特征被选出，即 SNP 被选出；0 表示特征没有被选出。

(2)种群初始化

根据样本中 SNP 的数量，用随机数随机生成同等数量的二进制数表示一个个体，并确定种群的个体数。

(3)适应度选择

遗传算法中最重要的一步，便是确定一个适应度函数来模拟自然选择。本文使用分类算法作为评估函数来评估选取的 SNP 子集，评价指标为接受者操作特性曲线(ROC)下的面积(Area Under the Curve，AUC)。具体地，使用轮盘赌法来选择种群中的个体，其中 *AUC* 值大的适应度较高，更容易被选择存活下来。

(4)交叉

以一定概率(本文设置为 0.8)选择偶数个个体，让它们进行两两交叉互换。

(5)变异

每个个体以一定的概率(本文设置为 0.01)在某一位点上发生变异，由 0 变 1，或由 1 变 0。

(6)种群进化结束

当迭代过程达到一定次数，或者适应度值，也就是 *AUC* 不再发生大的改变时，让种群的进化结束。其中，*AUC* 值最大的个体就是最优的个体。

2.2.2 基于随机森林的递归聚类消除法

随机森林不仅可以用于构建预测模型，还可以用于判断所有特征的重要性。随机森林在构建一颗树的过程中，会随机地、有放回地从原始样本集合 *D* 中抽出相同个数的样本集合，当重复抽 *N* 次，可以得到 *N* 个新的样本集合，并构建 *N* 颗决策树。在这个过程中，每棵树大约有 37% 的样本未被抽中，没有作为训练集，这一部分样本数据称为袋外数据(Out of Bag，OOB)[11]，可以估测决策树的泛化能力。随机森林评估特征重要性的主要原理是当随机置换样本中的任意一个特征时，计算每棵树置换前后 OOB 样本的误差变化，再对这些树的误差变化求平均，得到平均误差变化。若平均误差变化越大，则特征重要性越高。具体分析过程如下：

(1)随机置换前第 *n* 颗树(n=1,2,⋯,N)的

OOB 准确率为 Acc_n，某一特征 f 置换后第 n 颗树的 OOB 准确率为 Acc_n^f，则决策树 T_n 特征置换前后的 OOB 准确率之差为：

$$Err_n^f=Acc_n-Acc_n^f,\ n=1,2,\cdots,N \tag{1}$$

(2)特征 f 对整个随机森林的 OOB 平均准确率之差为：

$$Err^f=\frac{1}{N}\sum_{n=1}^{N}Err_n^f \tag{2}$$

Err_n^f 的方差为：

$$S^2=\frac{1}{N-1}\sum_{n=1}^{N}(Err_n^f-Err^f)^2 \tag{3}$$

(3)特征 f 对整个样本的重要性为：

$$f_{imp}=\frac{Err^f}{S^2} \tag{4}$$

依次类推，得到原始样本中所有特征的重要性。

随机森林能够利用 OOB 样本数据评估特征的重要性，对高维小样本数据能够进行有效的处理。虽然随机森林能够得到特征的重要性排序，但并没有一个明确的阈值来对重要性较高的特征进行筛选。Yousef 等[13]提出了一种基于机器学习方法的递归聚类消除特征的方法来选出比较重要的特征，并在此基础上提出 RF-RCE[11]。本文采用 RF-RCE 进行特征选择，挑选出相关的 SNP 子集。

RF-RCE 算法流程如下：

(1)设置初始值。其中，初始聚类数为 M、终止聚类数为 N、特征集为 F、类间删除比例为 d、类间删除的聚类数阈值为 p、类内删除比例为 d_1、类内删除的特征数阈值为 p_1；

(2)构建随机森林模型，用 OOB 样本数据求出每个特征 f 的重要性 f_{imp}。

(3)运用 K-means 算法将特征集 F 聚成 M 类，并计算出每一类的得分(得分等于该类中最大的 f_{imp})。

(4)如果 $M\geqslant p$，删除得分最低的那些类(删除比例为 d)。如果 $M<p$，需要进一步判断，当类内特征数 $>p_1$，则删除重要性最低的特征(删除比例为 d_1)；当类内特征数 $\leqslant p_1$，则保留所有特征。

(5)修改 $M=M(1-d\%)$，剩下的特征集组合成新的特征集 F，重复(2)(3)(4)(5)步骤，直到 $M\leqslant N$。

2.2.3 Gini 指数

机器学习算法中的决策树用特征分割数据集生成树时，通常会选择最优的特征进行划分。选出最优特征的方法一般有 3 种：一是信息增益，二是信息增益比，三是 Gini 指数。其中，信息增益只能用于离散数据，并且更容易选择多值的特征分裂，导致树模型效果变差；信息增益比需要多次遍历数据，并按照数值排序选择排位靠前的特征，计算量较大；Gini 指数既可用于离散数据，也可用于连续数据，无需对数据排序，提高了计算的效率[14]。因此，本文采用 Gini 指数作为特征选择的方法。

数据集中类别的多少可以衡量该数据的纯度。如果类别越少，那么就说明该数据集的纯度越高。Gini 指数可以衡量数据集的纯度。其中，Gini 指数的大小与数据集的纯度呈负相关。假设有数据集 S，当中有 Z 个类别，则 Gini 指数的计算公式如下：

$$Gini(S)=1-\sum_{k=1}^{Z}p_k^2 \tag{5}$$

其中，p_k 是指第 k 类数据的个数占总数据个数的比例。

基于 Gini 指数的 CART 决策树是二叉树。假设以某一特征 α 分裂时分为 C_1 和 C_2 子集，以特征 α 分裂后的 Gini 指数计算公式如下：

$$Gini(S,a)=\frac{|C_1|}{|S|}Gini(C_1)+\frac{|C_2|}{|S|}Gini(C_2) \tag{6}$$

每个 SNP 都可以算出其对应的 Gini 指数，然后根据 Gini 指数从小到大进行排序，其中排得越靠前的 SNP 与结果越相关。

2.2.4 评价指标

本文采用 *AUC* 作为衡量上述各方法性能的指标，主要是衡量 SNP 子集的相关性程度的指标。如果定义真阳性率(True Positive Rate，TPR)为被正确分类为正样本的正样本数占所有实际正样本数量的百分比，假阳性率(False Positive Rate，FPR)为被错误分类为正样本的负样本数占所有实际负样本数量的百分比，那么 *AUC* 的定义是：TPR 和 FPR 在坐标图形成一条曲线，称为 ROC 曲线，曲线下的面积便是 *AUC*。其中，TPR 和 FPR 的计算如公式(7～8)所示。

$$TPR=\frac{TP}{TP+FN} \tag{7}$$

$$FPR=\frac{FP}{FP+TN} \tag{8}$$

其中，*TP* 表示测试集中被正确分类为正样本的正样本数；*FN* 表示测试集中被错误分类为负样本的正样本数；*FP* 表示测试集中被错误分类为正样本的负样本数；*TN* 表示测试集中被正确分类为负样本的负样本数。

本文涉及的是一个二分类问题，用 *AUC* 可以很好地评估模型分类的性能。若 *AUC* 越大，表明分类效果越好，说明特征子集对标签的贡献越大，相关性也越高。其中，*AUC* 的取值介于 0.5～1。

3　实验与结果

本文将心脑血管疾病的数据集以 7∶3 的比例随机划分为训练集和测试集。本文所使用的特征选择方法一共有 5 种，其中属于单步筛选的有 3 种，分别是回归分析、卡方检验和 RF-RCE；属于多步筛选的有两种，分别是基于 Gini 指数的遗传算法和基于 Gini 指数的 RF-RCE。此处使用 Gini 指数作为初始筛选主要有 3 个原因：(1)Gini 指数能对每个 SNP 位点进行评分[15]，通过排序删除大量噪声 SNP；(2)Gini 指数能适应小样本、超高维的基因数据，且对参数的设置不太敏感[16]；(3)对于数据集中的噪声和假阳性 SNP，RF 表现出较强的鲁棒性[17]。本文所使用的分类算法有 4 种，分别是支持向量机(Support Vector Machine，SVM)、RF、自适应集成学习(Adaptive Boosting，AdaBoost)和反向传播神经网络(Back Propagation Neural Network，BPNN)。经过数据测试和参数调优，选取以下参数作为分类算法的参数：(1)选取线性核作为 SVM 的参数；(2)使用 200 颗决策树和 Gini 指数作为分裂规则，作为 RF 的参数；(3)使用 200 颗决策树和学习率 0.8，作为 AdaBoost 的参数；(4)使用两层隐藏层(5 层和 2 层)作为 BPNN 的参数。

本文所有算法程序均使用 Python 语言实现，应用 numpy 和 pandas 进行数据的读取写入，应用 sklearn 实现分类算法，特征选择根据算法的流程设计编程实现。程序运行环境有：(1)个人笔记本电脑(宏碁 Aspire E5-571G-500A)，内存 8 G，双核 CPU，Win10 系统；(2)服务器，内存 128 G，八核 CPU，Ubuntu 16.04 系统。

3.1　查找关联 SNP 子集

3.1.1 回归分析

本文所解决的问题是一个二分类问题，故用到的是逻辑回归。SNP 位点是自变量，通过逻辑回归分析可以给每个 SNP 位点计算出对应的 *P* 值，通过 *P* 值筛选出小于显著性水平 β 的 SNP 位点，得到相关的 SNP 子集，如图 1 所示。这一过程使用 Plink 软件来完成。

当取不同阈值时，得到不同数量的 SNP 子集，具体如表 3 所示。当 *P* 值越小时，$-\log_{10}(P)$ 越大，得到的 SNP 子集数越少。

采用表 3 中的 5 个阈值选出的 SNP 子集对测试集进行预测。由图 2 可知，在所有阈值情况下，RF 和 BPNN 求出的 *AUC* 比较稳定，基本在 0.5～0.51；而 SVM 和 AdaBoost 在 $-\log_{10}(P)=$

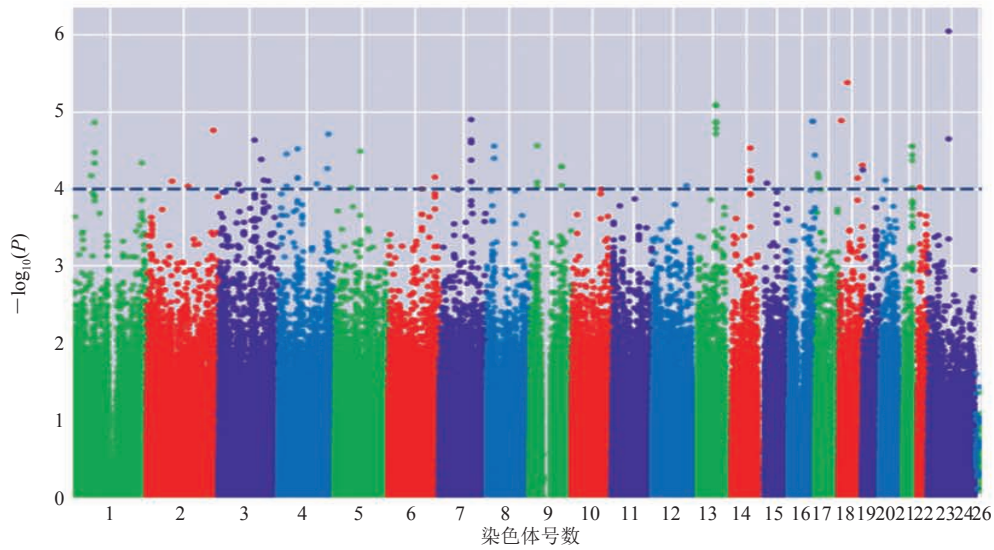


图 1 SNP 位点的曼哈顿图

Fig. 1 Manhattan plot of SNP loci

表 3 回归分析不同阈值对应的 SNP 子集数量

Table 3 The number of SNP subsets corresponding to different thresholds by regression analysis

$-\log_{10}(P)$	SNP 子集数
3.50	206
3.75	120
4.00	69
4.25	40
4.50	25

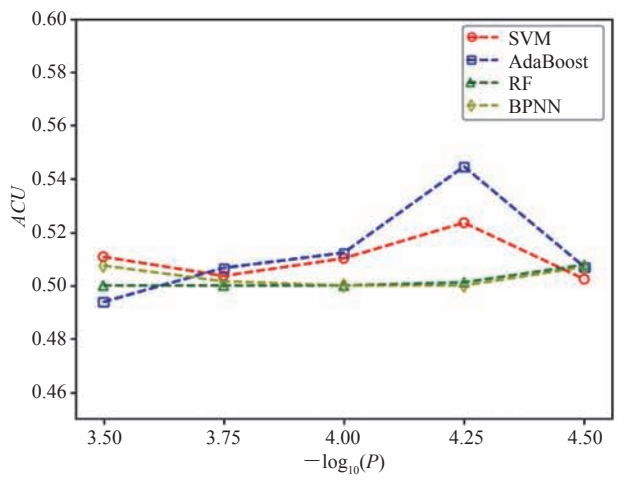


图 2 回归分析不同阈值下的 SNP 子集在测试集上的预测结果曲线

Fig. 2 Prediction curves of SNP subsets on *t*-test sets under different thresholds of regression analysis

4.25 时取得较大的突破，分别达到 0.523 6 和 0.544 7；综合 4 种算法的结果来看，可以认为当$-\log_{10}(P)=4.25$ 时得到的 SNPs 子集(数量为 40)相关性最强，可作为回归分析的最相关的 SNP 子集。

3.1.2 卡方检验

当卡方值不同时，*P* 值也不同。其中，卡方值越大，*P* 值则越小。利用不同的阈值，对 SNP 子集进行选取，选取出满足≤该阈值的 SNP 位点，具体如表 4 所示。

表 4 卡方检验不同阈值对应的 SNP 子集数量

Table 4 The number of SNP subsets corresponding to different thresholds by chi-square test

P 值	SNP 子集数	*P* 值	SNP 子集数
1.00e-06	9	0.001	204
2.00e-06	9	0.002	402
3.00e-06	9	0.003	604
4.00e-06	10	0.004	800
5.00e-06	11	0.005	986
1.00e-05	12	0.01	2 195
2.00e-05	14	0.02	5 003
5.00e-04	118		

使用表 4 中的 15 个阈值选出的 SNP 子集对测试集进行预测。由图 3 可知，在不同阈值条件

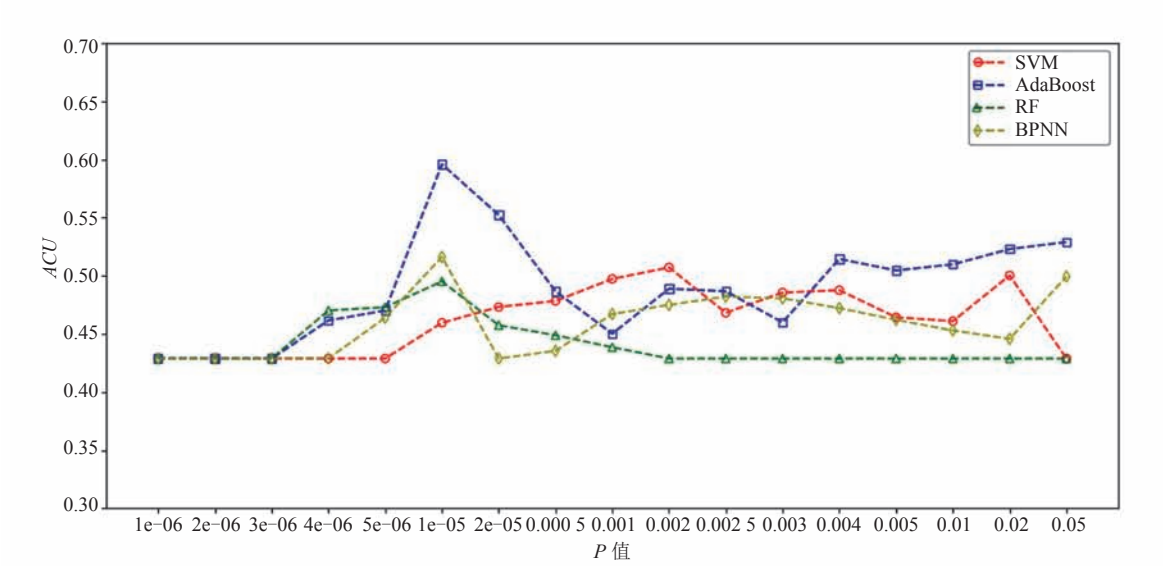


图 3　卡方检验不同阈值下的 SNP 子集在测试集上的预测结果曲线

Fig. 3　Prediction curves of SNP subsets on test sets under different thresholds of Chi-Square test

下，各种分类算法在测试集上预测出的平均准确度并不是稳定的，而是像过山车一样不断变化的。相对而言，当 $P=1\times10^{-5}$ 时，大多数算法的表现比其他情况好，故认为卡方检验在 $P=1\times10^{-5}$ 时可以取到最相关的子集。

3.1.3　基于 Gini 指数的遗传算法

通过 Gini 指数对所有 SNP 位点进行筛选，然后用筛选后的子集作为遗传算法的输入，进一步筛选出相关度更高的 SNP 子集。一般来说，与疾病相关的 SNP 子集相对于整个 SNP 数据集来说占极少数。因此，先取 Gini 指数排序前的 100 个 SNP 来做遗传算法，进一步查找高相关子集。在 Gini 指数进行粗粒度的筛选之后，由遗传算法进行细粒度查找。此过程为全局寻优，目的是让算法快速收敛，得到最优个体，进而解码得到相关 SNP 子集。

由图 4 可知，最终迭代收敛之后，四种分类算法作为适应度函数得到了各自恒定不变的适应度值，即 *AUC* 也得到了各自取得最优值的个体。假如最优个体为 {1,0,0,1,1,⋯,1,0,1}，则 1 代表选择该 SNP，0 代表抛弃该 SNP，最终得到高相关的 SNP 子集。其中，SVM 的 *AUC* 为 0.550 3；AdaBoost 的 *AUC* 为 0.564 8；RF 的 *AUC* 为 0.525 9；BPNN 的 *AUC* 为 0.566 2。

3.1.4　基于随机森林的递归聚类消除法

前文中有提到，直接使用 RF-RCE 对基因组数据进行特征选择会遇到计算量的问题，导致耗时过长。但为了观察其特征选择的效果，这里依然用了几个小时对其进行实验。RF-RCE 的初始聚类数 *M* 选为 500、400、300、200，通过不断迭代缩小聚类数，当聚类数缩小为 1 时，初始聚类数 $M=500$ 得到了 182 个 SNP，初始聚类数 $M=400$ 得到了 193 个 SNP，初始聚类数 $M=300$ 得到了 206 个 SNP，初始聚类数 $M=200$ 得到了 212 个 SNP。用前文提到的 4 种分类算法对以上 SNP 子集进行预测，结果如图 5 所示。

由图 5 可知，当初始聚类数为 300 时，SVM、AdaBoost、BPNN 的表现最好；当初始聚类数为 400 时，AdaBoost 的表现最好。相对而言，初始聚类数为 300 时，四种算法的总体表现最好，$AUC_{SVM}=0.5473$、$AUC_{AdaBoost}=0.5576$、$AUC_{RF}=0.5086$、$AUC_{BPNN}=0.5533$。

3.1.5　基于 Gini 指数的 RF-RCE

前文提到，在使用 RF-RCE 前先用 Gini 指

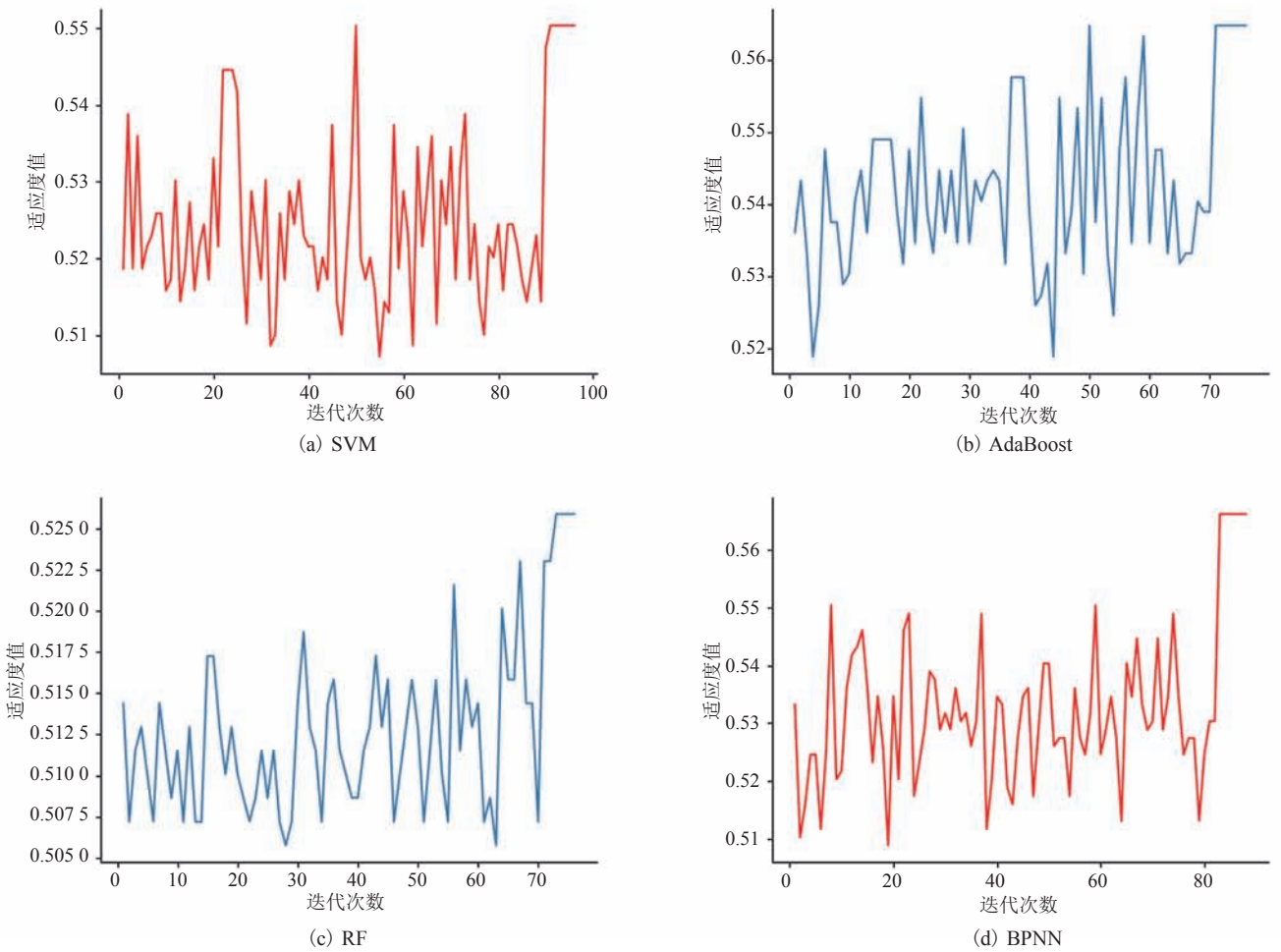


图 4 不同适应度函数下遗传算法的迭代寻优过程

Fig. 4 Iterative optimization process of genetic algorithm under different fitness functions

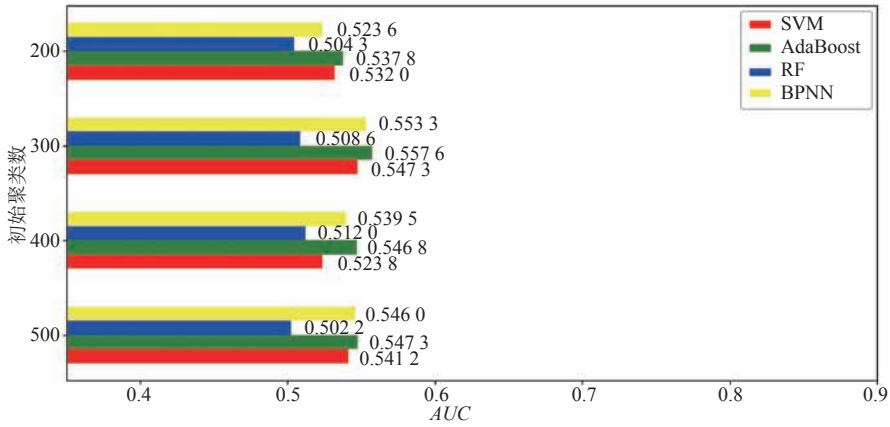


图 5 RF-RCE 不同初始聚类数下各种算法的 *AUC*

Fig. 5 *AUC* of various algorithms under different initial cluster numbers

数可以减少计算的复杂性。在这里同样先利用 Gini 指数做一个初始筛选，选出了 2 000 个靠前的 SNP 子集，再用 RF-RCE 从 2 000 个中选出相关的 SNP 子集，最后用机器学习分类算法进行预测。初始聚类数 M 选为 500、400、300、200，不断迭代缩小聚类数，最后当聚类数缩小为 1 时，初始聚类数 M=500 得到了 118 个 SNP，初始聚类数 M=400 得到了 129 个 SNP，初始聚类数 M=300 得到了 139 个 SNP，初始聚类数 M=200 得到了 151 个 SNP。用前文提到的 4 种分类算法对这些 SNP 子集进行预测，结果如图 6 所示。

由图 6 可知，在不同的初始聚类数条件下，各种算法的准确率基本在 0.6 以上。当初始聚类数为 200 和 300 时，AdaBoost 的表现最好；当初始聚类数为 400 和 500 时，SVM 的表现最好。相对而言，初始聚类数为 200 时，四种算法的总体表现最好，AUC_{SVM}=0.6792、$AUC_{AdaBoost}$=0.6899、AUC_{RF}=0.6486、AUC_{BPNN}=0.6380。

3.2　交叉验证

为了更好地体现各种方法的优劣性，此处采用交叉验证的方法来说明数据的拟合情况。将数据以 7∶3 的比例再次随机分为训练集和测试集，操作 5 次，得到 5 组不同的训练集和测试集，重复上述 3.1 中的所有算法查找关联 SNP 子集的过程，结果如图 7 所示。

4　讨　论

4.1　结果对比分析

上述 5 种特征选择方法都选出了各自最相关的 SNP 子集，然后用这些子集来对疾病状态进行分类预测。由图 7 可知，在 5 组随机分配的训练集和测试集上进行交叉验证时，基于 Gini 指数的 RF-RCE 进行 SNP 筛选的效果基本上好于其他的特征选择方法。下面对每组数据分类预测得出的 *AUC* 求平均值并进行对比，结果如图 8 所示。可以看到，两种多步筛选法和 RF-RCE 的 *AUC* 都要比回归分析、卡方检验的高，特别是基于 Gini 指数的 RF-RCE 的 *AUC* 比回归分析、卡方检验都高出 5%～8%。因此，我们认为逻辑回归和卡方检验在选择关联性特征上不如其他 3 种算法。值得注意的是，RF-RCE 的 *AUC* 都低于两种多步筛选法(Gini-GA 和 Gini-RF-RCE)，原因可能是 RF-RCE 在聚类的过程中过多考虑了与噪声 SNP 的相互作用，降低了筛选出的 SNP 子集的关联性。这也充分说明了多步筛选法的优越性，可以剔除较多的冗余 SNP 位点，提高与阳性位点的关

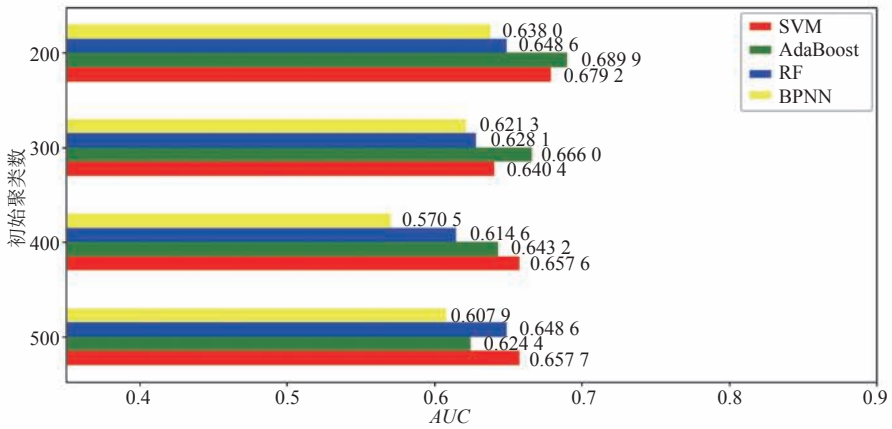
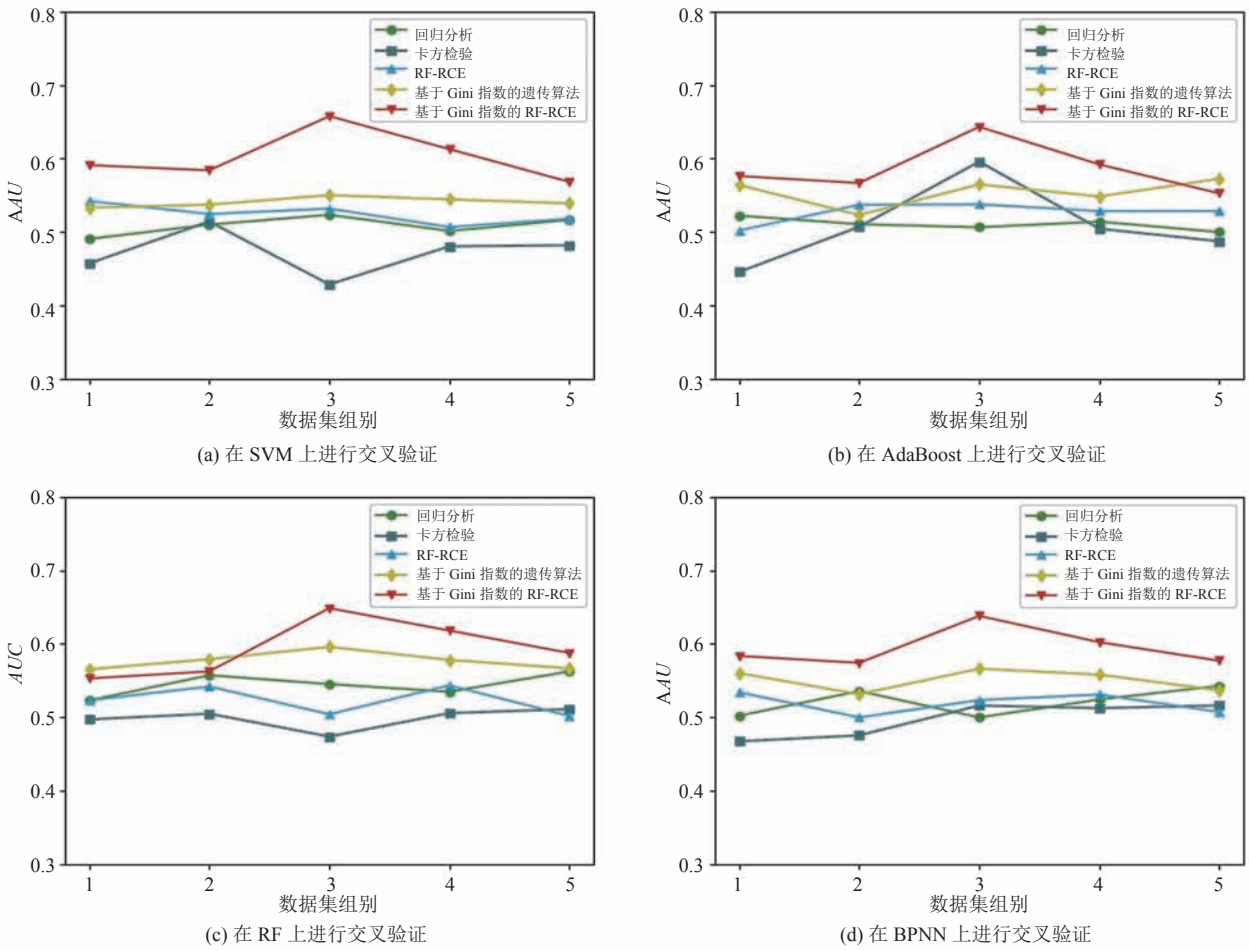


图 6　Gini-RF-RCE 不同初始聚类数下各种算法的 *AUC*

Fig. 6　*AUC* of various algorithms under different initial cluster numbers



(a) 在 SVM 上进行交叉验证　(b) 在 AdaBoost 上进行交叉验证

(c) 在 RF 上进行交叉验证　(d) 在 BPNN 上进行交叉验证

图 7　五组训练集和测试集对不同特征选择方法进行交叉验证

Fig. 7　Cross-validation of different feature selection methods

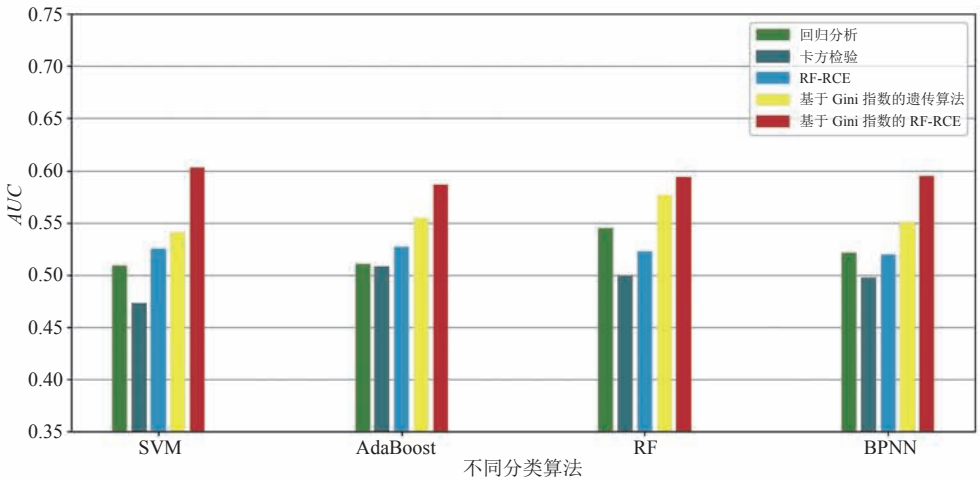


图 8　多种特征选择在 *AUC* 上的比较

Fig. 8　Comparison of multiple feature selection on *AUC*

联性。

通过对 *AUC* 值的比对分析，得出以下 3 个结论：(1) 在心脑血管疾病的数据集上，用本文提出的多步筛选法进行特征选择比传统的统计检验 *AUC* 更高，获得的 SNP 子集的关联性更强；(2) 本文提出的多步筛选法比单步特征选择的效果更好；(3) 本文提出的基于 Gini 指数的 RF-RCE 的分类预测 *AUC* 比传统的统计检验、Gini-GA 和 RF-RCE 提高了很多，说明其筛选的 SNP 子集与疾病的关联性最高。

4.2　定位关联的易感基因

根据上述 4.1 的对比分析，5 种特征选择方法表现最好的是基于 Gini 指数的 RF-RCE，接下来使用这个表现最好的特征选择方法来定位与疾病关联的 SNP 可能作用的易感基因。由 3.1.5 可知，4 种不同初始聚类数分别找出了 118、129、139、151 个 SNP 子集，为了最大程度提升所得 SNP 位点的关联性，本文认为只有在所有的 SNP 子集中都出现的 SNP 位点才是关联的 SNP 位点，所以对 4 个不同聚类得到的 SNP 子集取交集，得到了 13 个本文认为最相关的 SNP 位点，具体如表 5 所示。

在表 5 中，有两个加(*)的基因，表示已有研究发现其与心脑血管疾病有关联。*PCLAF* 是一种与后外侧心肌梗死有关蛋白编码基因[18]；*FBLN5* 是一种编码细胞外基质蛋白的基因，与动脉粥样硬化有关，其编码的蛋白可能在血管发育和重构中发挥作用[19]。这也进一步说明了本文提出的多步筛选法的有效性。其他匹配到的基因暂时没有找到公开的文献表明其与心脑血管疾病的关联性，作为潜在的相关基因，为后续复杂疾病致病机制的研究提供了方向。

4.3　与国内外类似研究的比较分析

现阶段复杂疾病的全基因组关联研究，主要是统计检验的多元回归分析、卡方检验等方法。文献[3-6]利用统计检验计算 *P* 值对 SNP 位点进行排序，并结合相关的生物实验找到了许多阳性位点。然而，统计检验未考虑多 SNP 间的相互作用，因此挑选出的高于阈值的 SNP 位点不一定是阳性 SNP，容易出现假阳性和假阴性的问题，且在阈值的选取上没有稳定的标准。机器学习的算法替代传统的统计检验进行全基因组关联研究，

表 5　相关 SNP 位点及其匹配基因

Table 5　Related SNP loci and their matching genes

序号	SNP ID	碱基对的位置	所属染色体	基因
1	Rs12692222	239413172	2	*MLPH*
2	Rs7591026	170271626	2	*TANC1*
3	Rs3732933	184911151	3	*EHHADH*
4	Rs6818184	118560727	4	*LINC01378*
5	Rs378932	5609695	5	*PLPP1*
6	Rs17074404	145208507	6	*RMND1*
7	Rs1010795	22720023	8	*PEBP4*
8	Rs1690627	31294981	10	*ZNF438*
9	Rs741198	92367411	14	*FBLN5**
10	Rs2062250	64672002	15	*PCLAF**
11	Rs2008344	3741562	16	*TRAP1*
12	Rs4793360	69543626	17	*SEPT9*
13	Rs844956	140450241	23	*FAM9B*

注：“*”表示已有研究发现该基因与心脑血管疾病有关

也慢慢成为了主流。Sun 等[7]应用岭回归方法，分析具有已知类风湿性关节炎易感性位点的 1、6 和 9 号染色体区域的 SNP；Kim 等[8]将 RF 应用在模拟数据集上，找到一个跟心肌梗死相关的风险 SNP；Arshadi 等[9]使用梯度提升机的相对影响测量得到高排名的 SNP 位点作为关联位点。与统计检验相比，机器学习考虑了特征的关联，可以减少假阳性的出现。Maciukiewicz 等[20]结合变量优先级和 LASSO 回归的方法挑选出预测抗抑郁症药物治疗效果的预测因子，并用 SVM 搭建预测模型，*AUC* 最高达到 0.66。但是，该方法依然需要在外部样本中进行进一步的验证和复制。本文的多步筛选法也是利用算法结合的方式，将 Gini 指数和 RF-RCE 结合来筛选 SNP，计算出的 *AUC* 高于传统的统计检验和单步筛选法，且基于心脑血管疾病的真实数据集，比仿真数据得出的结论更加真实可信。当然，本文的研究还存在不足之处，如数据量不足、多步筛选 SNP 时存在算法的排序问题。未来工作将重点解决这些问题，并尝试与其他算法的结合取得更好的效果。

5 结 论

本文主要针对传统的统计检验和现有的一些机器学习特征选择方法在挑选关联 SNP 上的不足，选择了两种既考虑特征相互作用，又不用设定人工阈值的特征选择方法来研究。针对机器学习方法计算的复杂性和可行性的问题，在两种特征选择算法的基础上提出多步筛选法来适应基因组数据。通过实验分析，我们认为在心脑血管基因组数据集上，使用本文提出的多步筛选法进行特征选择比传统的统计检验和单步特征选择所得到的 *AUC* 更高，所以其所选出的 SNP 子集关联性更强，效果更好；用两种多步筛选法中表现最好的基于 Gini 指数的 RF-RCE 来找出与疾病关联的易感基因，为心脑血管疾病的致病机制的研究提供方向和指导。在未来的研究中，希望可以在更多复杂疾病基因组数据集上验证多步筛选法的有效性和普适性。

参 考 文 献

[1] Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies [J]. Nature Reviews Genetics, 2019, 20: 467-484.

[2] Martin LS, Eskin E. Population structure in genetic studies: confounding factors and mixed models [Z]. bioRxiv, 2016, doi: https://doi.org/10.1101/092106.

[3] Hadji-Turdeghal K, Andreasen L, Hagen CM, et al. Genome-wide association study identifies locus at chromosome 2q32.1 associated with syncope and collapse [J]. Cardiovascular Research, 2019, https://doi.org/10.1093/cvr/cvz106.

[4] Nielsen JB, Rom O, Surakka I, et al. Loss-of-function genomic variants with impact on liver-related blood traits highlight potential therapeutic targets for cardiovascular disease [Z]. bioRxiv, 2019, http://dx.doi.org/10.1101/597377. T.

[5] Matsukura M, Ozaki K, Takahashi A, et al. Genome-wide association study of peripheral arterial disease in a Japanese population [J]. PLoS One, 2015, 10(10): e0139262.

[6] Klarin D, Lynch J, Aragam K, et al. Genome wide association study in the million veteran program identifies a novel role for thrombosis in the pathogenesis of peripheral artery disease [J]. Arteriosclerosis, Thrombosis, and Vascular Biology, 2018, 38(Suppl 1): A126.

[7] Sun YV, Shedden KA, Zhu J, et al. Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression [C] // BMC Proceedings, 2009, 3(Suppl 7): S67.

[8] Kim YH, Wojciechowski R, Sung HJ, et al. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects [C] // BMC Proceedings, 2009, 3(Suppl 7): S64.

[9] Arshadi N, Chang B, Kustra R. Predictive modeling

in case-control single-nucleotide polymorphism studies in the presence of population stratification: a case study using genetic analysis workshop 16 problem 1 dataset [C] // BMC Proceedings, 2009, 3(Suppl 7): S60.

[10] Vafaie H, De Jong K. Robust feature selection algorithms [C] // Proceedings of 1993 IEEE Conference on Tools with Al (TAI-93), 1993: 356-363.

[11] 宁永鹏. 高维小样本数据的特征选择研究及其稳定性分析 [D]. 厦门: 厦门大学, 2014.

[12] 雷英杰, 张善文, 李续武, 等. MATLAB 遗传算法工具箱及应用 [M]. 西安: 西安电子科技大学出版社, 2005.

[13] Yousef M, Jung SG, Showe LC, et al. Recursive cluster elimination (RCE) for classification and feature selection from gene expression data [J]. BMC Bioinformatics, 2007, 8(1): 144.

[14] 尹儒, 门昌骞, 王文剑, 等. 模型决策树: 一种决策树加速算法 [J]. 模式识别与人工智能, 2018, 31(7): 643-652.

[15] Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures [J]. Computational Statistics & Data Analysis, 2008, 52(4): 2249-2260.

[16] Zeng P, Zhao Y, Qian C, et al. Statistical analysis for genome-wide association study [J]. Journal of Biomedical Research, 2015, 29(4): 285.

[17] Bureau A, Dupuis J, Falls K, et al. Identifying SNPs predictive of phenotype using random forests [J]. Genetic Epidemiology, 2005, 28(2): 171-182.

[18] Zema MJ. Electrocardiographic tall R waves in the right precordial leads: comparison of recently proposed ECG and VCG criteria for distinguishing posterolateral myocardial infarction from prominent anterior forces in normal subjects [J]. Journal of Electrocardiology, 1990, 23(2): 147-156.

[19] Nakamura T, Ruiz-Lozano P, Lindner V, et al. DANCE, a novel secreted RGD protein expressed in developing, atherosclerotic, and balloon-injured arteries [J]. Journal of Biological Chemistry, 1999, 274(32): 22476-22483.

[20] Maciukiewicz M, Marshe VS, Hauschild AC, et al. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder [J]. Journal of Psychiatric Research, 2018, 99: 62-68.