遗传 Hereditas (Beijing) 2018 年 3 月, 40(3): 218－226
www.chinagene.cn

综 述

机器学习方法在基因交互作用探测中的研究进展

彭哲也，唐紫珺，谢民主

湖南师范大学物理与信息科学学院，长沙 410081

摘要：复杂疾病是基因与基因、基因与环境交互作用的结果，高维基因交互作用的探测给计算带来了极大的挑战。在过去 20 年间，机器学习方法被用于探测基因–基因交互作用，并取得了一定的效果。本文综述了机器学习方法在基因交互作用探测中的研究进展，系统地介绍了神经网络(neural networks, NN)、随机森林(random forest, RF)、支持向量机(support vector machines, SVM)和多因子降维法(multifactor dimensionality reduction, MDR)等机器学习方法在全基因组关联研究(genome wide association study, GWAS)中探测基因交互作用的原理和局限性，并对未来的研究进行了展望。

关键词：机器学习；基因交互；全基因组关联分析；单核苷酸多态性；上位性

Research progress in machine learning methods for gene-gene interaction detection

Zheye Peng, Zijun Tang, Minzhu Xie

College of Physics and Information Science, Hunan Normal University, Changsha 410081, China

Abstract: Complex diseases are results of gene-gene and gene-environment interactions. However, the detection of high-dimensional gene-gene interactions is computationally challenging. In the last two decades, machine-learning approaches have been developed to detect gene-gene interactions with some successes. In this review, we summarize the progress in research on machine learning methods, as applied to gene-gene interaction detection. It systematically examines the principles and limitations of the current machine learning methods used in genome wide association studies (GWAS) to detect gene-gene interactions, such as neural networks (NN), random forest (RF), support vector machines (SVM) and multifactor dimensionality reduction (MDR), and provides some insights on the future research directions in the field.

Keywords: machine learning; gene-gene interactions; genome wide association studies; single nucleotide polymorphism; epistasis

收稿日期: 2017−09−20; **修回日期:** 2017−12−28
基金项目: 国家自然科学基金(编号 61772197, 61370172)资助[Supported by the National Natural Science Foundation of China (Nos. 61772197, 61370172)]
作者简介: 彭哲也，硕士研究生，专业方向：生物信息学。E-mail: hnsfdxpzy@yeah.net
唐紫珺，硕士研究生，专业方向：生物信息学。E-mail: tangzijun0531@yeah.net
彭哲也和唐紫珺并列第一作者。
通讯作者: 谢民主，教授，博士生导师，研究方向：生物信息学。E-mail: xieminzhu@sina.com
DOI: 10.16288/j.yczz.17-254
网络出版时间: 2018/2/1 13:25:26
URI: http://kns.cnki.net/kcms/detail/11.1913.R.20180201.1325.001.html

全基因组关联研究(genome wide association studies, GWAS)在全基因组范围内检测 DNA 变异与特定疾病或者性状之间的相关性，进而发现与之相关的遗传变异。目前，全基因组关联研究已经发现了与各种疾病或性状(表型)相关的数千个单核苷酸多态性(single nucleotide polymorphism, SNP)位点。然而，对于绝大部分复杂疾病而言，这些 SNP 位点上的变异导致的患病风险增加往往较小，即只有少部分人可以用这些位点上的变异解释其患病状态，这种现象被称为“遗传性缺失(missing heritability)”[1]。对于这种现象，研究人员提出了多种解释，其中被广泛认同的解释是：复杂疾病是由基因与基因，基因与环境之间的交互作用引起的，其中基因与基因的交互作用通常表现为 SNP 位点的上位性效应，即两个或两个以上的 SNP 位点对表型具有协同影响[2]。单个 SNP 通过改变单个基因的表达对疾病发病率的影响通常会很小，新出现的证据表明，许多稀有的 DNA 变异与多个风险等位基因的交互作用会导致患病风险增加[3]。而目前的全基因组关联研究主要探测单个 SNP 位点与疾病的相关性，缺乏探测多个基因交互作用的能力。

探测基因交互作用有助于识别基因功能，对发现潜在的药物靶点和人类复杂疾病的遗传机理尤为重要[4]。探测基因交互作用通常采用的方法是计算多个 SNP 位点上的等位基因组合与表型的统计相关性。但随着交互作用基因数目的增加，对应 SNP 位点上可能的等位基因组合数量呈指数增长，假定每个 SNP 位点上可能的基因型不同取值为 3，则 n 个 SNP 位点上不同的基因型组合数量高达 3^n。探测高维基因交互作用在计算上面临巨大的挑战[5]。机器学习(machine learning, ML)是让计算机模拟人类认知过程对问题进行求解的一种方法，利用机器学习方法探测基因交互作用的优点是不需要事先假定位点或基因间交互作用的模型，不是通过穷尽搜索而是让模拟人类认知过程的计算机算法通过大量数据进行学习，从而获得发现非线性高维交互作用的能力[6]。近 20 年来，众多机器学习方法已被用于基因交互作用探测，并取得了一定的成功[7]，然而遗传异质性、群体分层和涉及交互作用的 SNP 位点数量众多是影响机器学习方法探测基因交互作用性能的主要因素。本文将对探测基因交互作用的机器学习方法进行综述，并对未来研究方向进行展望。

1　机器学习方法的原理和特点

在过去的 20 年中，一系列机器学习方法被用来探测基因与基因的交互作用。目前，已经应用于基因交互探测的机器学习方法主要包括神经网络(neural networks, NN)，随机森林(random forest, RF)，支持向量机(support vector machines, SVM)和多因子降维法(multifactor dimensionality reduction, MDR)，本节将综述这些机器学习方法的原理和特点，及其在基因交互作用的探测上取得的一些成果。

1.1　神经网络

神经网络基于神经元模型，其中“前馈/反向传播”神经网络最为常见，它具有出色的模式识别和模式分类能力，并且能够处理大量数据[8]。神经网络的结构采用多层有向图，前馈神经网络由输入层、隐藏层和输出层组成[9]，输入层和隐藏层包含众多节点，而输出层只有一个节点。当用于探测基因和基因的交互作用时，输入层的节点代表遗传变异，通常是 SNP，输入层节点通过弧(有向边)连接隐藏层的节点，最后隐藏层的节点通过弧连接输出节点，控制输出节点的输出。神经网络中每条弧对应一个不同的权重，神经网络中弧的不同权重配置对应着 SNP 之间不同的交互作用，弧的权重是通过大量数据对神经网络进行训练得到。对神经网络进行训练时，每个弧被分配一个随机的初始权重，使用训练数据已知的基因型给输入层的节点设置输入值，然后观察神经网络输出节点的状态，根据该状态与已知的表型的差异，对弧的权重进行调整，期望使神经网络输出的错误率达到最小[10]。最终分析训练后的神经网络的内部权重结构，识别真实数据中所隐含的基因交互模式[11,12]。

Tomita 等[13]利用神经网络对 172 个患过敏性哮喘的儿童和 172 个对照组的正常人的 17 个基因的 25 个 SNPs 进行基因交互作用分析，发现了日本人群中与过敏性哮喘相关的 10 个易感 SNPs，测试结果显示总准确率达到了 74.4%。

构建合适的神经网络内部权重结构是探测基因与基因交互作用成功的关键。目前神经网络内部权重的构造方法有反向传播(back propagation, BP)、遗传编程(genetic programming, GP)和语法演化(grammatical evolution, GE)[14~18]。Ritchie 等[15]比较了遗传编程神经网络(GPNN)和反向传播神经网络(BPNN)探测基因与基因交互作用的能力，其结果表明当测试数据包含功能性和非功能性的 SNPs 时, GPNN 表现优于 BPNN[16]。Motsinger 等[17]也对 GPNN 探测基因–基因交互作用的能力进行了测试：对 1600 个样本(case 和 control 各占 1/2, SNP 位点总数为 10 个)中的 2 个 SNP 位点交互作用的探测结果显示, GPNN 对遗传效应(heritability)低至 0.5%的基因–基因交互模型的探测能力也达到了 86%；在真实的帕金森病的数据上 GPNN 也探测到线粒体基因与性别(mitochondrial gene-sex)的交互作用，该交互作用导致帕金森病发病率显著上升。Campos 等[18]利用语法演化技术对 GPNN 进行改进提出了 GENN 神经网络，用于存在噪声情况下基因–基因交互作用的探测。GENN 利用进化搜索策略，并在语法中使用布尔运算，在模拟数据上的测试显示 GENN 在处理基因分型错误和数据遗漏等问题上具有很强的鲁棒性。

1.2 随机森林

随机森林是由 Leo Breiman 提出[19]，是一种由随机向量生成的分类树或回归树的集合所构成的高维非参数预测模型，包括 4 个主要部分：随机选择样本；随机选择特征；构建决策树；随机森林投票分类。随机森林通过自助法(bootstrap)重采样技术进行采样，给定一个训练样本集，数量为 N，使用有放回的采样得到 N 个样本，从而构成一个新的训练集。随机森林的优点在于它们不会“过度拟合(overfit)”数据，随着随机森林中的树的数量增加，预测误差将不会超过一个给定值[20]。

随机森林为每个 SNP 提供重要性分数，使其能识别与表型相关的 SNPs，进而探测交互作用的 SNPs[21]。随机森林方法在基因交互作用的探测中有很多成功的应用[21~25]。Chen 等[23]使用随机森林方法对遗传性球形细胞增多症(hereditary spherocytosis, HS)的相关数据进行分析，探测到了 41 个已知的与 HS 相关的基因，发现了 150 个新的与 HS 相关的基因及这些基因构成的交互网络中的核心基因。Bureau 等[24]利用随机森林从 131 个哮喘病人和 217 个正常人的 42 个 SNP 数据中找到了能有效预测哮喘病的 SNP 对 ST+4 和 BC+1。

随机森林算法是一种有效的分类工具，具有发现没有强主效应的基因之间交互作用的潜力，在低维数据(100 个 SNP 和 10 000 个观测值)中已经显示出较好地性能，然而，它们探测交互作用的能力实际上取决于主效应是否存在，不管存在的主效应是多么弱，因此，这种方法可能缺乏发现没有任何主效应的基因之间的交互作用的能力[26]。

SNPInterForest 是对随机森林方法改进而来的，它在发现与疾病相关的 SNP 的能力比随机森林更强，并且具有同时识别多种交互作用的能力，SNPInterForest 对具有主效应的 SNP 比随机森林更为敏感[27]。Pan 等[28]把随机森林和互信息网络(mutual information network, MIN)集成，提出了互信息网络引导的随机森林方法 MINGRF(MIN guided RF)，其目的是减少边际效应对 RF 的影响。

1.3 支持向量机

支持向量机，也称为支持向量网络[29]，是一种监督式的机器学习方法，用于求解二分问题(binary classification)，广泛应用于分类和回归分析(regression)。支持向量机通常是设计一个合理的核函数，对数据进行变换，通过已知类别的数据对向量机进行训练，在变换的空间寻找一个超平面，期望能最大限度地把不同类别的数据隔离在超平面的两侧。支持向量机的学习过程其实是寻求一个既能最小化经验损失、又能最大化不同类别数据之间的几何间距的超平面的过程，因此 SVM 又被称为最大间距分类器。SVM 可以通过学习已知存在交互作用的基因的特点，来预测哪些基因在遗传上有交互作用。为了实现这一点，支持向量机的训练数据是两组特征向量，它们被标记为阳性(存在遗传交互作用)和阴性(无遗传交互作用)，在模拟数据集和真实数据集上的测试都显示出 SVM 具有较强的探测基因交互作用能力[30~34]。

早在 2004 年, Listgarten 等[32]利用 SVM 鉴定出许多与乳腺癌风险相关的基因变异，该文结果表明，

当使用具有二次核函数的 SVM 预测乳腺癌患者时，多个 SNP 位点的组合比单一 SNP 位点预测乳腺癌患者的精度更高。Chen 等[33]把 SVM 和局部搜索、遗传算法结合起来构建了一个探测基因交互作用的平台，在大量模拟数据上的测试结果表明该平台虽然需要较大的计算资源，但该平台能在 case 和 control 两组人数严重不对称的数据也能有效探测高维的基因交互作用。

Shen 等[34]提出了一种两阶段探测基因与基因的交互作用的方法。第一阶段，Shen 等利用 L1 惩罚 SVM(模型选择法)识别最有可能有交互作用的 SNP 位点；第二阶段在第一阶段识别出的 SNP 位点的基础上，应用逻辑回归(logistic regression)和 Bonferroni 校正排除非候选 SNPs。结果表明，L1 惩罚 SVM 在病例对照组数据上的 SNP 交互作用探测是有效的，多变量 logistic 回归分析比传统的 logistic 回归分析对 SNP 的交互作用分析效果要好。Ban 等[35]利用 SVM 方法分析韩国 462 个 2 型的糖尿病人和 456 个正常人在 87 个基因上的 408 个 SNP 位点上基因型的数据集，获得了一个由 14 个 SNP 交互作用的组合，该组合识别糖尿病的准确率大于 70%。

1.4　多因子降维

2001 年 Ritchie 等[36]提出了一种分析基因交互作用方法–多因子降维法。MDR 是一种非参数的分析方法，适用于病例–对照组(case-control)研究，只需提供各遗传变异位点的遗传数据(如 SNP 等)，即可进行基因交互作用分析。在 MDR 的第一阶段，从数据集中选择 x 个变异位点(在 GWAS 中为 SNP 位点)，其中 x 为需要分析的交互作用的维数。对于 SNP 位点上的基因型数据而言，这 x 个位点上有 3^x 个不同的基因型组合，MDR 的第二阶段则用一个 3^x 行 2 列的列联表统计出在这 x 个变异位点上所有不同取值组合的病例人数和对照组人数。第三阶段，利用列联表，计算出每个基因型组合对应的病例人数与对照组人数的比值，若该比值大于某个阈值 t (例如 t =总病例人数/总对照组人数)，则标记为高危因子，反之则标记为低危因子，这样就把 x 维的数据精简到一维两水平(即高危或低危)的数据，获得了一个基于这 x 个变异位点预测疾病状态的基因交互作用模型，然后通过交叉验证该模型的精确度，选择预测误差最小的模型作为最终的模型。最后通过置换测试(permutation test)评价最终模型的统计显著性。

MDR 是一种无模式(model-free)的方法，不需提前对疾病模型进行假设，这使得 MDR 被大量用于分析发病机制未知的复杂疾病的遗传数据，获得了许多与复杂疾病相关的基因交互作用模型[37~45]，例如，Tsai 等[40]利用 MDR 方法发现了房颤中交互作用的基因对(*RAS-ACE*)，MDR 获得的最佳模型是由 3 个 SNP 组成，其中 2 个 SNP 来自 *RAS* 基因，1 个 SNP 来自 *ACE* 基因。这 3 个 SNP 的 10 重交叉验证显示有很好的一致性，100 次的置换测试得到的 *P*-value 为 0.001。

然而，在分析表型-遗传异质性率偏高(>50%)的遗传数据集时，MDR 发现基因交互作用模块的性能大大降低，尽管基因型组合分为“高危”或“低危”，但没有定量评价他们是危险程度，获得的最终模型很难解释[42]。MDR 可以很便捷地发现交互作用，但 MDR 却无法揭示主效应[43]。当基因型组合中的病例对照率与整个数据集病例对照率相近时，MDR 具有较高的假阳性和假阴性错误率[44]。为了解决这一问题，Leem 等[44]用最大似然度方法确定基因型组合的风险级别，提出了 EF-MDR(empirical fuzzy MDR, EF-MDR)。EF-MDR 在 WTCCC 的克罗恩病(Crohn's disease, CD)和躁郁症(bipolar disorder, BD)数据集中探测到了一些有趣的多 SNP 交互。

Gui 等[45]将 x 个位点上的基因型组合分为 3 组：高风险，低风险和未知风险，如果该组合上病例人数与对照组人数之比与所有病例人数与对照组人数之比相同或接近，则将其标记为未知风险，并从模型中排除，在此基础上提出了 RMDR(Robust MDR)。Gui 等使用膀胱癌数据集对 RMDR 和 MDR 进行测试，结果表明 RMDR 发现的基因交互模型更容易解释，其计算速度也较快。

为了使 MDR 能处理连续表型数据，Lou 等[46]提出了对 MDR 进行了扩展，提出了 GMDR(generalized MDR)。GMDR 用一个通用的线性模型表示表型数据，利用最大似然度估计确定多个位点上的基因型组合的风险类别，当数据除了包含基因型数据还包含其他协变量数据时，GMDR 能提高探测基因

交互作用的能力，并且能适用于随机采样获得的数据集。在此基础上，为了处理数据中的群体层化问题，Chen 等[47]提出了 UGMDR(unified GMDR)。

表 1 总结了目前用于探测基因交互作用的机器学习方法以及它们的优势和局限性。

2 现阶段模型的应用

全基因组关联研究在探测疾病相关的 SNP 上取得了大量的研究结果，但是在探测多基因的交互作用上还存在很多困难，这是由于基因组遗传数据具

表 1 机器学习方法的优势和局限性

Table 1 Advantages and limitations of machine learning methods

方法	优势	局限性	参考文献
Neural networks (NNs)	1. 优秀的模式识别/分类功能 2. 有能力处理大数据 3. 适应遗传异质性/多基因遗传/高表型率/不完全外显率	不能枚举所有可能的神经网络架构，并且改变架构会改变数据分析的结果，无法确定正在使用的架构是否是最佳的	[8]
GPNN	1. GP 优化的 NN 体系结构 2. 在非功能性 SNP 存在下，探测交互作用时具有较高效能 3. 当功能性 SNP 未知，且变量选择和模型拟合所需一样时，优选结果 4. 不会过度拟合数据 5. 在弱边际效应的上位模型中具有较高的效能 6. 模型灵活：不需要选择最优的输入，权重，连接或是隐形层	1. 在三位点的模型中具有高假阳性率 2. 需要并行计算环境 3. 输出是二元表示树，它可能很大(多至 500 个节点)，并难以解释	[15]
GENN	1. GE 优化的 NN 体系结构 2. 可用于从有噪声(例如，基因分型错误，缺失数据，拟表型，遗传异质性)的高维遗传病学数据中发现基因-基因交互作用	1. 数据集中拟表型的存在导致 GENN 的效果大大降低	[18]
RF	1. 能发现没有强主效应的基因之间的交互作用 2. 不会过度拟合数据，且误差收敛有上限值 3. 能鉴定预测表型的 SNP	1. 探测交互作用的能力取决于主效应 2. 无法探测没有边际效应的基因之间的相互作用 3. 从随机森林中提取有用的生物信息时相对困难	[19]
SNPInterForest	1. 可同时识别多个交互作用 2. 在没有边际效应时，不会低估 SNP 的重要性分数 3. 没有边际效应的情况下，每个节点上的多个 SNP 选择提高了探测疾病相关 SNP 的能力 4. 能评估 SNP 组合的交互作用强度 5. 具有较高的召回率和较低的假阳性率 6. 能发现存在遗传异质性的交互作用	计算量很大	[27]
SVM	1. 比 MDR 有更多可解释的输出结果 2. 可以应用到新的数据结构 3. 分类时无需用户自定义	1. 无法处理不完整的数据 2. 处理存在遗传异质性的数据时效能降低	[33]
MDR	1. 同时探测多个基因位点，保持低误报率 2. 无模式，适应于机制未知的遗传基因数据	1. 在高(50%)表型/遗传异质性下，检验效能显著降低 2. 当 SNP 的数量超过 10 时，需要大量的计算资源	[36]
RMDR	1. 获得的交互模型比较容易解释 2. 多位点上基因型组合模型分类为高风险、未知风险和低风险三类，降低了假阳性率	比 MDR 需要更大的计算资源	[45]
GMDR	1. 使用最大似然法给基因型组合模型分类 2. 给基因型组合模型分类是能考虑协变量的影响，可提高分类的准确性	比 MDR 需要更大的计算资源	[46~48]

有高度的异质性，还有拟表型、表型变异性和不完全外显率等诸多因素造成的[49]。机器学习法在探测基因交互作用上可以用来解决这些局限性，例如，随机森林方法能够成功处理某些类型的异质性的问题，神经网络的一些特性能够解决遗传异质性，多基因遗传，高拟表率和不完全外显的问题[50]。

帕金森病(Parkinson's disease, PD)是老年人常见的一种神经退化性疾病，在 65 岁以上的人口有约 2%的发病率，在 85 岁以上的老年人中，发病率上升至约 5%，目前帕金森病的发病机制尚不清楚，但有假设认为帕金森病是由影响能量代谢和蛋白质合成的复杂的基因–环境的交互作用导致的，Mellick 等[51]对 306 个 PD 病人和 321 个正常人测定了与线粒体复合体 I 相关的 31 个基因上的 70 个 SNP 数据，并进行了分析，没有发现单个 SNP 与 PD 有显著的统计相关性，而遗传编程神经网络(GPNN)则在该数据集中，探测到了 DLST 基因与性别之间的交互作用[17]。

唇裂，伴有或不伴有腭裂(CL/P)，是人类最常见的一种脸部先天性缺陷，非综合征型 CL/P 得到了广泛的研究，发现了大量与 CL/P 相关的候选基因组区域。Li 等[52]对 891 个亚洲裔 Trio(一个 Trio 由父亲、母亲和患有非综合征型 CL/P 的小孩组成)和 681 欧洲裔 Trio 的 SNP 数据进行了分析，他们利用随机森林(RF)探测与 WNT 信号通路相关的 18 个基因上 360 个 SNP 和其他候选基因组区域上 153 个 SNP 位点之间的交互作用，结果发现 WNT5B 和 MAFB 有显著的交互作用(亚洲裔 Trio 的 P =0.0076，欧洲裔 Trio 的 P = 0.018)。类风湿关节炎(rheumatoid arthritis，RA)是一种慢性的主要体现为炎性滑膜炎的系统性疾病。WTCCC 有一个 RA 数据集，该数据集包含了 3499 个人(1999 个 RA 患者，2000 个正常人)的 500K 个 SNP 数据。Yoshida 等[27]首先利用单位点关联分析方法从该数据集的 500K SNP 中选出 10K 个 SNP 位点，然后利用 SNPInterForest 探测这些 SNP 之间的交互作用。SNPInterForest 在 1 台 6GB 内存的计算机上运行 98 个小时候后发现了两个新的 SNP 交互作用(rs17665418，rs2121526)和(rs17665418，rs4799934)。rs17665418 位于 3p13，rs2121526 位于 10q21.1，而 rs4799934 位于 18q12.2。

在欧美国家，前列腺癌(prostatic cancer)的发病率高居男性肿瘤的首位，死亡率仅次于肺癌、结直肠癌。Chen 等[33]利用 SVM 方法分析来自瑞典的前列腺癌数据集，该数据集包含 1355 个病例和 765 个对照个体的位于 18 个基因中的 57 个 SNP 位点上的基因型数据，其中数据的缺失率低于 5%。由于对照个体数少于病例数，他们从对照组中随机选择 590 个对照个体，加上原来的对照个体获得平衡数据集。分析结果显示，SVM 方法即使在存在 5%基因分型错误，5%缺失数据或两种错误都存在的情况下也具有较好的探测基因–基因交互作用的能力，在分析 4 阶或 5 阶交互作用时，SVM 方法也展示较好的性能。

MDR、RMDR 和 GMDR 也在真实生物数据上有成功的应用，但是由于其计算复杂度较高，通常用于 SNP 个数不是很多的场合。乳腺癌(breast cancer)最常见的形式是散发性乳腺癌，其致病原理仍然不明，但是有临床证据显示雌激素会影响其发病率。Ritchie 等[36]将 MDR 应用于散发性乳腺癌的病例对照数据集，该数据集包含 200 个白人病例和对照个体的位于 COMT、CYP1A1、CYP1B1、GSTM1 和 GSTT1 基因上的 10 个 SNP 位点上的基因型数据，分析结果显示位于 3 个不同雌激素代谢基因 COMT、CYP1A1 和 CYP1B1 上的 4 个 SNP 位点之间存在高度交互作用，与散发性乳腺癌的犯病风险显著相关。膀胱癌(Bladder cancer)是泌尿系统中常见的恶性肿瘤，其发病机制十分复杂。Gui 等[45]利用 MDR 与 RMDR 对美国新罕布什尔州 355 例膀胱癌病例和 559 例对照个体的数据集进行研究。该数据包含了与 DNA 修复有关的 5 个基因上 7 个 SNP 位点的基因型。分析结果发现 MDR 与 RMDR 都能找到相同的最佳多位点交互作用模型，但 RMDR 标记为高风险或低风险的基因型组合数量比 MDR 少很多，使模型更易解释，RMDR 能比 MDR 提供了更加清晰的多位点交互作用模型。

Lou 等[46]利用 GMDR 和 MDR 对 191 名吸烟者和 191 名不吸烟者的脑源性神经营养因子(BDNF [MIM 113505])、Ⅱ型神经营养性酪氨酸激酶受体(NTRK2[MIM 600456])、胆碱能受体烟碱 α4(CHRNA4 [MIM 118504])和胆碱能受体烟碱 β2(CHRNB2 [MIM 118507])这 4 个基因的 23 个 SNP 位点基因型数据进行分析。分析结果发现了 CHRNA4 的 1 个 SNP (rs2-229959)和 NTRK2 的 3 个 SNP(rs993315，rs1122530

表 2 机器学习方法在真实遗传数据的应用

Table 2 Application of machine learning approaches to real genetic data

方法	应用案例	参考文献
GPNN	应用于帕金森病数据集，该数据集包含与线粒体复合体 I 相关基因的 70 个 SNPs，探测到了 DLST 基因与性别之间的交互作用	[17]
RF	应用于非综合征性唇腭裂(CL/P)的真实数据，发现了 WNT5B-MAFB 等有统计显著性的基因交互	[52]
SNPInterForest	应用于风湿关节炎的 GWAS 数据(约 500000 SNPs)，发现了两个新的交互作用	[27]
SVM	应用于前列腺癌研究中 18 个基因中的 57 个 SNP 位点，识别高达 5 个 SNP 之间的高阶交互作用	[33]
MDR	应用于与乳腺组织中雌激素代谢相关的 5 个基因中的 10 个 SNP 位点，确定了与乳腺癌风险相关的四位点交互作用	[36]
RMDR	测试了与 DNA 修复有关的 5 个基因中的 7 个 SNP 位点；结果与使用相同数据的 MDR 研究相同，但是提供了更清晰的高风险交互作用模型	[45]
GMDR	应用于 4 个基因中的 23 个 SNP 位点，以鉴定尼古丁依赖症的易感基因；GMDR 和 MDR 确定了相同的交互作用	[46]

和 rs736744)的交互作用与尼古丁依赖症有显著的统计相关性。GMDR 和 MDR 都能发现该 4 位点交互作用模型，但在模拟数据上的测试结果显示 GMDR 具有更好的预测能力。

表 2 汇总了上述机器学习方法在真实遗传数据上的应用及相关的结果。

3 结语与展望

在全基因组关联研究中，多种机器学习方法被用来探测基因–基因交互作用，这些方法在模拟数据中能够成功地发现基因–基因交互作用，有些方法也用来分析一些真实遗传数据并发现了一些相关的多基因交互作用(表 2)。机器学习算法在识别非线性复杂关系中具有优势,但机器学习算法也存在很多共性问题如计算资源需求大、可扩展性不强、给出的最优模型难以解释等局限性。探测基因-基因交互作用所需的计算量随着需要考虑的 SNP 位点数交互的维数指数增长，本文所讨论的大多数方法能从包含几百个 SNP 的数据集中探测多基因交互作用，但无法扩展到包含几十万 SNP 位点的数据集，当尝试发现大于 2 的高阶交互作用时，许多方法的效能显著降低。另外通过神经网络、随机森林、支持向量机等发现的基因交互作用模块很难给出合理的生物学解释。为了解决这些问题，可以考虑采用多阶段策略，在不同的阶段采用不同的机器学习方法，在前面的阶段采用神经网络、随机森林、支持向量机等寻找可能具有交互作用的候选 SNP 位点集，后续阶段则在这些 SNP 位点集的基础上，采用基于 MDR 的方法发现高阶基因交互作用，形成具有可扩展且结果容易解释的基因交互作用探测框架。

参考文献(References)：

[1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*, 2009, 461(7265): 747–753. [DOI]

[2] Pecanka J, Jonker MA, Bochdanovits Z, Van AW. A powerful and efficient two-stage method for detecting gene-to-gene interactions in GWAS. *Biostatistics*, 2017, 18(3): 477–494. [DOI]

[3] Li FG, Wang ZP, Hu G, Li H. Current status of SNPs interaction in genome-wide association syudy. *Hereditas (Beijing)*, 2011, 33(9): 901–910.
李放歌, 王志鹏, 户国, 李辉. 全基因组关联研究中的交互作用研究现状. 遗传, 2011, 33(9): 901–910. [DOI]

[4] Li J, Malley JD, Andrew AS, Karagas MR, Moore JH. Detecting gene-gene interactions using a permutation-based random forest method. *Biod Min*, 2016, 9(1): 14–31. [DOI]

[5] Young JH, Marcotte EM. Predictability of genetic interactions from functional gene modules. *G3*, 2017, 7(2): 617–624. [DOI]

[6] Wang XG, Lv C, Xu Q, Liu YF. Interactions among polymorphisms of NER genes prompt the risk of transplantation rejection. *Hereditas(Beijing)*, 2017, 39(1): 22–31.
王本刚, 吕执, 徐倩, 刘永峰. 多 NER 基因多态的交互作用与移植排斥的发病风险相关. 遗传, 2017, 39(1): 22–31. [DOI]

[7] Zhao JY, Zhu Y, Xiong MM. Genome-wide gene-gene interaction analysis for next-generation sequencing. *Eur J Hum Genet*, 2016, 24(3): 421–428. [DOI]

[8] Anusha AR, Vinodchandra SS. Probabilistic neural network inferences on oligonucleotide classification based on oligo: target interaction. In: Nguyen N, Tojo S, Nguyen L, eds. Intelligent Information and Database Systems. Cham: Springer, 2017: 733–740. [DOI]

[9] Li RW, Dudek SM, Kim D, Hall MA, Bradford Y, Peissig PL, Brilliant MH, Linneman JG, McCarty CA, Bao L, Ritchie MD. Identification of genetic interaction networks via an evolutionary algorithm evolved bayesian network. *BioData Min*, 2016, 9: 18. [DOI]

[10] Tong DL, Boocock DJ, Dhondalay GK, Lemetre C, Ball GR. Artificial neural network inference (ANNI): a study on gene-gene interaction for biomarkers in childhood sarcomas. *PLoS One*, 2014, 9(7): e102483. [DOI]

[11] De Poswar FO, Farias LC, De Fraga CA, Bambirra W Jr, Brito-Júnior M, Sousa-Neto MD, Santos SHS, De Paula AMB, D'Angelo MFSV, Guimarães AL. Interaction network analysis, and neural networks to characterize gene expression of radicular cyst and periapical granuloma. Journal of Endodontics. *J Endod*, 2015, 41(6): 877–883. [DOI]

[12] Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet Epidemiol*, 2008, 32(4): 325–340. [DOI]

[13] Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, Honda H. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. *BMC Bioinf*, 2004, 5: 120. [DOI]

[14] Leung FHF, Lam HK, Ling SH, Tam PKS. Tuning of the structure and parameters of a neural network using an improved genetic algorithm. *IEEE Trans Neural Netw*, 2003, 14(1): 79–88. [DOI]

[15] Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics*, 2003, 4(1): 28–42. [DOI]

[16] Manshad AK, Manshad MK, Ashoori S. The application of an artificial neural network (ANN) and a genetic programming neural network (GPNN) for the modeling of experimental data of slim tube permeability reduction by asphaltene precipitation in Iranian crude oil reservoirs. *Petroleum Science and Technology*, 2012, 30(23): 2450–2459. [DOI]

[17] Motsinger AA, Lee SL, Mellick G, Ritchie MD. GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinf*, 2006, 7: 39. [DOI]

[18] De Campos LML, De Oliveira RCL, Roisenberg M. Optimization of neural networks through grammatical evolution and a genetic algorithm. *Expert Systems with Applications*, 2016, 56: 368–384. [DOI]

[19] Breiman L. Random forests. *Machine Learning*, 2001, 45(1): 5–32. [DOI]

[20] Yoo W, Ference BA, Cote ML, Schwartz A. A comparison of logistic regression, logic regression, classification tree, and random forests to identify effective gene-gene and gene-environmental interactions. *Int J Appl Sci Technol*, 2012, 2(7): 268. [DOI]

[21] Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol*, 2013, 66(4): 398–407. [DOI]

[22] Nguyen TT, Huang JZ, Wu Q, Nguyen T, Li M. Genome-wide association data classification and SNPs selection using two-stage quality-based random forests. *BMC Genomics*, 2015, 16 (Suppl.2): S5. [DOI]

[23] Chen J, Zhou Y, Gao YQ, Cao WJ, Sun H, Liu YF, Wang C. A genetic features and gene interaction study for identifying the genes that cause hereditary spherocytosis. *Hematology*, 2017, 22(4): 240–247. [DOI]

[24] Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol*, 2005, 28(2): 171–182. [DOI]

[25] Chen X, Ishwaran H. Pathway hunting by random survival forests. *Bioinformatics*, 2013, 29(1): 99–105. [DOI]

[26] Winham SJ, Colby CL, Freimuth RR, Wang X, De Andrade M, Huebner M, Biernacka JM. SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinf*, 2012, 13: 164. [DOI]

[27] Yoshida M, Koike A. SNPInterForest: a new method for detecting epistatic interactions. *BMC Bioinf*, 2011, 12(1): 469–479. [DOI]

[28] Pan QX, Hu T, Malley JD, Andrew AS, Karagas MR, Moore JH. Supervising random forest using attribute interaction networks. In: Vanneschi L, Bush WS, Giacobini M, eds. Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. EvoBIO 2013. Lecture

Notes in Computer Science. Berlin, Heidelberg: Springer, 104–116. [DOI]

[29] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, 20(3): 273–297. [DOI]

[30] Sehhati M R, Dehnavi A M, Rabbani H, Javanmard SH. Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence. *J Med Signals Sens*, 2013, 3(2): 87–93. [DOI]

[31] Qi ZQ, Tian YJ, Shi Y. Robust twin support vector machine for pattern classification. *Patt Recognit*, 2013, 46(1): 305–316. [DOI]

[32] Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, Mackey J, Wishart D, Greiner R, Zanke B. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clin Cancer Res*, 2004, 10(8): 2725–2737. [DOI]

[33] Chen SH, Sun JL, Dimitrov L, Turner AR, Adams TS, Meyers DA, Chang BL, Zheng SL, Grönberg H, Xu JF, Hsu FC. A support vector machine approach for detecting gene-gene interaction. *Genet Epidemiol*, 2008, 32(2): 152–167. [DOI]

[34] Shen YY, Liu Z, Ott J. Support vector machines with L_1 penalty for detecting gene-gene interactions. *Int J Data Min Bioinform*, 2012, 6(5): 463–470. [DOI]

[35] Ban HJ, Heo JY, Oh KS, Park KJ. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. *BMC Genet*, 2010, 11: 26. [DOI]

[36] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 2001, 69(1): 138–147. [DOI]

[37] Lee S, Son D, Yu WB, Park T. Gene-gene interaction analysis for the accelerated failure time model using a unified model-based multifactor dimensionality reduction method. *Genom Inform*, 2016, 14(4): 166–172. [DOI]

[38] Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 2007, 9(1): 30–50. [DOI]

[39] Gui J, Moore JH, Williams SM, Andrews P, Hillege HL, Van Der Harst P, Navis G, Van Gilst WH, Asselbergs FW, Gilbert-Diamond D. A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS One*, 2013, 8(6): e66545. [DOI]

[40] Tsai CT, Lai LP, Lin JL, Chiang FT, Hwang JJ, Ritchie MD, Moore JH, Hsu KL, Tseng CD, Liau CS, Tseng YZ. Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation*, 2004, 13(6): 1640–1646. [DOI]

[41] Su MW, Tung KY, Liang PH, Tsai CH, Kuo NW, Lee YL. Gene-gene and gene-environmental interactions of childhood asthma: a multifactor dimension reduction approach. *PLoS One*, 2012, 7(2): e30694. [DOI]

[42] He H, Oetting WS, Brott MJ, Basu S. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. *BMC Med Genet*, 2009, 10(1): 127–144. [DOI]

[43] Yu W, Lee S, Park T. A unified model based multifactor dimensionality reduction framework for detecting gene–gene interactions. *Bioinformatics*, 2016, 32(17): i605–i610. [DOI]

[44] Leem S, Park T. An empirical fuzzy multifactor dimensionality reduction method for detecting gene-gene interactions. *BMC Genom*, 2017, 18(S2): 115–127. [DOI]

[45] Gui J, Andrew AS, Andrews P, Nelson HM, Kelsey KT, Karagas MR, Moore JH. A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann Hum Genet*, 2011, 75(1): 20–28. [DOI]

[46] Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *AJHG*, 2007, 80(6): 1125–1137. [DOI]

[47] Chen GB, Liu NJ, Klimentidis YC, Zhu XF, Zhi DG, Wang XJ, Lou XY. A unified GMDR method for detecting gene-gene interactions in family and unrelated samples with application to nicotine dependence. *Hum Genet*, 2014, 133(2): 139–150. [DOI]

[48] Kwon MS, Kim K, Lee S, Chung W, Yi SG, Namkung J, Park T. GWAS-GMDR: a program package for genome-wide scan of gene-gene interactions with covariate adjustment based on multifactor dimensionality reduction. IEEE International Conference on Bioinformatics and Biomedicine Workshops. Washington, DC: IEEE, 2011: 703–707. [DOI]

[49] Wang XS, Cheng YH, Zhang L. Machine learning method in bioinformatics. Beijing: Science Press, 2014.
王雪松, 程玉虎, 张林. 生物信息学中的机器学习分析方法. 北京: 科学出版社, 2014. [DOI]

[50] Li SY, Cui YH. Gene-centric gene-gene interaction: a model-based kernel machine method. *Annals of Applied Statistics*, 2012, 6(3): 1134–1161. [DOI]

[51] Mellick GD, Silburn PA, Prince JA, Brookes AJ. A novel screen for nuclear mitochondrial gene associations with Parkinson's disease. *J Neural Transm*, 2004, 111(2): 191–199. [DOI]

[52] Li Q, Kim Y, Suktitipat B, Hetmanski JB, Marazita ML, Duggal P, Beaty TH, Bailey-Wilson JE. Gene-gene interaction among *WNT* genes for oral cleft in trios. *Genet Epidemiol*, 2015, 39(5): 385–394. [DOI]

(责任编委: 赵方庆)