基于机器学习的microRNA预测方法研究进展

王 颖[1,2] 李 金[1] 王 磊[1] 徐成振[1] 才忠喜[1]

(哈尔滨工程大学自动化学院 哈尔滨 150001)[1] (齐齐哈尔大学网络信息中心 齐齐哈尔 161006)[2]

摘 要 传统的克隆方法受组织和环境影响显著,且实验成本高,而计算方法中的比较方法对进化距离远的microRNA敏感性低,无法预测无同源的microRNA,机器学习方法解决了比较方法依赖同源基因的问题。首先总结了基于机器学习预测microRNA的相关生物学知识;其次,给出基于机器学习的microRNA预测方法的大体流程,列举了基于机器学习的microRNA预测方法的最新研究算法及软件;再次,从数据集选取、特征集选取、分类器设计、特征子集选择、类不平衡问题解决和评价标准等环节出发,归纳总结了各环节中采用的方法及技术,并详细阐述了它们的最新研究进展,部分环节对采用的方法及技术进行了对比分析,总结了各自的优势和不足;最后,总结和展望了基于机器学习的microRNA预测方法的研究工作。

关键词 microRNA,机器学习,分类器,特征选取,类不平衡,生物信息学

中图法分类号 Q3,Q6,TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.002

Research and Progress of microRNA Prediction Methods Based on Machine Learning

WANG Ying[1,2] LI Jin[1] WANG Lei[1] XU Cheng-zhen[1] CAI Zhong-xi[1]

(College of Automation, Harbin Engineering University, Harbin 150001, China)[1]

(Network Information Center, Qiqihar University, Qiqihar 161006, China)[2]

Abstract Traditional cloning experimental approaches are affected by the organizational and environmental impact, and the cost is high. The comparative method that belongs to the computational method is low sensitivity to the far evolutionary distance genes, and can't predict the no homologous microRNAs. The machine learning method can resolve the restraints that comparative method dependents on the homologous gene. Firstly, this paper summarized the microRNA relevant biological knowledge which the machine learning is related to. Secondly, it outlined the general process and the latest research software and algorithms of machine learning based on microRNA prediction. Thirdly, starting from data selection, feature selection, classifier design, feature subset selection, class imbalance problem, performance evaluation and other aspects in terms of the essential elements of microRNA prediction based on the machine learning, it summarized the method and technology in each process, described their latest research progress. The approaches were contrasted and analyzed respectively in some process, and their respective advantages, disadvantages were summarized. Finally, summary and prospect of the research work on microRNA prediction based on machine learning were given.

Keywords microRNA, Machine learning, Classifier, Feature extraction, Class imbalance, Bioinformatics

1 引言

microRNA是一类内源性、长度为19～24个碱基的小分子RNA,microRNA在转录后水平的基因表达中起到调控作用,在生物学和疾病学进程中扮演着重要角色[1,2],包括细胞发育、组织分化、细胞循环[3-6]。microRNA的调节异常与一些疾病有密切关系,尤其是超过50%人类的microRNA与癌症基因片段相关。要进一步研究microRNA对生命体的调控功能,首要的是对microRNA进行有效的预测,microRNA的预测成为生物信息学研究领域的一个重要课题。

到目前为止,miRbase数据库中包括microRNA发夹前体序列24521条、成熟microRNA30424条,microRNA序列共涵盖206个物种[7],但相对于尚未被发现的microRNA的数量,这些只是很少的一部分。microRNA识别采用的是传统的cDNA克隆测序方法,由于microRNAs为短序列、表达水平低、表达水平受组织和环境条件变化显著、实验成本高等原因,识别microRNAs基因变得十分困难,而microRNA前体具有稳定的茎环结构、物种之间具有很强的保守性、最小折

到稿日期:2014-07-01 返修日期:2014-09-08 本文受国家重大仪器专项(2012YQ0401401001),黑龙江省教育厅科学技术研究项目(12541898)资助。

王 颖(1980－),女,博士生,工程师,主要研究领域为模式识别与智能系统、生物信息学,E-mail:wangying0129@126.com;李 金(1962－),女,教授,博士生导师,主要研究领域为高维数据场可视化、生物信息学、模式识别,E-mail:lijin@hrbeu.edu.cn(通信作者);王 磊(1983－),男,博士,讲师,主要研究领域为生物信息学、高维数据场可视化;徐成振(1988－),男,博士生,主要研究领域为生物信息学、机器学习、模式识别;才忠喜(1973－),男,博士生,副研究员,主要研究领域为生物信息学、机器学习、模式识别。

叠自由能，因此采用计算方法预测 microRNA 成为当前 micro-RNA研究领域的一个研究课题。基于比较方法是研究早期采用的一种计算方法，它根据二级结构折叠成发夹结构和相近基因保守特征预测基因序列。比较方法在 microRNA 挖掘领域取得巨大成功，能够挖掘出基因组范围内相近物种保守性好的 microRNA，但是这种方法找不到那些无同源的 micro-RNA，且对于进化距离远的 microRNA 敏感性低，尤其是对于快速进化和特异性物种无能为力，因此采用机器学习方法预测 microRNA 应运而生。

本文总结了基于机器学习预测 MicroRNA 的相关生物学知识，给出了基于机器学习的 microRNA 预测方法的大体流程，列举了基于机器学习的 microRNA 预测方法最新研究算法及软件，从数据集选取、特征集选取、分类器设计、特征子集选择、类不平衡问题解决和评价标准等环节出发，归纳总结了各环节中采用的方法及技术，并详细阐述了它们的最新研究进展，部分环节针对采用的方法及技术进行对比分析，总结了各自的优势、不足，最后，总结和展望了基于机器学习的 microRNA 预测方法的研究工作。

2 基于机器学习的 microRNA 预测方法概述

采用机器学习方法预测 microRNA 解决了比较方法对同源基因的依赖，只是采用它们提取数据集和特征集，训练分类器，采用机器学习方法预测 microRNA 的大体流程如图 1 所示。

采用机器学习方法预测 microRNA 大致分为数据集选取、特征集选取、分类器设计、特征子集选择、类不平衡问题解决和评价标准等几环节，首先研究正样本集、负样本集和测试集；然后研究能代表 microRNA 的生物特征，确定特征集；再根据特征集数量或分类器性能要求以及生物特征分析，确定是否进行特征选择，如果需要，进行特征选择，则采用特征选择算法选取最优特征子集；然后针对分类数据特点，决定是否对类不平衡问题进行解决，如果需要，则采用类不平衡问题算法选取平衡训练子集；然后，根据数据特点，选用分类器；接着将提取的特征集和训练集在分类器反复训练，以获得性能最优的分类器；然后采用评估标准对算法在精度、特异性、敏感性等方面进行性能评估。分类器应用时，输入候选 microRNA或 pre-miRAN 序列即可。

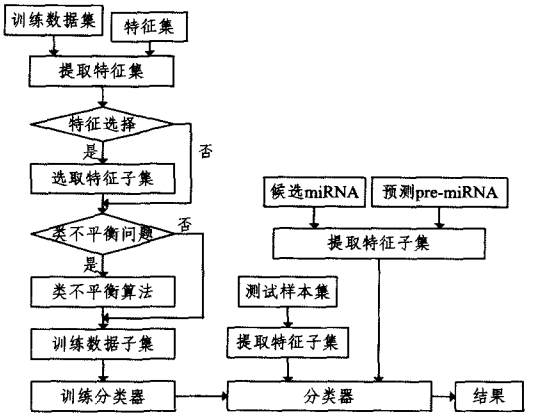

图 1

根据预测对象不同，机器学习方法对 microRNA 预测又可分为 pre-microRNA 预测、成熟 microRNA 预测等。采用机器学习的 microRNA 预测方法多数为对 pre-microRNA 的预测，主要是由于 microRNA 序列较短，而且是由～120nt 的 pre-microRNA 的分裂得来，pre-microRNA 具有保守的发夹结构等特征，而且是成熟 microRNA 的中间产物和必经阶段。根据算法依据不同，对成熟 microRNA 预测又分为基于 Drosha 和 Dicer 位点特征预测和基于 microRNA 茎环特征预测。如 Microprocessor SVM [20]、MiRmat[21] 和 miRRim2[22] 基于 Drosha 和 Dicer 位点特征预测成熟 microRNA，maturepred[23]、maturebayes[24]、mirdup[25] 基于 microRNA 茎环特征预测成熟 microRNA。

基于机器学习的 microRNA 预测算法及软件如表 1 所列。

表 1 基于机器学习的 microRNA 预测算法及软件

软件名	预测对象	分类器	网址	敏感性/特异性/精确性(ACC)
Triplet[26](2005)	pre-microRNA	SVM	http://bioinfo.au.tsinghua.edu.cn/software/mirnasvm/Triplet-svm-predictor.htm	93.30%/88.10%
Mipred[27](2007)	pre-microRNA	RF	http://www.bioinf.seu.edu.cn/miRNA/	89.35%/93.21%
mipred[28](2007)	pre-microRNA	SVM	http://web.bii.a-star.edu.sg/～stanley/Publications	84.55%/97.97%
micropred[29](2009)	pre-microRNA	SVM	http://web.comlab.ox.ac.uk/people/ManoharaRukshan.Batuwita/microPred.htm	90.02%/97.28%
miridenf[30](2012)	pre-microRNA	SVM	http://www.regulatoryrna.org/pub/miri-dent/index.html	99.2%/97.6%
PlantMiRNAPred[31](2011)	pre-microRNA	SVM	http://nclab.hit.edu.cn/PlantMiRNAPred/	ACC>90%
HeteroMirPred[32](2013)	pre-microRNA	SVM、RF、KNN	http://ncrna-pred.com/premiRNA.html.	94.8%/98.3%
MatureBayes[24](2010)	成熟 microRNA	贝叶斯	http://miRNA.imbb.forth.gr/MatureBayes.html	
MaturePred[23](2011)	成熟 microRNA	SVM	http://nclab.hit.edu.cn/maturepred/	
Mipara[33](2011)	成熟 microRNA	SVM	http://www.whiov.ac.cn/bioinformatics/mirpara	ACC 达到 80%
miRdup[25](2013)	成熟 microRNA	随机森林	http:// www.cs.mcgill.ca/～blanchem/mirdup/	
miRRim2[22](2012)	Pre-microRNA、成熟 microRNA	条件随机场	http://mirrim2.ncrna.org(预测结果)	
mirExplorer[34](2011)	Pre-microRNA、成熟 microRNA	Adaboosting	biocenter.sysu.edu.cn/mir/	93.71%/95.53%
Microprocessor SVM [20](2007)	Drosha 位点	SVM	https://demo1.interagon.com/miRNA/	
miRmat[20](2012)	成熟 microRNA	随机森林	http://mcube.nju.edu.cn/jwang/lab/soft/MiRmat/	

3 基于机器学习的 microRNA 预测方法各环节研究进展

3.1 数据集的选取

数据集选取对象可以划分为：ppri-microRNA 数据集、pre-microRNA 数据集、microRNA-microRNA * 数据集、成熟 microRNA 数据集和下一代测序数据集。因为 pre-microRNA具有保守的发夹结构，所以数据集多数选用 pre-microRNA数据集，随着 microRNA-microRNA * 、成熟 microRNA 特征发掘以及研究功能进一步发展到对成熟 microRNA 起始位点预测、microRNA-microRNA * 双体预测和成熟 microRNA 预测，数据集增加了 microRNA-microRNA * 数据集[48]、成熟 microRNA 数据集和 ppri-microRNA 数据集[33]，而下一代测序技术产生的大量短序列片段也为 microRNA 的预测提供了更好的契机，如 mirExplorer [34]的 mirExplorer-NGS 分类器采用 NCBI 的 GEO 数据库中人的下一代测序数据 GSM416753 到 GSM416761，联合作为下一代测序数据集。

数据集包括正样本集、负样本集和测试集，负样本集的选取尤为重要，负样本集选取方法分为伪 pre-microRNA 提取法、单训练样本方法和随机-起始序列法[33]。

伪 pre-microRNA 提取法。负样本集来自蛋白质编码区、基因组、mRNA 和 ncRNA 中非 pre-microRNA 的发夹结构[29,35]，数据来源包括 UCSC、miRbase 等，提取方法为将数据库中候选序列以一定的滑动窗口[23]，如 90nt[36]、100nt[37,38]、110n[39]、500nt[40-42]进行扫描，按照一定的规则选取与真 pre-microRNA 相似的伪 pre-microRNA 序列[23,26,27,30,40]，选取规则一般为序列长度、茎长度、凸起数量、环数量、自由能、多环等限制条件，由于 ncRNA 中除 microRNA 外的其他 RNA 也具有发夹结构，因此它作为负样本集的一部分[29,34]，增加了软件预测真伪 microRNA 的能力。

单训练样本方法主要考虑到负样本集选取不当会影响分类器性能提高，产生评估偏差，于是出现了不提取负样本集、只采用正样本集的方法，如阳性样本学习方法 PSoL(positive sample only learning) [44]和单类方法(one class methods)[45]，因为正样本集多是已被证实的，所以此方法消除了负样本集特征集无明确定义的缺陷，但是同采用负样本集方法相比性能较低。

随机-起始序列法将每条正样本集序列根据偏移起始提取位置的方法加以改造，构成负样本集，如 miRdup[25]方法训练集负样本集的选取将正样本集重新定位，保留了 microRNA 的长度，但改变了真实 microRNA 的确切位置。Maturepred[23]根据 pre-microRNA 发夹结构上同一臂上不会产生多折叠的 microRNA 理论，将正样本集以非成熟 microRNA起始位点为滑动窗口起始点，采用两个窗口联合方法选取伪 microRNA-microRNA * 双体数据集。Mirpara[33]负样本集通过正样本集数据将成熟 microRNA 起始位点随机移动至少 5bp 的方法取得。

3.2 特征集选取

triplet-svm[26]首次提出三元编码方法，它将序列-结构结合，应用序列中核苷酸与结构上连续配对关系作为重要辨别特征；Mipred[27]将三元编码特征与二级结构最小能和 P 值随机化测试相结合作为特征集；mipred[28]选取了 17 个序列特征，6 个折叠特征，1 个拓扑特征，dG、dP、dQ、dD、dF 5 个标准化转化特征；micropred[29]除了选用 mipred 中 29 个全局内在特征外，提出了 19 个自由能、RNAfold 相关、Mfold 相关、碱基相关特征。三元组特征应用广泛，如 maturepred[23]在数据集 microRNA 和 microRNA-microRNA * 双体上分别提取了三元组编码特征，PlantMiRNAPred[31]是针对植物 pre-microRNA 的预测方法，特征集采用了来自 triplet-SVM 的 32 个特征和 microPred 的 49 个特征，还提取了 3 个自由能和碱基配对相关的新特征。

主要常用特征归纳如下：大小(长度、距离)；稳定性(二级结构最小自由能、热力学特征)；序列(核苷酸频率、核苷酸类别、二核苷酸含量，GC 含量和 microRNA 第一个配对碱基)；结构(最大内环大小、内环个数、碱基配对类型、连续碱基配对个数、不配对碱基数、不配对碱基率、microRNA 位置、上部茎位置、凸环属性、不配对数量、不配对率、不稳定 GU，低端茎处 microRNA 起始点的 3′悬垂部分)；其他特征(三元组、位子特异性特征)等。

随着研究的深入，一些新的特征和特征选取方法被采用。

专门针对序列、结构或二者相结合的方法，如 Mirident[30]采用序列结构相结合特征，采用软件 Teiresias 选取不定长特征模块，最后选取 1300 个特征；FOMmir[46]方法将 RNAfold 产生的点-括号表示法转化为茎环结构，接着再转化为茎-凸起-空白结构，这样避免了茎-分支结构的噪声，将最长茎作为主茎，而其他分支划分为环、凸起和空白；Xiao 的 network-level 方法[47]将茎环二级结构翻译成网络，采用网络参数来构造预测模型，提供了结构解析的新方法；mirExplorer[34]采用了自由能特征和两个概率转移矩阵(TPM)特征、6 个 TPM 扩展特征和 15 个生物相关特征。

强健性特征。因为 microRNA 的茎环结构比其他 RNA 茎环结构更有高内在强健性，一些强健性特征在预测算法上得到应用，如 Z-score、P-value、SC(self-containment) score、PhastCons score[48]、PhyloP score[49]。研究表明 SC score 在动物、植物 pre-microRNA 发夹结构中具有较高的 SC 分值，与其他 RNA 发夹结构相比，它具有右倾斜分布，HeteroMirPred[32]用 SC 分值作为区分其他 RNA 发夹结构和伪发夹结构的特征，miRdup[25]和 Mipred[27]则采用 P-value 特征值，PhastCons score、PhyloP score 是两个进化保守相关特征，PhastCons 考察相邻位置独立性、每一位置进化保守程度，PhyloP score 考察连续区域保守程度，miRRim2[22]采用了这两个特征。

从特征选取的对象来看，最初研究提取 pri-microRNA 和 pre-microRNA 的特征，而最近研究中采用了专门针对 microRNA-microRNA * 和 Drosha 剪切位点的生物特征。由于 microRNA-microRNA * 双体的高度保守性，microRNA-microRNA * 双体特征主要用于预测成熟 microRNA 方法，如 PMirP[50]采用双体结构和碱基配对特征来预测 pre-microRNA；miRRim2[22]应用碱基配对特征；miRdup[25]采用双体中碱基配对数量、碱基配对类型、最小自由、凸出的数量/大小、成熟 microRNA 在 pre-microRNA 中的位置特征；MaturePred[23]采用了双体和它的侧翼位置特异特征、最小自由能特征、稳定性特征、距离和三元组特征；Philip H 的 C5.0 Decision Trees [42]采用 microRNA：microRNA * 双体的自由

能、不匹配数量、GC 含量特征。Drosha 剪切位点相关特征是建立在对 Drosha 剪切位点生物学实验基础上的，J han 在文献[19]通过计算 pri-microRNA 二级结构的热力学稳定性资料，提出 DRB(Drosha recognition base pair)下游大约 11～13bp 位置为 Drosha 剪切位点，而上游的终环位置意义甚微，并提出剪切位点主要决定于茎中第一个不匹配氨基酸位置。根据以上生物特征，miRRim2[22] 针对这一特征提取了 DRB 的进化保守特征，采用 PhastCons 和 PhyloP 分值来表示；miRmat[21] 则计算了 DRB 处的能量特征，发现在该处存在“低能向高能转换点”，利用该特征并结合结构特征预测了 Drosha 和 Dicer 剪切位点，进而预测成熟 microRNA；Microprocessor SVM[20] 对每个候选 Drosha 剪切位点提取除了环的大小外 686 个序列结构相关特征，采用 Drosha 剪切位点特征存在一个问题，即 mirtron 类型的 microRNA 不经 Drosha 剪切，这一部分 microRNA 不能被预测到。

综上所述，特征集的选取是预测性能的一个重要保证，有待于生物学进一步研究发展，同时在现有基础上，更大程度地挖掘潜在特征也是机器学习方法的一个重要研究课题。

3.3 分类器设计

microRNA 预测构建分类器应用的机器学习算法比较多，常采用的方法如下。

支持向量机方法，主要思想是：采用核方法寻找最优分类面，使得不同类别样本之间的间隔最大化。microRNA 预测采用支持向量机方法取得较好预测性能，使用也最广泛，如 mipred[28]、micropred[29]、mirident[30]、MaturePred[23]、Triplet[26]、Microprocessor SVM[20]、Mipara[33] 等均采用支持向量机方法。该方法的优点是：特有的分隔面模式对样本分布、冗余特征和过拟合问题处理效果较好，其泛化能力使其效果和稳定性能较佳；缺点是处理的开销较大，不适合大数据集情况。

决策树方法，基本思想是：对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析，Philip H[42] 采用决策树程序 C5.0，该程序结合 boosting 技术来提高分类器性能，该方法适用于基因组短序列。决策树具有速度优势，但其基于局部信息的样本过滤特性会影响分类结果的稳定性，而且在维度较高情况下，复杂度过高也是一个重要问题。

随机森林，由很多决策树组成，主要思想是：通过自助法(boot-strap)重采样技术，不断生成训练样本和测试样本，由训练样本生成多个分类树组成随机森林，miRdup[25]、miRmat[21]、Mipred[27] 采用随机森林方法，随机森林性能与支持向量机相当，应用也较广泛，优点是定量判断特征重要性，更利于特征选择，算法简单、速度快，适用于部分数据丢失情况。

贝叶斯算法先进行贝叶斯网络分类器的学习，然后进行贝叶斯网络分类器的推理，MatureBayes[24] 和 Dang[63] 采用该方法取得了较好的预测性能，该方法适用于样本特征条件独立且满足高斯分布的特征集，对于不满足此条件的样本，预测准确率低，相对于上述方法预测能力较弱，优点是模型简单、效率较高，它们的改进算法值得研究。

另外，常用方法还有：CSHMM[35]、Nam[37]、HHMMiR[40] 采用隐马尔可夫，xu[36] 采用随机游走，Abeel[51] 采用神经网络方法分别用于 microRNA 预测。

除了上述方法得到广泛使用，基于机器学习的 microRNA 预测最新研究中还采用了其他方法：MiRRim2[22] 采用 Adaboost 方法[34] 构建分类器模型，应用优化特征权重方法构建模型预测真伪发夹结构 Adaboost 方法[34]；MiRRim2[22] 采用条件随机场(conditional random field)[22] 方法构建分类器模型，应用优化特征权重方法构建模型预测真伪发夹结构；G^2DE[52] 将密集评估算法与一般高斯组件相结合构建分类器，与决策树和 SVM 相比，它具有更高的分类器性能；HeteroMirPred[32] 采用多分类器方法，将支持向量机(折中预测负样本集类和正样本集类)、K-近邻方法(在过滤负样本集类上具有高性能)、随机森林(预测正样本集类上具有高性能)3 种方法相结合来建立分类器，取得了较高的预测性能；mirExplorer[34] 首次采用 Adaboost 方法构建了两个模块，其一用于在全基因组中预测 pre-microRNAs 模块，其二用于在下一代测序数据中预测 microRNAs，Boosting 方法在提升较弱分类方法方面效果优于支持向量机，受到广泛关注。

3.4 特征子集选择方法

特征选择技术分为过滤法、缠绕法和嵌入法[53]。

过滤法：过滤法最早是在数据集基础上，将每一特征分别考虑，最终组合成最有特征集。如 Maturepred[23] 采用信息增益方法，PlantMiRNAPred[31] 采取了信息增益及特征相似度方法，HeteroMirPred[32] 采用信息增益、基于关联规则的特征选择算法。由于缠绕法在大数据集上的训练时间长，micropred[29] 方法采用了过滤法，考虑到相互作用的特性更优于特征过滤方法选用了反向淘汰算法(backward elimination algorithm)。MiRpara[33] 采用贪婪算法和 SPSS 方法进行特征过滤。

缠绕法(wrapper methods)：将特征集与学习算法结合考虑，根据评估标准选取特征集。

嵌入法(embedded methods)：特征选择算法本身作为组成部分嵌入到学习算法里。如 Mirdup[25] 随机特征子集的训练采用了决策树与随机森林联合的方法，并且联合了 Adaboost M1 方法，采用信息增益对特征进行排序。决策树方法在每一节点选择分类性能最佳特征，然后对整个选取特征集进行分割，决策树形成即为特征选择选取过程。

总的来讲，过滤法去除冗余特征速度快，计算效率高，但由于没有考虑到特征集之间的相互作用，选取特征集不一定是最佳的，效果不理想，在实际应用中，可作为预筛选器或者在大数据集情况下采用；缠绕法选取特征集规模相对较小，准确率高，但是速度较慢，泛化性差，而且时间复杂度高，适用于对特征选取精度高且数据集较小的情况；嵌入法的优势在于与分类模型的相互作用，它能选到一个理想的特征集，但是时间复杂度太高。

3.5 类不平衡问题解决

类不平衡导致的倾斜会导致大量假阴性预测，解决方法分为基于数据层面(随机欠抽样和随机重抽样)和基于算法层面(分类器集成、代价敏感学习和特征选择方法)[54]，具体算法包括 SMOTE[55]、informed sampling methods、聚类方法、集成学习方法、基于 SVM 集成系统[56]、不同的 boosting 方

法[57]、级联的神经网络[14]、决策树、模糊系统、different error costs (DEC)方法、zSVM 方法等。下面重点介绍 microRNA 预测研究中应用的几种方法。

SMOTH 方法。micropred[29]和 mirExplorer[34]同样采用了 SMOTH 方法。microPred 针对类不平衡问题考察 5 种方法,包括:随机欠采样/随机过采样方法[58]、SMOTE 和多分类器系统 (MCS)[59]、Different error costs (DEC)[60]和 zSVM[61],最后证明 SMOTH 方法效果最佳。Rok Blagus[62]对采用 SMOTH 方法处理 microRNA 预测问题中高维类不平衡问题进行研究,证实 K-邻近方法采用 SMOTH 在处理高维数据不平衡问题中比较有效。这种方法可以避免过拟合问题,但有可能会引入噪声。

SMOTH 改造方法。HeteroMirPred[32]的方法采用改进的 SMOTHbagging 再平衡方法,首先应用 SMOTH 以 50% 比率增加少数类频率,然后采用欠采样方法增加子集内多数类达到数据平衡,共得到 4 个最优平衡数据子集。Xuan Tho Dang[63]提出基于过采样方法 SMOTH 的改进 incre-mental - SMOTH 方法,其在敏感性和 G-平均值方面优于 SMOTH 及其他 SMOTH 改造方法,如 safe-level-SMOTE 和 bor-der-line-SMOTE。

KNN 方法。MaturePred[23]采用两阶段采样算法解决类不平衡问题。首先评估负样本集样本在 K-邻近区域的密度,样本选择算法选择符合数据分布的典型负样本,第二阶段反复选择那些导致当前预测模型产生最大偏离的样本作为负样本集,从而构建平衡训练集。

3.6 评价标准

精确评估预测能力是预测方法和模型高性能的保证,在生物信息学中广泛应用的评价标准包括:式(1)即 Sn 为敏感性指标,是指每类样本中被正确预测的比例;式(2)即 SP 为特异性指标,也就是第 i 类的样本中真正属于第 i 类的比例;式(3)即 ACC 为总预测精度;式(4)即 MCC 为 Matthews 相关系数;式(5)即 PVV 为阳性预测值;式(6)即 FPR 为被预测为正的负样本结果数;式(7)即 Gm(Geometric-mean)几何均值;式(8)即 Fm(F-measure);式(9)即 AGm(Adjusted Geometric-mean)调整的几何均值,另外还有 AUC-ROC 为 ROC 曲线下面积,AUC-PR 为 PR 曲线下面积等。

主要对应公式如下:

$$\text{Sensitivity (Sn)}=\frac{TP}{TP+FN} \tag{1}$$

$$\text{Selectivity or Specificity (Sp)}=\frac{TN}{TN+FP} \tag{2}$$

$$\text{Accuracy (ACC)}=\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$\text{MCC}=\frac{TP\times TN+FP\times FN}{\sqrt{(TP+FP)\times(TN+FN)\times(TP+FN)\times(TN+FP)}} \tag{4}$$

$$\text{Positive Predictive Value (PVV)}=\frac{TP}{TP+FP} \tag{5}$$

$$\text{False Positive Rate(FPR)}=1-\text{Specificity}=\frac{FP}{TN+FP} \tag{6}$$

$$\text{Geometric Mean (Gm)}=\sqrt{(Sn\times Sp)} \tag{7}$$

$$\text{Fm }\beta=\frac{(1+\beta^2)(PPV\times SE)}{\beta^2 PPV+SE} \tag{8}$$

$$\text{AGm}=\text{Gm}-(100\%-SP)\times \text{Nn} \tag{9}$$

其中,TP 为被模型预测为正的正样本,TN 为被模型预测为负的负样本,FP 为被模型预测为正的负样本,FN 为被模型预测为负的正样本。

上述标准中,当 TN 远大于 TP、FP 远大于 FN 时,采用 ACC 评价意义不大,MCC 受到样本比例影响较大,当比例失衡时,无法反映真实分类性能,所以 ACC 和 MCC 用于平衡数据集的性能评估效果较好,Fm、Gm、AUC-ROC、AUC-PR 和 AGm 则用于不平衡数据集机器学习方法的性能评估,但 ACU-ROC 曲线下面积用来作为一个数值进行模型性能评估,不能应用于高度倾斜数据集的性能评估,AUC-PR(Precision Recall)解决了 ROC 曲线存在的问题,但增加了假阳性率的敏感性,AGm 是一个新的评价标准,具有较低的 SP 降低率,增加了敏感性,降低特异性的效果较好[64]。

4 总结和展望

4.1 算法及软件总结

从预测物种上看,多数预测软件提供人的 microRNA 预测,PlantMiRNAPred、MaturePred、HeteroMirPred、Mipara 针对植物提供预测服务,HeteroMirPred 是对动物和植物训练分类器,miRmat 用于脊椎动物预测,miRdup 可用于多物种预测,MatureBayes 除了用于人外还可用于鼠的预测。

从使用上讲,多数提供给定候选 microRNA 预测,还有全基因组序列的预测[33,34]、下一代测序数据的预测[34],进一步细分,如 mirExplorer[34]含多环的发夹结构预测,MatureBayes 和 MaturePred 分别预测长度为 22nt 和 21nt 的成熟 microRNA,Microprocessor SVM 和 miRmat 不能预测 mitron microRNA 类型成熟 microRNA,miRmat 和 miRRim2 预测 pre-microRNA 长度为 50～80nt 范围内的成熟 microRNA。

从友好性讲,分为单机运行和在线服务,运行系统包括 windows 和 linux。miRdup[25]、Triplet[26]、micropred[29]、mirExplorer[34]只提供单机运行,PlantMiRNAPred、Maturepred[23]只提供在线服务,MatureByes[24]、Microprocessor SVM[20]提供在线服务和单机运行,miRRim2[22]只提供文献预测结果。

从研究角度讲,分为提供源代码和提供预测服务两种。miRdup[25]、MatureByes[24]、Triplet[26]、micropred[29]、miri-dent[30]、HeteroMirPred[32]、Microprocessor SVM[20]免费提供源代码,可根据需要进行修改编译。

4.2 面临的挑战和展望

机器学习方法在 microRNA 预测问题上取得较高预测性能,对 pre-microRNA 达到 90%以上的预测精度,对成熟 microRNA 的预测虽然准确度不高,但是在一定偏离标准下预测精度较高;展开了针对多物种的预测方法研究,部分算法实现全基因组序列的预测功能,但是研究还面临一些挑战,主要有如下方面:(1)实验验证的 microRNA 训练样本相对于海量的未验证数据数据量有限,这样对样本空间的模拟较困难,存在过拟合和数据偏斜问题;(2)特异性问题,虽然现有方法在训练集和测试集上测试达到很高预测精度,但是在全基因组

预测时，会预测成千上万的 microRNA 候选基因，这其中存在假阳性结果，而且真正 microRNA 的漏检率也是很高的，所以提高特异性是当前面临的问题之一；(3)数据集瓶颈，实验验证的 microRNA 序列有限，而基于下一代测序数据，由于低表达 microRNA 具有种族特异性和条件特异性，Dicer 产物和 microRNA * 完全衰退，因此很难被探测到，而数据库中数据除了高端错误(如序列错误)以外，还会出现一些短序列映射错误。所以如何利用少量的阳性样本和大量预测样本是当前研究的一个方向。

今后机器学习方法在 microRNA 预测研究上的主要研究方向有以下几个方面：(1)从功能上讲，最初的研究都是针对 pre-microRNA，随着研究深入，成熟 microRNA 的预测有待于进一步研究。(2)算法的强健性提高。虽然已经有很多预测算法取得良好性能，但是假阳性高是目前面临的重要问题，预测精度也有待继续提高。(3)全基因组、多物种预测。当前算法多为针对一种或几种物种的预测，多物种、提高预测算法计算能力、实现全基因组范围预测功能也是研究方向之一。(4)当前研究都是基于序列和二级结构等特征，随着三维结构预测技术的成熟，将三维结构信息应用到 microRNA 预测中，将提高预测性能和精度。(5)结合下一代测序数据的机器学习方法应用，下一代测序技术为 microRNA 的预测提供了契机，如今，第三代测序技术[65]的研究已经展开，基于新的测序数据的机器学习方法预测 microRNA 有待研究。

总之，microRNA 预测对 microRNA 功能研究意义重大，在实验方法未取得下一步进展之前，基于机器学习方法对 microRNA 预测将得到全方位、多角度的研究。

参 考 文 献

[1] Baltimore D, Boldin M P, Connell M O, et al. MicroRNAs: new regulators of immune cell development and function[J]. Nature, 2008, 9(8): 839-845

[2] Song L, Tuan R S. MicroRNAs and cell differentiation in mammalian development[J]. Birth Defects Research Part C: Embryo Today: Reviews, 2006, 78(2): 140-149

[3] Bartel D P. MicroRNA: genomics, biogenesis, mechanism, and function[J]. Cell, 2004, 116(2): 281-297

[4] Hammond S M. MicroRNAs as oncogenes [J]. Current opinion in genetics & development, 2006, 16: 4-9

[5] Brennecke J, Hipfner D R, Stark A, et al. Cohen SM: bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila[J]. Cell, 2003, 113(1): 25-36

[6] Rogaev E I. Small RNAs in human brain development and disorders[J]. Biochemisry, 2005, 70 (12): 1404

[7] Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data[J]. Nucleic Acids Res., 2009, 39: D152-D157

[8] Büssing I, Slack F J, Großhans H, et al. let- microRNAs in development, stem cells and cancer[M]. Trends Mol Med, 2008

[9] Nielsen C B, Shomron N. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs[J]. RNA, 2007, 13: 1894-1910

[10] Hertel J, Lindemeyer M, Missal K, et al. The expansion of the metazoan microRNA repertoire[J]. BMC Genomics, 2006: 7-25

[11] Lee Y, Jeon K, Lee J T, et al. MicroRNA maturation: Stepwise processing and subcellular localization[J]. The EMBO Journal, 2002, 21(17): 4663-4670

[12] Eugene B. Evolution of microRNA diversity and regulation in animals[J]. Nature, 2011: 846-860

[13] Lee Y, Ahn C. The nuclear RNase III Drosha initiates microRNA processing[J]. Nature, 2011, 425: 415- 419

[14] Elsebet L, Stephan G. Nuclear Export of MicroRNA Precursors[J]. Science, 2004, 303: 95-98

[15] Hutvágner G, Zamore P D. A microRNA in a Multiple-Turnover RNAi Enzyme Complex[J]. Science, 2002, 297: 2056-2060

[16] Reinhart B J, Bartel D P. Small RNAs Correspond to Centromere Heterochromatic Repeats[J]. Science, 2012, 297 (5588): 1831

[17] Gregory R I, Chendrimada T P, Cooch N, et al. Human RISC couple microRNA biogenesis and posttranscriptional gene silencing[J]. cell, 2005, 123: 631-640

[18] Filipowicz W, Bhattacharyya S N, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? [J]. Nature Reviews Genetics, 2008, 9: 102-114

[19] Han J, Lee Y, Yeom K H, et al. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex[J]. Cell, 2009, 125: 887-901

[20] Helvik S A, Snove O, Sætrom P, et al. Reliable prediction of Drosha processing sites improves microRNA gene prediction [J]. Bioinformatics, 2003, 23: 142-149

[21] He C, Li Y-X, Zhang G X, et al. MiRmat: Mature microRNA Sequence Prediction[J]. PLoS ONE, 2012, 7(12): e51673

[22] Terai G, Okida H, Asai K, et al. Prediction of Conserved Precursors of microRNAs and Their Mature Forms by Integrating Position-Specific Structural Features[J]. PLoS ONE. 2012, 7(9): e44314

[23] Xuan P, Guo M, Huang Y C, et al. MaturePred: Efficient Identification of MicroRNAs within Novel Plant Pre-microRNAs[J]. PLoS ONE, 2011, 6(11): e27422

[24] Gkirtzou K, Tsamardinos I, Tsakalides P, et al. MatureBayes: A Probabilistic Algorithm for Identifying the Mature microRNA within Novel Precursors[J]. PLoS ONE, 2010, 5(8): e11843

[25] Leclercq M, Diallo A B, Blanchette M. Computational prediction of the localization of microRNAs within their pre-microRNA [J]. Nucleic Acids Research, 2013, 41(15): 1-12

[26] Xue C, Li F, He T, et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine[J]. BMC Bioinformatics, 2005(6): 310

[27] Jiang P, Wu H, Wang W, et al. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features [J]. Nucleic Acids Research, 2007, 35: 339-344

[28] Ng K L S, Mishra S K. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures[J]. Bioinformatics, 2007, 23: 1321-1330

[29] Batuwita R, Palade V. microPred: Effective classification of pre-microRNAs for human microRNA gene prediction[J]. Bioinformatics, 2009, 25(8): 989-995

[30] Liu X, He S, Skogerbø G, et al. Integrated Sequence-Structure Motifs Suffice to Identify microRNA Precursors [J]. PLoS ONE, 2012, 7(3): e32797

[31] Xuan P, Guo M Z, Liu X Y, et al. PlantMiRNAPred: efficient classification of real and pseudo plant pre-microRNAs[J]. Bioinformatics, 2011, 27 (10): 1368-1376

[32] Lertampaiporn S, Thammarongtham C, Nukoolkit C, et al. Heterogeneous ensemble approach with discriminative features and modified-SMOTE bagging for pre-microRNA classification[J]. Nucleic Acids Research, 2013, 41: 1-21

[33] Wu Y G, Wei B, Liu H Z, et al. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences[J]. BMC Bioinformatics, 2011(12): 107

[34] Guan D G, Liao J Y, Qu Z H, et al. mirExplorer Detecting microRNAs from genome and next generation sequencing data using the Adaboost method with transition probability matrix and combined features[J]. RNA Biology, 2011, 8(5): 922-934

[35] Agarwal S, Vaz C, Bhattacharya A, et al. Prediction of novel precursor microRNAs using a context-sensitive hidden Markov model (CSHMM)[J]. BMC Bioinformatics, 2010, 11(Suppl 1): S29

[36] Xu Y, Zhou X, Zhang W. MicroRNA prediction with a novel ranking algorithm based on random walks[J]. Bioinformatics, 2008, 24: 50-58

[36] Nam J W, Shin K R, Han J, et al. Human microRNA prediction through a probabilistic co-learning model of sequence and structure[J]. Nucleic Acids Res, 2005, 33: 3570-3581

[38] Brameier M, Wiuf C. Ab initio identification of human microRNAs based on structure motifs [J]. BMC Bioinformatics, 2007, 8: 478

[39] Yousef M, Nebozhyn M, Shatkay H, et al. Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier[J]. Bioinformatics, 2006, 22: 1325-1334

[40] Kadri S, Hinman V, Benos P V. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models[J]. BMC Bioinformatics, 2009, 10(Suppl 1): S35

[41] Pfeefer S, Sewer A, Lagos-Quintana M, et al. Identification of microRNAS of the herpesvirus family [J]. Nature Methods, 2005, 2(4): 269-276

[42] Williams P H, Eyles R, Weiller G. PlantMicroRNA Prediction by Supervised Machine Learning Using C5. 0 Decision Trees [J]. Journal of Nucleic Acids, 2012: 652979

[43] Zhang Y W, Yang Y Y, Zhang H, et al. Prediction of novel pre-microRNAs with high accuracy through boosting and SVM[J]. Bioinformatics, 2011, 27: 1436-1437

[44] Wang C, Ding C, Meraz R F, et al. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes[J]. Bioinformatics, 2006, 22: 2590-2596

[45] Yousef M, Jung S, Showe L C, et al. Learning from positive examples when the negative classis undeter mined microRNA gene identification[J]. Algorithms for Molecular Biology, 2008(3): 2

[46] Shen W, Chen M, Wei G, et al. MicroRNA Prediction Using a Fixed-Order Markov Model Based on the Secondary Structure Pattern[J]. PLoS ONE, 2012, 7(10): e48236

[47] Xiao J X, Tang X J, Li Y Z, et al. Identification of microRNA precursors based on random forest with network-level representation method of stem-loop structure[J]. BMC Bioinformatics, 2011(12): 165

[48] Siepel A, Bejerano G, Pedersen J S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes [J]. Genome Res, 2005, 15: 1034-1050

[49] Siepel A, Pollard K S, Haussler D. New methods for detecting lineage-specific selection[C]// Proceedings of the Tenth Annual International Conference on Computational Molecular Biology (RECOMB 2006). 2006: 190-205

[50] Zhao Dong-yu, Wang Yan. PMirP: A pre-microRNA prediction method based on structure-sequence hybrid features[J]. Artificial Intelligence in Medicine, 2010, 49: 127-132

[51] Abeel T, Helleputte T. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods[J]. Bioinformation, 2010, 26: 392-398

[52] Hsieh C H, Chang D T H. Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm[J]. BMC Bioinformatics, 2010, 11(Suppl 1): S52

[53] 张丽新，王家钦，赵雁南，等. 机器学习中的特征选择[J]. 计算机科学，2004，31(11)：180-184

[54] Batuwita R, Palade V. FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning[J]. IEEE Transactions on Fuzzy Systems, 2010, 18(3): 558-571

[55] Chawla N, Bowyer K, Kegelmeyer P. SMOTE: Synthetic minority over-sampling technique, Artif[J]. Intell. Res, 2002, 16: 321-357

[56] Li P, Wang X L, Liu Y C, et al. A classification method for imbalance data set based on hybrid strategy[J]. Acta Electronica Sinica, 2007, 35(11): 2161-2165

[57] Guo H, Viktor H L. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach[J]. ACM SIGKDD Explorations, 2004, 6(1): 30-39

[58] Weiss G. Mining with rarity: a unifying framework[J]. SIGKDD Expl, 2004(6): 7-19

[59] Molinara M, Ricamato M T, Tortorella F, et al. Facing imbalance classes through aggregation of classifiers [C] // Proc. of 14th ICIAP. IEEE Comp. , 2007: 43-48

[60] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets[C]// Proc. of 15th ECML. 2004: 39-50

[61] Imam T, Ting K M, Kamruzzaman J. z-SVM: an SVM for improved classification of imbalanced data [C] // Proc. of 19th AUS-AI. 2006: 264-273

[62] Blagus R, Lusa L. SMOTE for high-dimensional class imbalanced data[J]. BMC Bioinformatics, 2013, 14: 106

[63] Dang X T, Osamu H, Saethang T, et al. A novel over-sampling method and its application to microRNA prediction[J]. Biomedical Science and Engineering, 2013, 6: 236-248

[64] Rukshan B. Adjusted geometric-mean: A novel performance measure for imbalanced bioinformatics datasets learning [J]. Journal of Bioinformatics and Computational Biology, 2012, 10: 125003

[65] Rusk N. Cheap third-generation sequencing[J]. Nat Methods, 2009, 6(4): 244-245