文章编号：2095-2163(2020)01-0257-05 中图分类号：TP391 文献标志码：A

基于高通量测序数据的插入/删除新突变检测方法

邢文昊，刘永壮，王亚东
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 插入/删除新突变是一种重要的新突变形式，与多种人类疾病的发生密切相关。随着高通量测序技术的迅猛发展，基于高通量测序数据进行插入/删除新突变检测已成为常规手段，但由于测序错误以及 reads 比对错误的影响，已有的检测方法通常存在错误率较高的问题。本文提出一种基于 Adaboost 的插入/删除新突变检测方法，旨在对常用的新突变检测方法产生的插入/删除新突变检测结果进行过滤，在确保基本不损失敏感度的前提下，显著降低错误发现率。

关键词： 高通量测序数据；新突变检测；Adaboost

De Novo Indel detection method based on high-throughput sequencing data

XING Wenhao，LIU Yongzhuang，WANG Yadong
（School of Computer Science and Technology，Harbin Institute of Technology，Harbin 150001，China）

【Abstract】 De Novo Indel is an important form of de novo mutation，and it is closely related to the occurrence of a variety of human diseases. With the development of high-throughput sequencing technology，using high-throughput data to detect De Novo Indels has become routine. However，due to the sequencing errors and the reads alighment errors，existing detection methods usually yield high error. This paper proposes a new De Novo Indel detection method based on Adaboost. This method is designed to filter De Novo Indels detected by common De Novo Indel detection methods，which can significantly reduce the false discovery rate without sacrificing the sensitivity.

【Key words】 high-throughput sequencing data；De Novo mutation detection；Adaboost

0 引 言

新突变是指并非遗传自亲代，而是在子代中首次出现的基因组变异，这种变异既可能来自于生殖细胞，也可能来源于胚胎发育早期的受精卵。大量研究表明新突变与人类疾病的发生密切相关，尤其是自闭症、癫痫等精神类疾病[1]。随着高通量测序技术的迅猛发展，基于高通量测序数据进行新突变检测已经成为常规手段。目前的基于高通量测序数据的新突变检测方法大概可以分为 2 类。一类是在亲代和子代样本中独立地进行基因组变异检测，然后通过比较子代和亲代的基因型发现新突变；另一类是对亲代和子代基因组数据建立联合检测模型直接检测新突变[2]。例如 PolyMutt [3] 和 Triodenovo [4] 均为基于似然度的新突变检测方法，该方法分别计算无孟德尔遗传约束和有孟德尔遗传约束下的最大基因型似然度，然后根据似然比确定候选位点是否是新突变，似然比越大，候选位点是新突变的可能性越高；DeNovoGear [5] 是一种基于贝叶斯的新突变检测方法，该方法利用贝叶斯公式计算候选位点是新突变的后验概率，并根据后验概率的大小来判断候选位点是否是新突变；DNMFilter[2] 是一种基于 Gradient Boosting [6] 的新突变检测方法，该方法利用 Gradient Boosting 作为分类器，从子代和亲代序列比对上下文选择序列特征，该方法相比其它方法能够显著降低新突变的错误发现率。在上述 4 种方法中，PolyMutt、Tridenovo 均能够检测单核苷酸新突变和插入/删除新突变，但插入/删除新突变的检测准确率明显低于单核苷酸新突变的检测准确率；而 DNMFilter 只能够检测单核苷酸新突变而不能检测插入/删除新突变。

本文提出一种基于 Adaboost 的插入/删除新突变检测方法。该方法利用 Adaboost 作为分类模型；该方法对 read 比对之后的家系基因组数据进行局部重新拼接，并从局部拼接后的 read 比对上下文中提取序列特征；该方法利用已经验证的插入/删除新突变和随机选取的错误插入/删除新突变构建训练集。最后本文将训练的 Adaboost 模型应用于千人基因组计划 CEU 家系基因组测序数据，并同常用的新突变检测方法进行比较。

作者简介：邢文昊(1992-)，男，硕士研究生，主要研究方向：生物信息学；刘永壮(1985-)，男，博士研究生，主要研究方向：生物信息学；王亚东(1964-)，男，教授，博士生导师，主要研究方向：机器学习、知识工程、生物信息学等。

收稿日期：2017-06-15

1 Adaboost 算法

Adaboost 算法[7]是一种经典的 boosting 算法，其主要思想是将一系列较弱的分类器结合起来，形成具有较强分类能力的组合分类器(强分类器)。

假设 N 个被标记的样本数据集合$\{(X_1,Y_1),(X_2,Y_2),(X_3,Y_3),\cdots,(X_N,Y_N)\}$为训练集，$X_i$ 为特征向量，Y_i 为类别标记，定义 $W_k(i)$ 为第 K 次循环样本 i 的权重(K 的取值范围为 1,2,3,$\cdots$,$K_{\max}$，其中 $K_{\max}$ 为最大循环次数)。Adaboost 算法流程可表述如下。

首先，初始化 $W_1(i)= 1/N$(其中 $i=1,2,3,\cdots,N$)。

然后，在每一轮循环中根据样本权重 $W_k(i)$，在训练集中进行样本重采样生成子集 D_k。

最后，在子集 D_k 上训练弱分类器 C_k，计算 C_k 在训练集上的训练误差 E_k，并且根据本轮的分类结果计算下一轮训练的权重 $W_{k+1}(i)$，研究推得其数学运算公式如下：

$$W_{k+1}(i) \leftarrow \frac{W_k(i)}{Z_k} * \begin{cases} e^{-\alpha_k}, & \text{if } h_k(x_i) = y_i; \\ e^{\alpha_k}, & \text{if } h_k(x_i) \neq y_i. \end{cases} \quad (1)$$

其中，Z_k 是归一化系数；$h_k(x_i)$ 是分类器 C_k 对特征向量 X_i 分类的标记；α_k 的数学求值则可写作如下形式：

$$\alpha_k \leftarrow \frac{1}{2}\ln\frac{1-E_k}{E_k}, \quad (2)$$

使用更新后的权重 $W_{k+1}(i)$ 重复上述过程，当循环次数达到最大循环次数 $K_{\max}$ 或者是误差率小于某一阈值时停止，样本被划分类别可使用以下公式做出判断：

$$H(x) = sign\left(\sum_{k=1}^{k_{\max}} \alpha_k h_k(x)\right). \quad (3)$$

Adaboost 算法可以有效降低分类中的偏差和方差[8]，而与其它 boosting 算法相比，Adaboost 对过拟合有着更强的抗性[9]。选择 Adaboost 算法可以有效降低假阳性错误，进而减少过滤结果中的假阳性位点的数量。

2 基于 Adaboost 的插入/删除新突变过滤方法

基于 Adaboost 的插入/删除新突变检测方法总体流程如图 1 所示，使用千人基因组计划的家系基因组数据作为实验数据。首先使用已有的新突变检测工具检测出候选的插入/删除新突变位点，并对候选位点进行局部从头拼接(local de novo assembly)；然后结合已经存在的真实突变位点，创建带标签的候选位点集合，根据候选位点从重组的家系基因组数据中提取特征构建训练集；最后使用训练集对模型进行调整优化，并且用测试集来验证模型的效果。

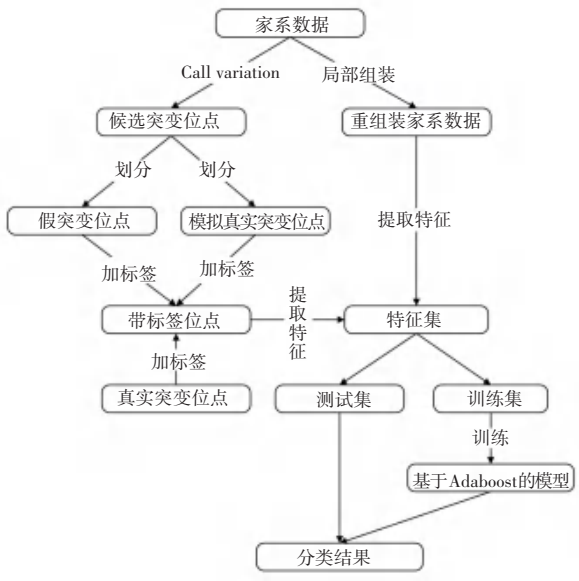

图 1 插入/删除新突变检测流程图

Fig. 1 The work flow to detect De Novo Indels

2.1 数据集

本文使用千人基因组计划 CEU 家系全基因组测序数据进行模型构建与实验结果分析，该数据集的 BAM 文件可以从此处：ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/下载得到。此外，该数据集包含了经过生物实验验证的 53 个插入/删除新突变(6 个插入突变，47 个删除突变)[5]，这些突变位点的位置信息可以在 http://www.nature.com/nmeth/journal/v10/n10/extref/nmeth.2611-S2.xls 获得。

2.2 训练集构建

本文构造训练集应该满足的条件是真实的新突变尽可能集中分布，假的新突变在空间分布上应该与新突变尽可能分离，并且在空间中尽可能分布广泛。

本文构造正例集的具体方法是：首先将 2.1 节中经生物实验验证的 53 个真实新突变加入到正例集；由于正例集规模较小，无法满足模型构建的需求，本文模拟部分插入/删除新突变并将其加入到正例集中，此后接续的模拟过程就是选择父本(或母本)和子代是野生型且母本(或父本)是杂合型的位点，将母本(或父本)与子代进行交换，即可模拟出满足亲代是野生型、子代是杂合型的新突变位点。

本文构造反例集的具体方法是：从除去经过生物实验验证的 53 个真实新突变的候选新突变位点集合中，随机选取一定量的位点加入到反例集；从胚系基因组变异(Germline Variation)位点集合，随机选取一定量的位点加入到反例集中。

2.3　序列特征选择

研究中选定了 55 个特征来训练模型，这些特征分为 4 个部分。就整体而言，第一部分是父本 read 的特征，第二部分是母本 read 的特征，第三部分是子代 read 特征，第四部分是公共特征。其中，前三部分特征都是相同的，所选用的特征见表 1。

表 1　插入/删除新突变检测选择的序列特征描述信息

Tab. 1　Description of selected sequence features to detect De Novo Indels

特征	说明
等位基因平衡性(Allele Balance)	3 个值，变异碱基比例
read 深度(Read Depth)	3 个值，候选位点 read 的数量
参考序列与变异序列的平均映射质量 (Mean Mapping Quality for ALT and REF)	6 个值，平均映射质量
参考序列与变异序列到 3’端的平均距离 (Mean Distance to Three Prime for ALT and REF)	6 个值，候选位点到 3’端的距离
链方向 (Strand Direction for ALT and REF)	6 个值，read 是否是同向
Soft Clipped Read 的比例 (Fraction of Soft Clipped Reads for ALT and REF)	6 个值，Soft Clipped Read 所占的比例
平均错配位置距离 (Mean Nearby Mismatches for ALT and REF)	6 个值，错配位置到候选位置距离
平均插入删除位置距离 (Mead Nearby Indels for ALT and REF)	6 个值，插入删除位置到候选位置距离
链偏性 (Strand Bias)	3 个值，read 方向的 Fisher 检验
P 值 (PValue)	2 个值(父本或母本与子代基因数目的 Fisher 检验)
MQ0 read 的比例 (Fraction of MQ0 Reads for ALT and REF)	6 个值，映射质量为 0 的 read 的比例
参考序列中是否有同聚物 (Homopolymer)	1 个值，是否含有同聚物
参考序列中是否含有短序列重复 (Short and Tandem Repeat)	1 个值，是否含有短序列重复

3　实验结果与分析

3.1　Adaboost 与 Gradient boosting 性能比较

Gradient boosting 和 Adaboost 一样，也是一种常用的 boosting 方法，两者的主要区别是损失函数的不同，直观上看前者是由梯度下降来决定，后者是由高权重的点来决定。Adaboost 可以看作是 Gradient boosting 的特殊情况[10]。

为了测试比较两者的运行效果，在同一组数据上分别使用 Gradient boosting 和 Adaboost 进行了实验，如图 2 所示，纵轴表示百分比，横轴表示 *cutoff*(*cutoff* 是分类时使用的值，小于此值被分为假的插入 / 删除突变，大于此值被划分为真的插入 / 删除突变，实验中需要对其进行设置)，*FP* 表示 *cutoff* - 假阳性率曲线，*TP* 表示 *cutoff* - 真阳性率曲线。

在图 2 所示的曲线中，GBFP 在 0~0.5 下降比较快，ADATP 在 0.5 以后下降比较快，但是所设定的 *cutoff* 都大于 0.5(只有 2 个类别：1 个是真的插入/删除突变，1 个是假的插入/删除突变)，在 0.5 以后的区域 Adaboost 的表现比较好，可以筛选掉大部分的假阳性的插入/删除新突变。在 *cutoff* - *TP* 曲线上，尽管 GBTP 表现比较好，在 0.97 左右才出现了突然下降，即大量的真实插入/删除新突变开始被错分为假的插入删除突变，而 ADAFP 在 0.84 左右就已经开始下降，但是考虑到设置 *cutoff* 时，为了防止原本数量就很少的真实的插入/删除新突变被错判为假的插入删除突变，所以这个值不会过高，尽管 Adaboost 在 *cutoff* - *TP* 曲线上下降较早，但是和

Gradient boosting 一样满足精度要求,可以选择设置在 0.6~0.8 之间。

3.2 基于 Adaboost 的插入/删除新突变检测模型评估

为了考察这些特征对于分类的重要性,本文选择部分正例和反例,使用 R 语言中的 adabag 包对其进行训练,然后提取各个特征对于分类的相对重要性,设计结果如图 3 所示,在本组数据中,父本等位基因平衡性、父本和子代 PValue 值、母本等位基因平衡性、Homopolymer、父本平均 mapping 质量等因素对于分类的影响比较大,而子代链方向等因素影响较小。

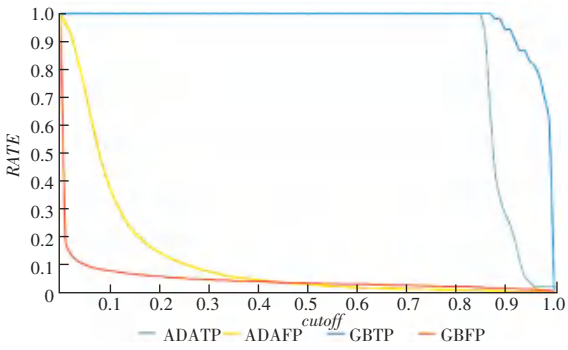

图 2 基于 Adaboost 与基于 Gradient boosting 构建模型对比

Fig. 2 Comparison of model based on Adaboost and Gradient boosting

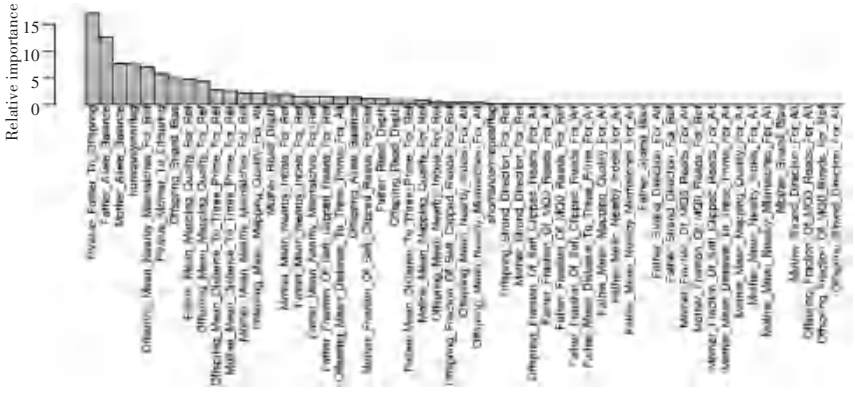

图 3 用于插入/删除新突变检测的各特征相对重要性

Fig. 3 Relative importance of all features for De Novo Indel detection

为了评价构造的训练集的分类性能,对所选择的训练集进行主成分分析,最终绘制后则如图 4 所示,前三个主成分之和达到了 75% 以上。其中,红色的点代表了正例集的数据,蓝色代表反例集的数据。可以看出正例样本与反例样本能够明显区分开,表明了构建的训练集和选择的分类特征能够较好区分真实的插入/删除新突变和假的插入/删除新突变。

3.3 千人基因组计划家系数据的实验结果与分析

将本模型与其它常用的模型进行了对比,研究得到的对比后结果详见表 2。其中,Triodenovo 是基于似然模型对新突变检测的工具,相较于 Polymutt 的运行,性能要更胜一筹。DeNovoGear 是基于似然模型对新突变进行检测的工具,GATK PhaseByTransmission 用于矫正家系中的模糊的位点。使用 4 种方法对同一个家系基因组数据进行处理,各个方法使用默认的参数,检出总数目中本文设计实现的 DNINDELFilter 具有最小的数目,而且 DNINDELFilter 检测出了数据中所有的 53 个真实的插入/删除新突变。从而达到了保留所有的真实插入/删除新突变,保证真阳性率,同时尽可能删除较多的假的插入/删除新突变,降低假阳性率的预期目标。

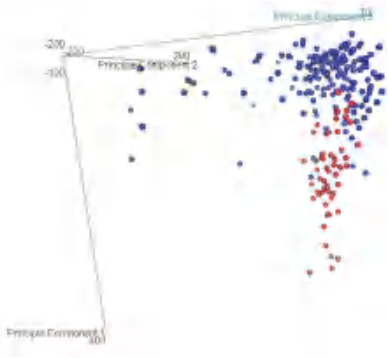

图 4 训练集数据主成分分析

Fig. 4 PCA for training data

表 2 新突变检测方法比较

Tab. 2 Comparison of De Novo mutation detection methods

方法	检测出的总变异数目	真实变异被检出数目
Triodenovo	990	52
DeNovoGear	389	47
PhaseByTransmission	1 054	53
DNINDELFilter	149	53

4　结束语

本文将 Adaboost 算法应用在插入/删除新突变筛选中，并且应用 Adaboost 算法，实现了 DNINDELFilter。在此研究过程中，选择了与插入/删除新突变相关的 55 个特征作为训练特征，对高通量测序数据进行了局部重新组装，通过构建训练集、继而训练、测试该模型、及对模型参数辅以一定调整，最后使得模型能够在保留大部分的真实插入/删除新突变同时，筛选掉绝大多数的假的突变，并且本文提出的模型具有一定的抗过拟合能力，筛选结果表现出较高的可信度。在本文的研究基础之上，未来的工作可以在如下方面展开：应深入探讨、并进一步优化模型中的特征，使得模型对插入/删除新突变具有更好的筛选能力；目前该模型在插入/删除新突变检测中取得了良好的效果，后续可以尝试将本文的模型做出研究改进，并应用到结构新突变筛选中。

参考文献

[1] VELTMAN J A, BRUNNER H G. De novo mutations in human genetic disease[J]. Nature Reviews Genetic, 2012, 13(8): 565-575.

[2] LIU Yongzhuang, LI Bingshan, TAN Renjie, et al. A gradient-boosting approach for filtering de novo mutations in parent-offspring trios[J]. Bioinformatics, 2014, 30(13): 1830-1836.

[3] LI Bingshan, CHEN Wei, ZHAN Xiaowei, et al. A likelihood-based framework for variant calling and de novo mutation detection in families[J]. PLoS Genet, 2012, 8(10): e1002944.

[4] WEI Qiang, ZHAN Xiaowei, ZHONG Xue, et al. A Bayesian framework for de novo mutation calling in parents-offspring trios [J]. Bioinformatics, 2015, 31(9): 1375-1381.

[5] RAMU A, NOORDAM M J, SCHWARTZ R S, et al. DeNovoGear: de novo indel and point mutation discovery and phasing[J]. Nature methods, 2013,10(10): 985-987.

[6] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[7] FREUND Y, SCHAPIRE R E. A desicion-theoretic generalization of on - line learning and an application to boosting [M]// VITÁNYI P. Computational learning theory. EuroCOLT 1995. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). Berlin/ Heidelberg:Springer,1995,904:23-37.

[8] BREIMAN L. Bias, variance, and arcing classifiers[R].Berkeley: University of California, 1996.

[9] MASON L, BAXTER J, BARTLETT P, et al. Boosting algorithms as gradient descent[C]//NIPS´99 Proceedings of the 12th International Conference on Neural Information Processing Systems. Denver, CO :ACM,1999:512-518.

[10] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. Annals of statistics, 2001,29(5):1189-1232.

(上接第 256 页)

来实现业财融合的优化方法。大数据和云计算的结合使得部门间上传的海量数据经过分析、筛选和存储后保证了数据的有效性和相关性；云计算和物联网的结合构造“云物联”平台使得数据的录入效率提高，数据传递的时效性大大提高，保证了管理层实时动态地获取企业经营状况和财务状况；移动互联网和人工智能的结合打破了传统的空间阻碍，使得管理层能够在移动端随时获取企业生产和交易信息以及财务报告，人工智能的深度学习又使得海量数据筛选的效率大大提高并且信息的相关性也提高了。“大智移云”技术的使用对于解决当下企业业财融合模式的困境有着积极意义，企业需要结合自身特点选择合适的技术来优化和落实企业的业财融合模式，提高企业竞争力，实现企业价值增值。

参考文献

[1] 范志英. “大智移云”背景下财务共享平台构建及应用—以 TCL 集团股份有限公司为例[J]. 财会通讯,2019(4):111-115.

[2] 郭永清. 中国企业业财融合问题研究[J]. 会计之友,2017(15): 47-55.

[3] 赵健,邱铁. 未来已来,你准备好了吗？—“大智移云”技术企业应用情况调查报告[J]. 财会通讯,2018(28):3-7.

[4] 朱亮. 基于财务共享平台的企业业财融合模式研究—以 W 保险公司为例[J]. 财会通讯,2018(35):79-82.

[5] 张玉缺. 云计算下的企业业财融合运作模式研究—以国家电网为例[J]. 会计之友,2018(24):58-60.

[6] 程平,施先旺,万章浩. 基于业财一体化的采购活动大会计研究[J]. 财会月刊,2017(34):3-10.

[7] 马贵兰. 基于大数据思维的“业财融合”管理会计体系应用—以通信行业为例[J]. 财会月刊,2015(32):24-26.

[8] 陆兴凤. 基于业财融合的新型财务信息化系统构建思考—以新零售为例[J]. 财会月刊,2018(9):98-102.

[9] 李闻一,李栗,曹菁,等. 论智慧财务的概念框架和未来应用场景[J]. 财会月刊,2018(5):40-43.

[10] 李闻一,王嘉良,陈桢. 基于“业财融合”的一体化管控——中石油湖北销售公司案例[J]. 财会月刊,2015(28):11-15.

[11] 王学(王乐),于璐. 基于财务职能定位的业财融合措施分析[J]. 会计之友,2016(22):34-36.

[12] 刘丽丽,毛庆. 财务业务一体化视角下企业财务流程优化[J]. 财会通讯,2016(17):41-43.

[13] 王建林. 财务业务一体化中小企业管理研究[J]. 财会通讯,2013(10):97-99.

[14] 周元元,贾晓柏. 人工智能在财务业务一体化中的应用[J]. 财会月刊,2007(31):60-62.

[15] 王斌. 论业财融合[J]. 财务研究,2018(3):3-9.

[16] 杜勇,李光辉. 京能集团业财一体化整合路径与机制[J]. 财务与会计,2017(13):14-16.

[17] 汤谷良,夏怡斐. 企业“业财融合”的理论框架与实操要领[J]. 财务研究,2018(2):3-9.