大豆科学 Soybean Science
http://ddkx.haasep.cn
2020,39(3):401-405
DOI:10.11861/j.issn.1000-9841.2020.03.0401

基于递归特征消除和随机森林融合算法的大豆前体 MicroRNA 预测模型研究

安　宇,陈桂芬,李　静

(吉林农业大学 信息技术学院,吉林 长春 130118)

摘　要:随着大豆 RNA 基因的生物调控作用研究的不断深入,利用数据挖掘技术对大豆前体 MicroRNA(pre-microRNA)进行有效的预测已成为该领域的重要发展方向。针对常规的随机森林算法在 pre-microRNA 预测模型中存在识别精度较低的问题,研究提出并构建基于递归特征消除(recursive feature elimination, RFE)与随机森林(random forest, RF)融合算法的大豆 pre-microRNA 预测模型。首先利用递归特征消除法筛选大豆 pre-microRNA 序列的最优特征子集;然后结合随机森林算法构建大豆 pre-microRNA 的预测模型;最后利用十折交叉验证法,将递归特征消除与随机森林(RFE-RF)融合模型的预测结果与单一随机森林和支持向量机分类模型的预测结果对比。研究结果表明:融合后构建的大豆 pre-microRNA 预测模型精度有明显提高,达到 84.62%,相比于支持向量机算法(support vector machine, SVM)构建的模型精度提高了 17.02%,相比于单独使用随机森林算法构建的模型精度提高了 14.58%。该研究方法为大豆的 pre-microRNA 基因预测提供了新思路。

关键词:大豆;Pre-microRNA;递归特征消除;随机森林;预测模型

Research on Soybean Pre-Micro RNA Prediction Model Based on Recursive Feature Elimination and Random Forest Fusion Algorithm

AN Yu, CHEN Gui-fen, LI Jing

(College of Information Technology, Jilin Agricultural University, Changchun 130118, China)

Abstract: With the continuous in-depth research on the biological regulatory effects of small genes in soybean, the use of data mining technology to effectively predict the pre-MicroRNA of soybean has become an important development direction in this field. To solve the problem that conventional Random Forest (RF) algorithm has low recognition accuracy in pre-MicroRNA prediction model, this study proposed and constructed a soybean pre-microRNA prediction model based on Recursive Feature Elimination (RFE) and RF fusion algorithm. Firstly, we used the RFE method to select the optimal feature subset of soybean pre-MicroRNA sequences. Then, we constructed a prediction model of soybean pre-MicroRNA based on RF algorithm. Finally, we compared the prediction results of the RFE-RF fusion model with the prediction results of the single RF and Support Vector Machine (SVM) classification model. The results showed that the accuracy of the soybean Pre-MicroRNA prediction model constructed after fusion was significantly improved, reaching 84.62%, 17.02% higher than the model constructed by SVM algorithm, and 14.58% higher than the model constructed by RF algorithm alone. This method provides a new idea for the prediction of pre-MicroRNA genes in soybean.

Keywords: Soybean; Pre-MicroRNA; Recursive Feature Elimination(RFE); Random Forest(RF); Prediction model

MicroRNA 是一类广泛存在于真核生物中、长度约为 19 ~ 25 个碱基的小分子非编码 RNA 基因。成熟的 microRNA 来源于其发夹前体(pre-microRNA),通常在转录后水平调整控制靶基因的表达。研究发现,microRNA 参与发育、代谢、抗病、逆境胁迫等多种生物途径,并起到非常重要的调控作用[1-2]。相比动物中的 microRNA,植物 microRNA 的研究还比较少。最早被发现的 MicroRNA 是 1993 年秀丽隐杆线虫中的 Lin-4,而关于植物 microRNA 的最早报道于 2002 年[3]。近年来,随着生物信息学的快速发展,大豆(*Glycine max*)、玉米(*Zea mays*)、水稻(*Oryzasativa*)、小麦(*Triticum aestivum*)等农作物中的 microRNA 不断被人们发现,然而研究主要集中在拟南芥、水稻等模式植物上,在大豆中的研究还比较少。对 microRNA 的预测及功能研究方面,2007 年金伟波等[4-5]根据已报道的水稻 pre-microRNA序列与结构信息在 microRNA 前体上预测成熟区构建模型,预测水稻成熟区的敏感性和特异

收稿日期:2019-10-09
基金项目:国家星火计划(2015GA660004);吉林省重点科技研发项目(20180201073SF)。
第一作者简介:安宇(1993-),女,硕士,主要从事生物信息学与计算机农业应用研究。E-mail:18626726453@163.com。
通讯作者::陈桂芬(1956-),女,博士,教授,主要从事人工智能与计算机农业应用研究。E-mail:guifchen@163.com。

性分别为 86.7% 和 100%。2009 年刘永鑫等[6]以大豆幼叶、老叶、根、茎组织为研究材料,采用 RT-PCR 的方法对大豆 microRNA 进行预测鉴定。同年陈旭等[7]通过计算机预测法和基因克隆法共找到 18 条玉米新 microRNA。2014 年 Huang 等[8]对多年生黑麦草进行研究,第一次通过计算方法检测到了 33 个潜在的 microRNA 标靶。同年,李小平等[9]通过挖掘大豆基因组信息并参考拟南芥同源基因序列,分析大豆的生长素响应因子 GmARF16 和 *Gm-miR160* 作用位点,设计出 *mGmARF16* 序列并成功构建了 pGmARF16::mGmARF16 双元表达载体。2016 年倪志勇等[10]采用生物信息学方法对大豆的 *gma-miR1510a* 靶基因和启动子中的顺式作用元件进行研究和分析,表明 *gma-miR1510a* 启动子具有一些参与非生物胁迫、植物激素、组织特异表达和光应答元件。可见,MicroRNA 的预测及功能研究的发展对推进相关科学研究具有重要意义,而现阶段利用机器学习的方法准确高效地挖掘出更多未知的 microRNA 同样具有十分重要的研究价值。

目前在 microRNA 预测模型中,多采用机器学习的算法构建模型的分类器。其中随机森林算法通过自助法(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,泛化能力强,计算速度快的同时具有良好的预测性能[11]。采用该算法构建 microRNA 预测模型的思路中,Jiang 等[12]利用 P 值和最小自由能等特征,设计了一种基于随机森林的方法,用于使用混合特征对 pre-microRNA 及伪 pre-microRNA 进行分类。Huang 等[13]提取病毒的 microRNA 序列特征和二级结构特征 54 个,建立了支持向量机和随机森林的模型,其平均精度达到 83%。递归特征消除的随机森林融合算法(RFE-RF)优势是将需要的特征集合初始化为整个数据集合,每次去除 1 个排序准则分数最小的数据,直到获得最后的特征集。首先是随机森林的过程,利用 bootstrap 重抽样方法从原始样本中抽取样本,对每个 bootstrap 样本构建决策树,所有的决策树构成随机森林;然后在回归模型中计算特征重要度,此时引入递归特征消除算法,删除特征重要度小的特征后再次构建模型,直至最后特征只剩下 1 个。

本研究以大豆的 pre-microRNA 作为训练及测试样本,构建一个适用于大豆 pre-microRNA 的预测模型。随机森林算法能够处理高维度数据,并且不用做特征选择,常被用来构建分类模型,但考虑到该算法处理数据时,可能会陷入局部最优解或在某些包含一定噪声数据的分类上出现过拟问题。本研究采用递归特征消除与随机森林融合的方法,以提高大豆 pre-microRNA 预测模型的精度,为预测新的大豆 pre-microRNA 提供一种新思路。

1 材料与方法

1.1 数据的获取与预处理

1.1.1 *数据获取* 本研究共获取了 873 条试验数据。从 MiRBase(Release 22.1)和 PMRD 数据库获取大豆 pre-microRNA 序列 433 条作为阳性样本数据集合;选取序列长度、GC 含量、自由能、不配对碱基数量 4 个特征值均与大豆 pre-microRNA 相似的伪 pre-microRNA 序列,在 NCBI 数据库中获取符合要求的大豆 mRNA、基因组、蛋白质编码区序列共 440 条,作为阴性样本数据集合。

在阳性样本中随机取出 2/3 的数据,在阴性样本中随机取出 2/3 的数据,共同作为训练集(train-set),剩余数据作为测试集(test-set)。

1.1.2 *数据预处理* 特征提取:将全部样本的碱基序列信息特征量化为可以被分类器学习的数字特征。首先提取序列特征,其中包括一、二、三联碱基组成以及序列长度,共 86 个;然后利用 RNAfold 计算大豆 Pre-MicroRNA 的二级结构,提取最小自由能(MFE)、配对碱基率、不配对碱基率、发夹结构尺寸、环个数 5 个结构特征;最后采用左三元编码法生成 1 组 32 维的特征信息,与前面的 86 个序列特征和 5 个结构特征组成 123 维的特征集合。

归一化处理:采用 0 均值归一化的方法对特征数据进行预处理,将原始特征数据化为均值为 0、方差为 1 的数据,归一化公式为:

$$Z=\frac{x-\mu}{\sigma} \tag{1}$$

其中,μ 为所有样本数据的均值,σ 为所有样本数据的标准差。

1.2 方 法

1.2.1 *随机森林算法* 随机森林在解决分类问题中其基本思想是一棵森林包含多个决策树,每一个决策树是由可放回的随机抽样样本构造的,通过最终投票对未知类别的样本进行分类[14-15]。随机森林的分类过程为:①用 Bootstrap 采样从样本集中选出 n 个样本;②从所有属性中随机选择 k 个变量,并利用这 k 个变量和 n 个样本构建一棵决策树;③重复前面两步,直到构建出 m 棵决策树;④用 m 棵决策树进行预测分类,并对 m 个结果采用加权或者投票的方式获取最终的预测结果。

1.2.2 *递归特征消除* 递归特征消除是一种寻找最优特征子集的贪心算法,其主要思想是反复构建

模型,最终选出分类的最佳特征集[16]。根据系数选出最好或者最坏的特征,把选出来的特征移除出来,然后在剩余的特征集合中重复这个过程,直到所有特征都遍历了为止[17-18]。其循环过程为:①训练分类器;②计算置换的重要性测度;③消除不相关变量;④用消除后的特征再次训练分类器。

1.3　基于递归特征消除的随机森林融合算法大豆 pre-microRNA 预测模型

本研究使用基于 RFE-RF 的特征选择方法,通过计算模型的均方根误差值确定最优子集,该特征选择方法选择过程包括输入和输出过程。

输入:训练数据集 F(n 个样本,p 个特征),类标签(n,1)。

第一步:初始特征集合 Fx 为原始数据集,最优特征集合 Fy 为空,最优特征子集均方根误差值为 Rx。

第二步:由 Fx 经过 bootstrap 重采样生成决策树建立随机森林分类模型,经过投票得到最终分类结果;计算均方根误差值 Rx,并且按照特征评分数的绝对值$|C|$降序排列。其中,第 i 个特征的特征评分分数计算公式为:

$$c_i = w_i^2 \tag{2}$$

第三步:删除子集 Fx 中排名靠后的特征 Fi,直到特征集合 Fx 为空。

若特征子集 Fy 的均方根误差值 Ry 小于 Rx,那么 $Ry = Rx$,否则执行第二步和第三步。

输出:最优特征子集 Fy。

RF-RFE 算法的流程图如图 1 所示。

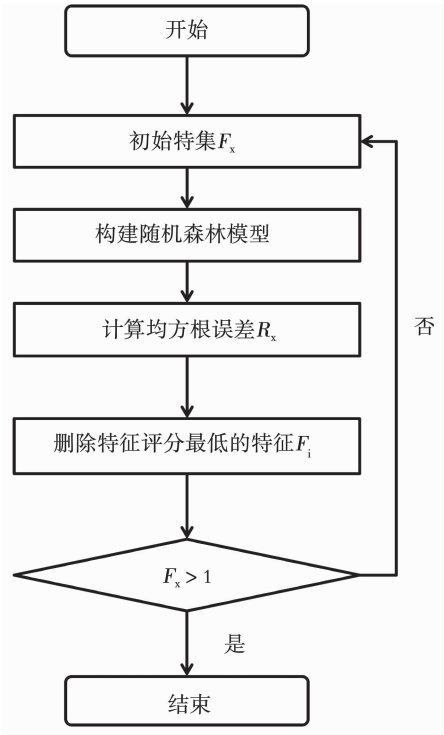


图 1　RFE-RF 算法流程图

Fig. 1　Flow chart of RFE-RF algorithm

2　结果与分析

2.1　RFE-RF 特征选择

RF-RFE 算法在大豆 pre-microRNA 特征选择的迭代过程中,重新评价当前的剩余特征集合,每个特征的得分在反复迭代过程中,通过不断地调整,可以克服单次随机森林的特征选择结果需要反复试验得到特征子集的缺点[19]。基于 RFE-RF 在寻找大豆 pre-microRNA 最优特征子集过程中的均方根误差(RMSE)变化图(图 2),随着递归特征消除反复构建随机森林模型,当特征值为 20 时,均方根误差达到最低值,此时分类效果最佳。

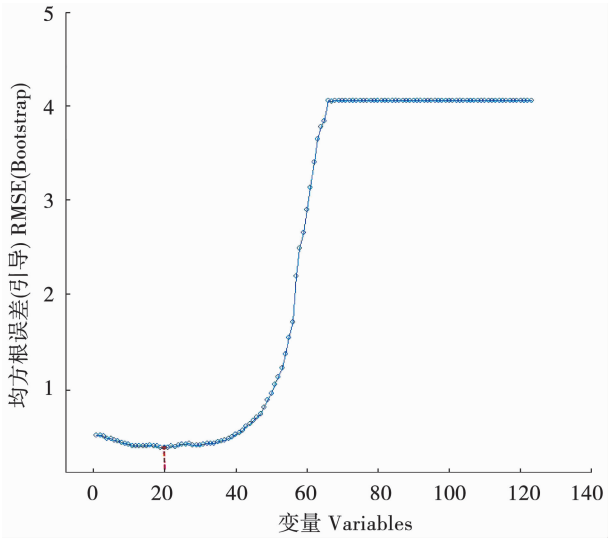


图 2　RFE-RF 大豆 pre-microRNA 特征选择与均方根误差变化曲线图

Fig. 2　RFE-RF soybean pre-microRNA characteristics selection and root-mean-square error variation curve

2.2　基于 RFE-RF 算法的大豆 pre-MicroRNA 预测模型

为比较支持向量机算法(SVM)、随机森林算法(RF)与递归特征消除与随机森林融合算法(RFE-RF)3 种算法在大豆 pre-microRNA 预测中的效果差异,分别构建 SVM、RF 和 RFE-RF 的大豆 pre-microRNA 预测模型。模型预测采用十折交叉验证法,重复计算 10 次,最后将 3 种算法预测模型的敏感度、特异性、误报率及精度取平均值进行对比研究,在相同的独立测试集中的对比试验结果。敏感度是阳性样本分类的准确度,即真阳性率;特异性是阴性样本分类的准确度,即真阴性率;误报率是阴性样本的错误率,即假阳性率;精度是预测值为阳性样本的正确率。其中,敏感度,特异性,精度 3 项

指标越大模型分类效果越好,误报率越小模型分类效果越好。由表 1 可以看出:对大豆 pre-microRNA 进行预测的模型中,SVM 算法构建的模型与 RF 算法构建的模型在敏感度、特异性、误报率及精度 4 项指标上结果均接近,基于 RF 算法构建的大豆 pre-microRNA 预测模型效果略优;基于 RFE-RF 融合算法构建的大豆 Pre-MicroRNA 预测模型敏感度、特异性、误报率和精度分别达到 88.00%、85.19%、14.39%、84.62%,与 RF 预测模型相比,4 项指标分别提高 2.67%、提高 1.06%、降低 37.38%、提高 14.58%。证明 RFE-RF 融合算法构建大豆 pre-microRNA 预测模型有更好的识别能力和泛化能力。

表 1 大豆 pre-microRNA 预测试验 3 种算法结果比较

Table 1 Comparison of experimental results of 3 algorithms on soybean pre-MicroRNA prediction

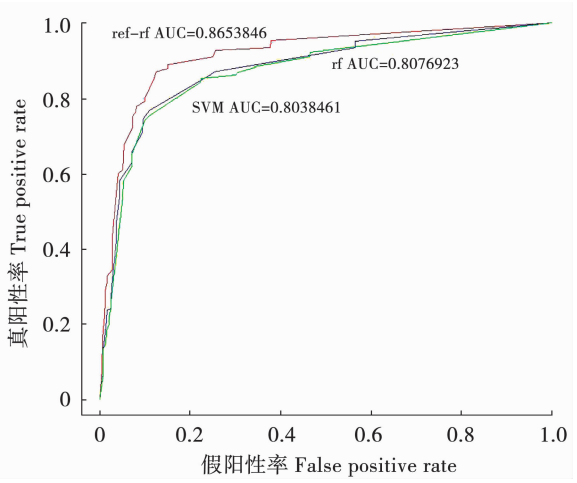
算法 Algorithm	敏感度 Sensitivity	特异性 Specificity	误报率 False positive rate	精度 Precision of measurement
SVM	0.8545	0.7600	0.2400	0.7231
RF	0.8571	0.7703	0.2298	0.7385
RFE-RF	0.8800	0.8519	0.1439	0.8462

2.3 个模型 ROC 曲线对比分析

利用试验结果分别绘制基于 SVM、RF 与 RFE-RF 算法的大豆 pre-microRNA 预测模型 ROC 曲线,曲线下方的面积即为 AUC 值,AUC 越大,模型效果越好。由 ROC 曲线可知,SVM 构建的大豆 pre-microRNA 预测模型 AUC = 0.803 846 1,RF 构建的大豆 pre-microRNA 预测模型 AUC = 0.807 692 3,RFE-RF 融合算法构建的大豆 pre-microRNA 预测模型 AUC = 0.865 384 6(图 3)。RF 算法与 SVM 算法构建的模型 AUC 值相近,说明在大豆 pre-microRNA 的预测中,RF 算法与 SVM 算法构建的模型效果接近,RF 算法构建的预测模型效果略好;RFE-RF 算法构建的模型 AUC 值较 RF 算法高0.057 692 3,表明随机森林算法与递归特征消除算法融合可以提高大豆 pre-microRNA 预测模型的预测效果;当真阳性率相同时,RFE-RF 融合算法构建的大豆 pre-microRNA 预测模型假阳性率更低。用 ROC 曲线综合权衡测试集上的分类性能,验证了基于 RFE-RF 算法的大豆 pre-microRNA 预测模型分类效果更好,能兼顾真阳性率和假阳性率。

3 讨 论

由于物种之间存在差异性,pre-microRNA 预测模型并不能通用[20],现有研究中缺少专门针对大豆



绿色:SVM; 蓝色:RF; 红色:REF-RF。

Green: SVM; Blue: RF; Red: REF-RF.

图 3 3 个模型 ROC 曲线对比图

Fig. 3 Comparison of the ROC curves of the three models

pre-microRNA 预测的模型,而本研究仅在大豆这一物种的基因序列及蛋白编码序列中筛选出阳性数据和阴性数据,利用这些从大豆中筛选出的数据训练预测模型,从而构建适用于大豆 pre-microRNA 预测的模型。本研究基于递归特征消除的自身变量多轮训练避免过度拟合的优势与随机森林算法结合,提出了基于递归特征消除和随机森林融合算法的大豆 pre-microRNA 预测方法,并进行试验结果对比分析,并验证了该融合算法的正确性和有效性,证明基于递归特征消除和随机森林融合算法的大豆 pre-microRNA 预测方法具有理论意义和应用价值。

递归特征消除能降低重要性度量相关性的效果,避免过度拟合,是一种有效的特征选择方法[21]。基于递归特征消除与随机森林融合算法的大豆 pre-microRNA 预测模型较单独使用随机森林算法、构建的大豆 pre-microRNA 预测模型精度提高 14.58%,证明递归特征消除方法与随机森林算法融合使用,应用到大豆 pre-microRNA 预测模型中可提高分类性能。

试验过程中大豆 pre-microRNA 的样本数据量较少,试验结果存在一定的局限性,导致分析结果可能存在一定偏差;构建的大豆 pre-microRNA 预测模型,只针对大豆的 pre-microRNA 数据进行了训练及测试,由于物种间 microRNA 的差别,对于其它作物 microRNA 的预测模型还有待研究。

4 结 论

随着对 MicroRNA 的不断深入研究,采用机器

学习的方法对挖掘出新的 pre-microRNA 已达到了明显的效果。本研究构建出一个基于递归特征消除与随机森林融合算法且针对大豆 pre-microRNA 的预测模型,将两种算法融合后,模型预测精度有了明显的提高,达到 84.62%。可见该研究方法为大豆的 pre-microRNA 基因预测提供了新思路,下一步可以通过优化算法提高模型的预测性能。

参考文献

[1] Bartel D P. MicroRNAs: Genomics, biogenesis, mechanism, and function[J]. Cell, 2004, 116: 281-297.

[2] Ambros V. The functions of animal MicroRNAs[J]. Nature, 2004, 431(76): 350-352.

[3] Reinhart B J, Weinstein E G. MicroRNAs in plant[J]. Gene Development, 2002, 16(13): 1616-1626.

[4] 金伟波, 李楠楠, 吴方丽, 等. 水稻 MicroRNA 的预测及实验验证[J]. 中国生物化学与分子生物学报, 2007, 23(9): 743-750. (Jin W B, Li N N, Wu F L, et al. Prediction and experimental verification of rice MicroRNA [J]. Chinese Journal of Biochemistry and Molecular Biology, 2007, 23 (9): 743-750.)

[5] 金伟波. 基于支持向量机方法的植物 miRNA 预测及小麦 miRNA 的克隆[D]. 杨凌: 西北农林科技大学, 2007. (Jin W B. Prediction of miRNA in plants and cloning of miRNA in wheat based on support vector machine [D]. Yangling: North West Agriculture and Forestry University, 2007.)

[6] 刘永鑫, 韩英鹏, 常玮, 等. 一种适合大豆 MicroRNA 鉴定的 RT-PCR 方法[J]. 大豆科学, 2009, 28(4): 600-604. (Liu Y X, Han Y P, Chang W, et al. A RT-PCR method suitable for identification of soybean MicroRNA [J]. Soybean Science,2009, 28(4):600-604.)

[7] 陈旭. 玉米 microRNA 的计算机预测与克隆及在干旱下的差异表达分析[D]. 雅安: 四川农业大学,2009. (Chen X. Computer prediction and cloning of maize microRNA and differential expression analysis in drought [D]. Ya'an: Sichuan Agricultural University, 2009.)

[8] Huang Y, Zou Q, Sun X H, et al. Computational identification of microRNAs and their targets in perennoal ryegrass (*Lolium perenne*)[J]. Applied Biochemistry and Biotechnology, 2014, 173(4): 1011-1122.

[9] 李小平,曾庆发,赵娟. 大豆生长素响应因子 *GmARF16* 器官表达特征及抗降解表达载体的构建[J]. 大豆科学, 2014, 33(5):661-666. (Li X P, Zeng Q F, Zhao J. Expression characteristics of soybean auxin response factor *GmARF16* organ and construction of anti-degradation expression vector [J]. Soybean Science, 2014, 33(5): 661-666.)

[10] 倪志勇,于月华,陈全家, 等. 大豆 *gma-miR1510a* 生物信息学分析及人工 microRNA 植物表达载体构建[J]. 大豆科学, 2016, 35(2): 239-244. (Ni Z Y, Yu Y H, Chen Q J, et al. Bioinformatics analysis of soybean *gma-miR1510a* and construction of artificial microRNA expression vectors [J]. Soybean Science, 2016, 35(2): 239-244.)

[11] 王颖, 李金, 王磊, 等. 基于机器学习的 microRNA 预测方法研究进展[J]. 计算机科学,2015,42(2):7-13. (Wang Y, Li J, Wang L, et al. Research progress of microRNA prediction method based on machine learning [J]. Computer Science, 2015, 42(2): 7-13.)

[12] Jiang P, Wu H, Wang W, et al. MiPred:Classification of real and pseudo MicroRNAs precursors using random forest prediction model with combined features [J]. Nucleic Acids Research, 2007, 35: 339-343.

[13] Huang K Y, Lee T Y, Teng Y C, et al. ViralmiR: A support-vector-machine-based method for predicting viral microRNA precursors[J]. BMC Bioinformatics, 2015,16(1): 1-7.

[14] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1-3): 389-422.

[15] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[16] 吴辰文,梁靖涵,王伟,等. 基于递归特征消除方法的随机森林算法[J]. 统计与决策,2017(21):60-63. (Wu C W, Liang J H, Wang W, et al. Random forest algorithm based on recursive feature elimination [J]. Statistics and Decision Making, 2017(21): 60-63.)

[17] 刘笑笑. 基于 RF-RFE 算法的森林生物量遥感特征选择方法研究[D]. 泰安: 山东农业大学,2016. (Liu X X. Research on forest biomass remote sensing feature selection based on RF-RFE algorithm[D]. Taian: Shandong Agricultural University,2016.)

[18] 魏小敏,徐彬,关佶红. 基于递归特征消除法的蛋白质能量热点预测[J]. 山东大学学报(工学版), 2014,44(2):12-20. (Wei X M, Xu B, Guan J H. Prediction of protein energy hotspots based on recursive feature elimination[J]. Journal of Shandong University (Engineering Science Edition), 2014, 44(2):12-20.)

[19] 董红斌, 石丽, 李涛. 一种改进的 microRNA 预测模型集成方法[J]. 计算机科学,2018,45(2): 69-75. (Dong H B, Shi L, Li T. An improved integrated method for microRNA prediction model [J]. Computer Science, 2008,45(2):69-75.)

[20] 林云光. 基于计算智能方法的 microRNA 预测[D]. 济南: 济南大学, 2013. (Lin Y G. MicroRNA prediction based on computational intelligence[D]. Jinan: Jinan University ,2013.)

[21] 张璇. 基于生物异构网络的疾病 microRNA 预测研究[D]. 厦门: 厦门大学,2017. (Zhang X. Prediction of disease microRNA based on biological heterogeneous network[D]. Xiamen: Xiamen University,2017.)