文章编号：2095-2163(2020)03-0202-07 中图分类号：TP391 文献标志码：A

基于 XGBoost 的质量性状基因互作检测方法

郭颖婕[1]，李 傲[1]，刘晓燕[1]，郭茂祖[1,2]
（1 哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001；
2 北京建筑大学 电气与信息工程学院 建筑大数据智能处理方法研究北京市重点实验室，北京 100044）

摘 要：在质量性状全基因组关联分析 GWAS 中，以基因作为研究单位的基因-基因相互作用检测方法，以其在统计效力与生物可解释性方面的优势备受关注。然而现有方法中多数对基因之间互作形式给出了强假设，降低了算法对互作关系的检测性能。针对已有方法存在的局限性，本文提出一种基于 XGBoost 的基因互作检测方法 geXGB。XGBoost 作为一种流行且高效的机器学习方法，可以拟合基因型数据与表型之间的作用关系，并利用预测概率与加和模型之间的偏差表征相互作用关系的程度。geXGB 对相互作用形式不作假设，增强该方法对不同形式相互作用的检测能力。仿真与真实实验结果表明：该方法能够有效进行不同类型相互作用的检测，可以应用于全基因组关联研究。

关键词：XGBoost；基因相互作用；单核苷酸多态性位点；质量性状；全基因组关联分析

A gene-based exchanged XGBoost method for detecting and ranking gene-gene interactions of qualitative trait

GUO Yingjie[1], LI Ao[1], LIU Xiaoyan[1], GUO Maozu[1,2]
(1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;
2 Beijing Key Laboratory of Intelligent Processing for Building Big Data, School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

【Abstract】Among the various statistical methods for identifying gene-gene interaction in qualitative genome-wide association studies (GWAS), gene-based methods have recently grown in popularity as they confer advantages in both statistical power and biological interpretability. However, most of these gene-based methods make strong assumptions on the form of the relationship between traits and SNPs, resulting in limited statistical power. The paper proposes a gene-based method based on XGBoost, a popular and highly effective method in machine learning, to model the relationship between genotype and traits, and then measure the interaction of gene pairs by the deviation of the predicted probability from a multiplicative model. This method makes fewer assumptions on the exact form of interaction, which may overcome some of the shortcomings in previous methods. In experiments with both simulation study on pure and strict disease models and real world data, the proposed method outperforms previous approaches in detecting interactions accurately.

【Key words】XGBoost; gene-gene interaction; single nucleotide polymorphism; qualitative trait; genome-wide association studies

0 引 言

研究基因-基因相互作用已被证实对于揭示复杂性状遗传调控机制至关重要。目前已有许多基于 SNP 位点间相互作用的检测方法。统计类的检测方法通过设计表征相互作用强度的统计量，检测显著的相互作用关系，例如基于优势比(odds ratio，OR)的统计量[1]、基于连锁不平衡(linkage disequilibrium，LD)的统计量[2-3]、基于单体型(haplotype)的统计量以及基于熵的统计量[4-5]等。另一类方法则采用人工智能方法的思想，例如采用将为技术的多因子降维方法(multifactor dimensionality reduction，MDR)[6]、基于树模型的 TEAM(tree-based epistasis association mapping)方法[7]、通过优化存储策略加速计算的 BOOST(Boolean operation-based screening and testing)方法[8]，以及基于贝叶斯理论的 BEAM(Bayesian epistasis association mapping)系列方法[9]等。这些基于位点的检测方法面临最大的挑战是维数灾难。由于算法需要考虑所有的 SNP 或 SNP 组，成对或者高阶的相互作用关系检测次数随着相互作用关系阶

基金项目：国家自然科学基金(61571163，61532014，61671189)；国家重点研发计划项目(2016YFC0901902)。

作者简介：郭颖婕(1987-)，女，博士研究生，主要研究方向：生物信息学；李 傲(1995-)，男，硕士研究生，主要研究方向：机器学习；刘晓燕(1963-)，女，博士，副研究员，主要研究方向：数据挖掘；郭茂祖(1966-)，男，博士，教授，博士生导师，主要研究方向：机器学习、智慧城市、生物信息学。

收稿日期：2019-06-28

数呈指数级增长，随之而来的对统计显著性的校正会导致统计效力的弱化。因此，本文研究以基因为单位，将一个基因中的所有 SNP 看做一个整体来检测基因-基因相互作用。

基因是生物功能表达的基本单位。基于基因的相互作用研究有 3 点明显的优势，可阐释分述如下。

(1)基于基因的方法可以大大减少所需的检验次数，20 000 基因之间成对检测互作关系运算量远远小于 300 万 SNP 之间成对检测互作关系。

(2)2 组基因之间可能存在多对 SNP 间的相互作用，组内的 SNP 之间也可能存在连锁不平衡关系，这些同时存在的作用关系会隐性地呈现在以基因为单位的模型中，更利于相互作用的检测。

(3)基于基因的方法可以更好地利用已有的生物学背景知识，缩小研究范围。例如可以检测那些蛋白质互作网络(protein-protein interaction，PPI)中已经呈现互作关系的蛋白质编码基因之间的关系，或者某个调控通路(pathway)内基因之间的相互作用关系。

目前，在以基因为单位的相互作用研究中，Peng 等人[10]在疾病组与对照组中分别对 2 个基因进行典型相关性分析(Canonical Correlation Analysis，CCA)，并设计统计量 CCU 来度量 2 个基因在疾病与对照组中相关性指标的差异程度，用于表征相互作用的强度。该方法的局限性在于 CCA 只能度量 2 个基因之间的线性关系。Larson 等人[11]和 Yuan 等人[12]针对上述方法存在的问题，将 CCU 扩展到 KCCU，在做典型相关性分析之前，将核函数作用在疾病和对照组中两个基因的数据上，从而增强模型对非线性关系的解释能力。Jin 等人[13]提出了 GBIGM，一种基于熵的非参数假设检验方法。通过分析 2 个基因共同作用时与考虑只有单个基因时的熵的变化(即信息增益)，并利用随机置换类标签的方式获得相互作用的显著性 p 值。Emily[14]开发了 AGGrGATOr，该方法首先计算两基因间所有 SNP 对的 Wald 统计值，并将一组 Wald 统计值结合成为一个显著性 p 值用于度量 2 个基因之间是否存在相互作用。此前，Ma 等人[15]成功地将这一策略用于数量型性状的基因互作检测中。

本文中，研究提出一种基于机器学习算法 eXtreme Gradient Boost(XGBoost)的相互作用检测方法 geXGB(gene-base exchanged eXtreme Gradient Boost)。该方法使用交换策略产生新的测试数据集，并通过计算该测试集在训练过的 XGBoost 模型上的预测值与加和模型之间的偏差来度量相互作用关系的强度。geXGB 无需对相互作用显式建模，因此可以检测到更多类型的互作关系。此外，geXGB 作为一个非参数化模型，在数据驱动的全基因组关联研究中的应用更为灵活有效。

1　方法

1.1　基因互作检测问题描述

本文以基因作为基本研究单位。假设基因 $\boldsymbol{G}_i$ 包含 p 个 SNP 位点，每个 SNP 位点由 2 种碱基构成，即：主等位基因(major allele)与次等位基因(minor allele)，分别记为 A 和 a。因此每个 SNP 存在 3 种可能的基因型 AA，Aa，aa，使用 0，1，2 编码上述 3 种基因型。研究中将 2 个基因之间不存在相互作用定义如下：

定义　令 $\boldsymbol{G}_i, i=1,2,\dots,n$ 表示 n 个基因组成的序列，随机变量 $y \in \{0,1\}$ 表示质量性状表型值。记 $l(\boldsymbol{G}_1,\dots,\boldsymbol{G}_n)$ 是对数优势比(log odds ratio)。其对应数学公式可表示为：

$$l(\boldsymbol{G}_1,\dots,\boldsymbol{G}_n)=\log(P(y=1\mid \boldsymbol{G}_1,\dots,\boldsymbol{G}_n)/P(y=0\mid \boldsymbol{G}_1,\dots,\boldsymbol{G}_n)), \tag{1}$$

当 $\boldsymbol{G}_j$ 与 $\boldsymbol{G}_k$ 之间的对数优势比可以写作只依赖于单个基因的函数和时，称 $\boldsymbol{G}_j$ 与 $\boldsymbol{G}_k$ 之间没有相互作用关系。

由上述定义可以获得以下性质。

性质　如果 $\boldsymbol{G}_j$ 与 $\boldsymbol{G}_k$ 之间不存在互作，记 N 是样本个数，$l(\boldsymbol{G}_j,\boldsymbol{G}_k)$ 是对数似然比，$\boldsymbol{G}_j^1,\dots,\boldsymbol{G}_j^N$ 是 $\boldsymbol{G}_j$ 在 N 个样本的取值，$\boldsymbol{G}_k^1,\dots,\boldsymbol{G}_k^N$ 是 $\boldsymbol{G}_k$ 在 N 个样本的取值。记 $l_{s,t}=l(\boldsymbol{G}_1,\dots,\boldsymbol{G}_{j-1},\boldsymbol{G}_j^s,\boldsymbol{G}_{j+1},\dots,\boldsymbol{G}_{k-1},\boldsymbol{G}_k^t,\boldsymbol{G}_{k+1},\dots)$，则有：

$$\frac{1}{N}\Big(\sum_s l_{s,1}+\sum_t l_{1,t}\Big)-\frac{1}{N^2}\sum_{s,t} l_{s,t}=l_{1,1}. \tag{2}$$

证明　根据没有相互作用的定义，$l_{s,t}$ 可以写作加和形式 $l_{s,t}=f_s+g_t$，因此：

$$\frac{1}{N}\Big(\sum_s l_{s,1}+\sum_t l_{1,t}\Big)-\frac{1}{N^2}\sum_{s,t} l_{s,t}=\frac{1}{N}\Big(\sum_s (f_s+g_1)+\sum_t (f_1+g_t)\Big)-\frac{1}{N^2}\sum_{s,t}(f_s+g_t)=f_1+g_1=l_{1,1}$$

1.2　XGBoost 方法概述

XGBoost 是一种基于梯度提升决策树(gradient boosting decision tree，GBDT)扩展的有监督机器学习方法[16]，在保证 GBDT 性能的同时优化其计算效率，是近几年 Kaggle 数据竞赛的主流算法。本节中，将简要介绍 XGBoost 在分类问题上的算法原理。

1.2.1 以 CARTs 为基分类器的集成方法

在集成学习部分,使用 CART(classifying and regression tree)作为基分类器。CART 类似决策树,区别在于叶子结点。决策树的叶节点分配的是类别信息,而 CART 则是实值得分。这使得算法的整体训练更加容易,且可以额外获得除类别外的更多信息。

记 F 是 CARTs 可表示的函数空间集合,集成算法预测值为 $\hat{y} = \sum_k f_k, f_k \in F$。在逻辑回归中有:

$$p(y = 1 \mid x) = \frac{1}{1 + e^{-\hat{y}(x)}}, \tag{3}$$

因此,集成学习的目标函数是:

$$obj = \sum_i l(y_i, \hat{y}(x_i)) + \sum_k \Omega(f_k). \tag{4}$$

其中, $l(y, \hat{y}) = y\log(1 + e^{-\hat{y}}) + (1 - y)\log(1 + e^{\hat{y}})$ 是逻辑回归的交叉熵损失函数, $\Omega(f_k)$ 是正则惩罚项。

1.2.2 梯度提升

XGBoost 并不是一次训练所有的树,而是采用递进的训练策略,即固定已经学习过的树,每次逐步增加新的树。记 $\hat{y}^{(t)}$ 为第 t 次迭代训练后的预测值,则:

$$\hat{y}^{(0)} = 0, \tag{5}$$

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + f_t, \tag{6}$$

其中, $f_t \in F$ 用于优化以下目标函数,该目标函数通过将逻辑回归中的损失函数进行二次泰勒展开获得:

$$obj^t = \sum_i \left[g_i(\hat{y}^{(t-1)}(x_i)) f_t(x_i) + \frac{h_i^2(\hat{y}^{(t-1)}(x_i))}{2} \cdot f_t^2(x_i) \right] + \Omega(f_t). \tag{7}$$

其中, $g_i(\hat{y}) = \frac{\mathrm{d}}{\mathrm{d}\hat{y}} l(y_i, \hat{y})$, $h_i(\hat{y}) = \frac{\mathrm{d}}{\mathrm{d}\hat{y}} l(y_i, \hat{y})$。

1.2.3 CARTs 的训练策略

对于任意的 $f \in F$, 记 T 表示 f 函数的树模型的叶子结点个数, $w_1, \cdots, w_T$ 是每个叶节点的分数。则正则惩罚项可写作:

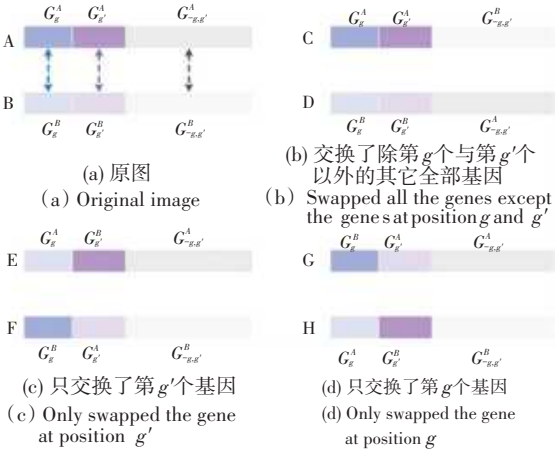
$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2. \tag{8}$$

其中,第二项的目的是为了平滑叶节点的得分。

可以看到, obj^t 是关于叶节点得分 w_j 的二次函数,当知道树结构时, obj^t 的最小值以及 obj^t 取得最小值时 w_j 的取值都可以确定。因此,为了优化 f_t,可以利用贪心算法获取树的结构。首先构造只有一个叶节点的树,然后沿着 obj^t 梯度的反方向重复地划分其叶节点。

1.3 基于 XGBoost 的基因互作检测方法 geXGB

根据前两节对相互作用的定义以及 XGBoost 的介绍可知,可以使用 XGBoost 建模数据,获得 l 的估计值 $\hat{l}$, 然后利用式(2)两边的差值来度量 2 个基因相互作用的强度。为了降低算法估计的误差,研究希望测试集中 $\boldsymbol{G}_j^s$ 和 $\boldsymbol{G}_k^t$ 的分布尽可能与训练集中 $\boldsymbol{G}_j$ 和 $\boldsymbol{G}_k$ 的分布相近,因此设计了随机置换策略(见图 1)。图 1 中, A, B 是原始数据集中的 2 个样本,通过按照箭头所示的方式交换 2 个样本中的部分数据,就获得了新的基因型数据样本 C, D, E, F, G, H。



(a) 原图 (a) Original image

(b) 交换了除第 g 个与第 g' 个以外的其它全部基因 (b) Swapped all the genes except the genes at position g and g'

(c) 只交换了第 g' 个基因 (c) Only swapped the gene at position g'

(d) 只交换了第 g 个基因 (d) Only swapped the gene at position g

图 1 生成测试集使用的交换策略示意图

Fig. 1 Schematic diagram of the exchange strategy used to generate the test set

geXGB 方法流程示意图如图 2 所示。由图 2 可知,geXGB 方法的流程步骤详述如下。

算法 基于 XGBoost 的基因互作检测方法

输入:基因型数据矩阵 $\boldsymbol{G}$、表现型数据向量 $\boldsymbol{y}$

输出:所有基因对相互作用强度列表,按照差异值降序排列

Step 1 利用 5 重交叉验证对训练 XGBoost 模型,获取最佳模型参数。

Step 2 令所有 $diff_{j,k}$ 为 0,其中 $1 \leqslant j < k \leqslant n$。

Step 3 随机将原始数据集分成训练集与测试集。

Step 4 利用获取的 XGBoost 参数,在训练集上获取优势对数比 $\hat{l}$。

Step 5　For $1 \leqslant j < k \leqslant n$, do:

Step 6　将测试集中的 $\boldsymbol{G}_j^s$ 的 N 元组随机重排，记 $\hat{l}_{s,t} = l(\boldsymbol{G}_1^1,\ldots,\boldsymbol{G}_{j-1}^1,\boldsymbol{G}_j^s,\boldsymbol{G}_{j+1}^1,\ldots,\boldsymbol{G}_{k-1}^1,\boldsymbol{G}_k^t,\boldsymbol{G}_{k+1}^1,\ldots)$。

Step 7　$\hat{I}'_{1,1} = \frac{1}{N}(\sum_s l_{s,1} + \sum_t l_{1,t}) - \frac{1}{N^2}\sum_{s,t} l_{s,t}$，计算 $diff_k = diff_{j,k} = diff_{j,k} + |\hat{l}_{1,1} - l_{1,1}|$。

Step 8　End。

Step 9　重复 Step 5～ Step 7，获得所有基因对的 $diff_{j,k}$，并按照该值从大到小的顺序输出。

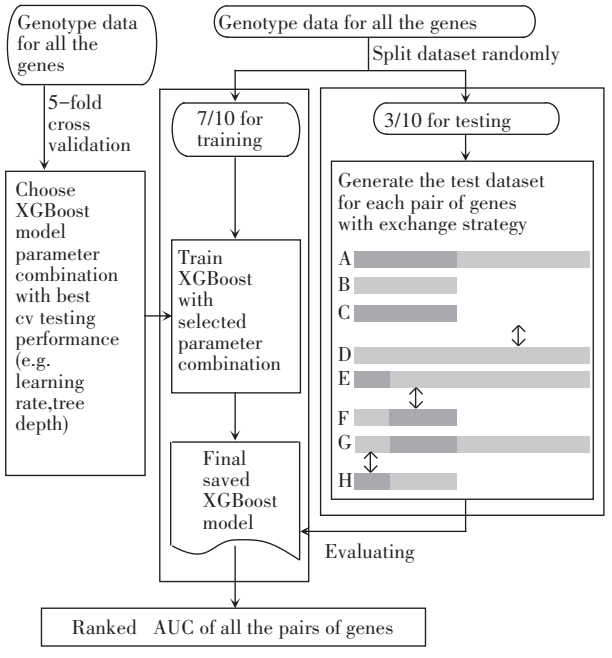


图 2　geXGB 方法流程示意图

Fig. 2　Flow diagram of geXGB method

2　实验与结果

2.1　仿真数据生成

为了评估 geXGB 方法检测基因-基因相互作用性能，所有的模拟数据集中均设置了 50 个 SNPs，其中有 2 个 SNP 之间有相互作用，此外 48 个 SNPs 是随机生成的。50 个 SNPs 被分为 5 个基因，每个基因包含 10 个 SNPs。2 个相互作用的 SNP 被分在不同的基因里。本次研究按照方法是否可以将相互作用的基因排在第一位来衡量方法的性能。模拟中使用 GAMETES 软件[17]来生成基因型数据，该工具可以生成严格的相互作用模型，即相互作用的 2 个基因均不存在主效应。

模拟实验中，为了研究遗传率和样本大小对方法性能的影响，研究设置了 2 组不同的实验环境。第一类情况下，测试了 5 个不同的遗传率值(0.01，0.025，0.05，0.1，0.2)和 2 种不同的次等位基因频率(minor allele frequence，*MAF*) 取(0.2，0.4)。这些模型的患病率均设置为 0.2，样本大小设置为3 000；对于遗传率和 *MAF* 的 10 种参数组合，均生成 10 个模型，由此获得 100 个模型。对于每个模型生成 100 个数据集，由此共获得 10 000 个数据集。第二类情况下，固定遗传率为 0.025，*MAF* 为 0.2 和 0.4，患病率为 0.2，样本量为 10 000。然后从 10 000 个样本中按照不同的样本大小无放回抽取样本生成新的样本集用于考察样本大小对方法性能的影响。数据集大小分别为 2 000，3 000，4 000 和 5 000。每个数据大小均生成 100 个数据集。

2.2　仿真实验结果

实验中选用了 3 种基于基因的基因互作检测方法作为对比方法，分别是：KCCU[11-12]，AGGrEGATOr[14]和 GBIGM[13]。对于每个模型下的 100 个数据集，如果方法将相互作用的一对基因排在第一位，则算作选中。方法在每个模型的统计效力用选中数据集的百分比来表示。

第一类模拟情况下各方法的统计效力值见表 1。图 3 是表 1 数据的盒图。图 4 为 4 种方法在不同模型下的平均效力比较。表 1 中，粗体为每个模型下最优的方法效力值，值越大表明方法检测性能越好。由图 4 可知，geXGB 具有最优的平均性能，在多数模型下都大幅超越其它对比方法。AGGrEGATOr 在 *MAF* = 0.2 且遗传率大于 0.05 的情况下可以达到与 geXGB 几乎相同的性能。但在更小的遗传率情况下，geXGB 表现出更好的检测性能。当遗传率为 0.01，*MAF* = 0.2 时，在 6 个模型上排位第一，在 3 个模型上排位第一；而相同遗传率情况下，当 *MAF* = 0.4时，geXGB 与 AGGrEGATOr 排位第一的模型个数比为 9∶2。当 *MAF* = 0.4 时，AGGrEGATOr 在各模型下的平均效力要高于 KCCU。但在某些模型下，当 AGGrEGATOr 效果不好时，统计效力甚至比 KCCU 还要低。从图 3 可以看出，相较于 geXGB，AGGrEGATOr 方法在各模型上的效力浮动各大，而 geXGB 则更为稳定。

此外，由图 4 可知 KCCU 与 AGGrEGTOr 具有相似的性能模式，但 AGGrEGTOr 普遍优于 KCCU。GBIGM 几乎无法检测到此类严格的相互作用关系，这个结果与 Emily 的模拟结果一致。

表 1 4 种对比方法在第一类模拟数据设置下的统计效力

Tab. 1 The statistical effectiveness comparison of four methods under the setting of the first type of simulated data

MAF	Heritability	Method	Model									
			M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
0.2	0.01	geXGB	0.14	**0.17**	**0.58**	0.75	**0.48**	**0.38**	0.71	0.91	**0.93**	**0.49**
		AGGrEGATOr	0.12	0.14	0.12	**0.89**	0.12	0.10	**0.89**	**1**	0.88	0.34
		KCCU	**0.15**	0.09	0.09	0.29	0.14	0.10	0.43	0.62	0.52	0.13
		GBIGM	0.09	0.08	0.11	0.13	0.12	0.17	0.11	0.08	0.10	0.09
	0.025	geXGB	0.98	**0.97**	**0.94**	**1**	**1**	**0.99**	**0.99**	**1**	**0.90**	**0.94**
		AGGrEGATOr	**1**	0.15	0.27	**1**	**1**	0.46	0.37	**1**	0.69	0.81
		KCCU	0.58	0.09	0.09	0.74	0.71	0.59	0.24	0.80	0.12	0.12
		GBIGM	0.08	0.11	0.07	0.11	0.10	0.12	0.13	0.20	0.14	0.10
	0.05	geXGB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		AGGrEGATOr	0.09	0.09	0.59	0.89	0.98	0.99	**1**	**1**	**1**	**1**
		KCCU	0.13	0.10	0.57	0.65	0.71	0.80	0.84	0.85	0.84	0.77
		GBIGM	0.18	0.16	0.08	0.22	0.23	0.12	0.17	0.19	0.12	0.12
	0.10	geXGB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		AGGrEGATOr	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		KCCU	0.81	0.88	0.93	0.90	0.90	0.83	0.86	0.91	0.84	0.93
		GBIGM	0.15	0.14	0.14	0.23	0.19	0.17	0.16	0.16	0.15	0.19
	0.20	geXGB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		AGGrEGATOr	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		KCCU	0.89	0.91	0.97	0.94	0.95	0.92	0.89	0.97	0.94	0.97
		GBIGM	0.19	0.21	0.31	0.18	0.29	0.23	0.22	0.21	0.23	0.22
0.4	0.01	geXGB	**0.90**	**0.74**	**0.82**	**0.82**	**0.83**	**0.96**	**0.96**	**0.66**	0.82	**0.89**
		AGGrEGATOr	0.71	0.73	0.09	0.10	0.77	0.96	0.94	0.17	**0.90**	0.81
		KCCU	0.34	0.34	0.05	0.08	0.40	0.29	0.77	0.16	0.73	0.61
		GBIGM	0.09	0.11	0.08	0.10	0.11	0.07	0.11	0.18	0.11	0.11
	0.025	geXGB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		AGGrEGATOr	0.99	**1**	0.56	0.12	0.06	0.26	0.91	0.51	0.12	**1**
		KCCU	0.58	**1**	0.24	0.08	0.06	0.11	0.24	0.32	0.12	**1**
		GBIGM	0.15	0.10	0.12	0.14	0.08	0.09	0.11	0.08	0.05	0.08
	0.05	geXGB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		AGGrEGATOr	**1**	0.68	0.97	0.91	0.15	0.42	0.35	0.19	0.91	**1**
		KCCU	0.86	0.15	0.90	0.95	0.10	0.37	0.41	0.14	0.74	**1**
		GBIGM	0.11	0.07	0.12	0.09	0.13	0.10	0.08	0.08	0.13	0.16
	0.1	geXGB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		AGGrEGATOr	0.98	0.06	**1**	0.96	**1**	**1**	**1**	**1**	**1**	**1**
		KCCU	0.62	0.17	**1**	0.95	**1**	**1**	**1**	**1**	**1**	**1**
		GBIGM	0.12	0.17	0.19	0.18	0.12	0.20	0.13	0.10	0.19	0.12
	0.2	geXGB	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		AGGrEGATOr	0.93	**1**	**1**	0.99	**1**	0.80	0.27	0.09	0.14	0.12
		KCCU	0.28	**1**	**1**	0.83	**1**	0.76	0.41	0.15	0.28	0.12
		GBIGM	0.19	0.20	0.25	0.31	0.23	0.26	0.26	0.29	0.22	0.10

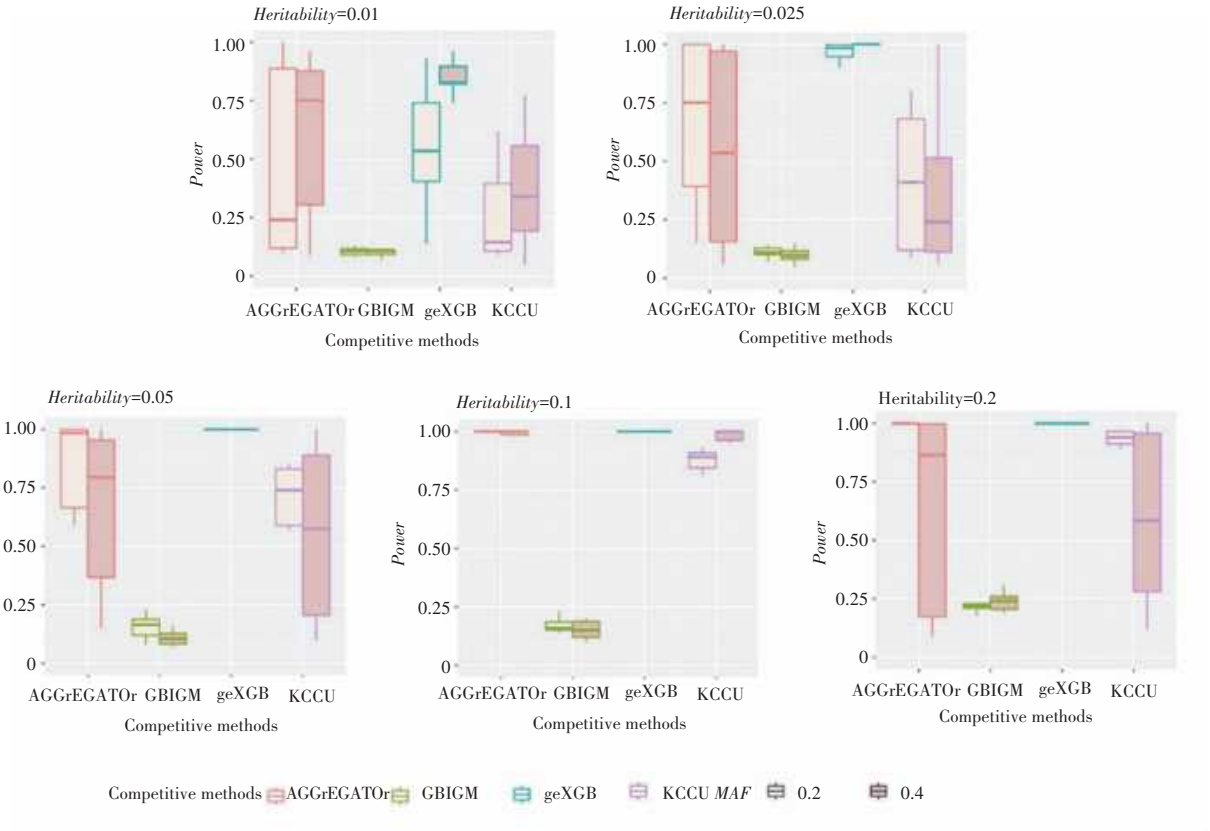


图 3　4 种对比方法在第一类模拟数据设置下的功效盒图

Fig. 3　The efficiency box comparison diagram of the four methods under the setting of the first type of simulation data

geXGB 算法的性能很大程度上取决于遗传率，即交互作用强度。随着遗传率从 0.01 到 0.025，geXGB 的统计效力成倍增加。其它方法也表现出了稳定的上升趋势(见图 4)。不仅如此，方法的性能还取决于相互作用位点的 *MAF*，例如 *MAF* = 0.2 时，不同模拟数据集上统计效力范围为 0.14～0.93 之间，而 *MAF* = 0.4 时，范围在 0.66～0.93 之间。前者平均效力为 0.554，远低于后者的 0.84。表 2 是 4 种方法在不同样本规模下的功效值，可以看出，增大样本量对方法性能有显著提升作用。

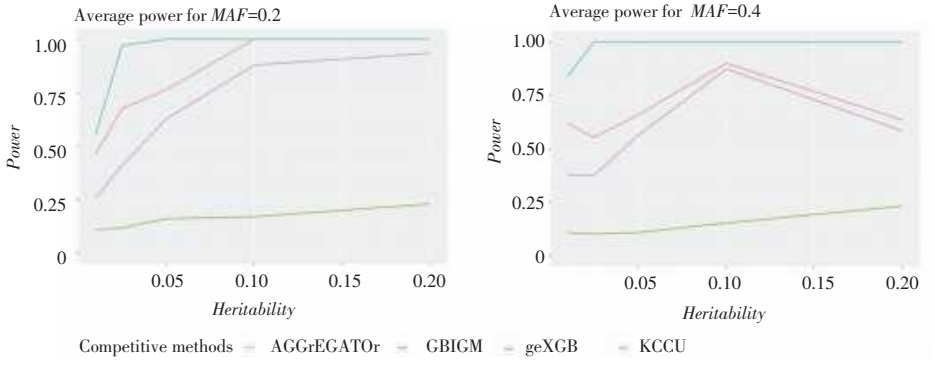


图 4　4 种方法在第一类模拟数据设置下的平均功效

Fig. 4　The average power of four methods under the setting of the first type of simulated data

表 2 4 种方法在不同样本规模下的功效值

Tab. 2 Power values of four methods under different sample sizes

Heritability	Sample Size	Method			
		geXGB	AGGrE-GATOr	KCCU	GBIGM
0.2	1 000	0.67	0.15	0.11	0.20
	2 000	1	0.18	0.38	0.16
	3 000	1	0.11	0.55	0.23
	4 000	1	0.31	0.76	0.21
	5 000	1	0.26	0.87	0.12
0.4	1 000	0.68	0.18	0.13	0
	2 000	0.97	0.16	0.11	0.04
	3 000	1	0.35	0.20	0.11
	4 000	1	0.54	0.37	0.11
	5 000	1	0.65	0.58	0.05

由模拟实验结果可知,本文提出的 geXGB 是一种十分有效的基因互作检测方法。较之其他对比方法,geXGB 可以适用于更为广泛的遗传模型下基因互作的检测。

3 结束语

检测基因-基因相互作用的研究在阐明人类复杂疾病致病机理方面具有重要意义。本文提出一种基于 XGBoost 的方法 geXGB 用于检验基因间相互作用。研究定义基因型数据的对数优势比,将基因之间的互作转化为基因联合的对数优势比与单独基因函数之和之间的偏差。这一假设对基因之间互作形式没有限定,增强了方法可检测基因相互作用的类型。仿真数据实验结果表明,geXGB 在遗传率、*MAF* 与样本规模三个参数的多种组合设定下,均有优于其它对比方法的统计效力,且方法效力随遗传率、*MAF* 和样本规模的增大呈现单调递增趋势。以上结果表明该方法在基因互作检测中的有效性。

参考文献

[1] EMILY M. IndOR: A new statistical procedure to test for SNP-SNP epistasis in genome-wide association studies[J]. Statistics in Medicine, 2012, 31(21): 2359.

[2] WU Xuesen, DONG Hua, LUO Li, et al. A novel statistic for genome-wide interaction analysis[J]. PLoS Genetics, 2010, 6(9): e1001131.

[3] UEKI M, CORDELLL H J. Improved statistics for genome-wide interaction analysis[J]. PLoS Genetics, 2012, 8(4): e1002625.

[4] DONG Changzheng, CHU Xun, WANG Ying, et al. Exploration of gene - gene interaction effects using entropy-based methods [J]. European Journal of Human Genetics: EJHG, 2008, 16(2): 229.

[5] KANG Guolian, YUE Weihua, ZHANG Jifeng, et al. An entropy-based approach for testing genetic epistasis underlying complex diseases[J]. Journal of Theoretical Biology, 2008, 250(2): 362.

[6] RITCHIE M D, HAHN L W, MOORE J H. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity[J]. Genetic Epidemiology, 2003, 24(2): 150.

[7] ZHANG Xiang, HUANG Shunping, ZOU Fei, et al. TEAM: Efficient two - locus epistasis tests in human genome - wide association study[J]. Bioinformatics, 2010, 26(12): i217.

[8] WAN Xiang, YANG Can, YANG Qiang, et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies[J]. American Journal of Human Genetics, 2010, 87(3): 325.

[9] CORDELL H J. Detecting gene-gene interactions that underlie human diseases[J]. Nature Reviews Genetics, 2009, 10(6): 392.

[10] PENG Qianqian, ZHAO Jinghua, XUE Fuzhong. A gene-based method for detecting gene-gene co-association in a case-control association study [J]. European Journal of Human Genetics: EJHG, 2010, 18(5): 582.

[11] LARSON N B, JENKINS G D, LARSON M C, et al. Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer[J]. European Journal of Human Genetics: EJHG, 2014, 22(1): 126.

[12] YUAN Zhongshang, GAO Qingsong, HE Yungang, et al. Detection for gene - gene co - association via kernel canonical correlation analysis[J]. BMC Genetics, 2012, 13: 83.

[13] JIN Li, HUANG Dongli, GUO Mazu, et al. A gene-based information gain method for detecting gene-gene interactions in case-control studies[J]. European Journal of Human Genetics, 2015, 23(11): 1566.

[14] EMILY M. AGGrEGATOr: A gene - based gene - gene interActTiOn test for case - control association studies [J]. Statistical Applications in Genetics and Molecular Biology, 2016, 15(2): 151.

[15] MA L, CLARK A G, KEINAN A. Gene-based testing of interactions in association studies of quantitative traits[J]. PLoS Genet, 2013, 9(2): e1003321.

[16] CHEN Tianqi, GUESTRIN C. XGBoost: A scalable tree Boosting system[C]//the 22nd ACM SIGKDD International Conference.San Francisco, CA, USA: ACM, 2016: 785.

[17] URBANOWICZ R J, KIRALIS J, SINNOTT-ARMSTRONG N A, et al. GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures[J]. BioData Mining, 2012, 5(1): 16.