Journal of Computer Applications
计算机应用,2015,35(2):374 -377
ISSN 1001-9081
CODEN JYIIDU
2015-02-10
http://www.joca.cn

文章编号:1001-9081(2015)02-0374-04
doi:10.11772/j.issn.1001-9081.2015.02.0374

基于特征聚类和随机子空间的 microRNA 识别方法

芮志良[1*],朱玉全[1],耿 霞[1],陈 耿[2]
(1.江苏大学 计算机科学与通信工程学院,江苏 镇江 212013; 2.南京审计学院 工学院,南京 210029)
(*通信作者电子邮箱 ruizhiliang@126.com)

摘 要:针对 microRNA 识别方法中过多注重新特征、忽略弱分类能力特征和冗余特征,导致敏感性和特异性指标不佳或两者不平衡的问题,提出一种基于特征聚类和随机子空间的集成算法 CLUSTER-RS。该算法采用信息增益率剔除部分弱分类能力的特征后,利用信息熵度量特征之间相关性,对特征进行聚类,再从每个特征簇中随机选取等量特征组成特征集用于构建基分类器,最后将基分类器集成用于 microRNA 识别。通过调整参数、选择基分类器实现算法最优化后,在 microRNA 最新数据集上与经典方法 Triplet-SVM、miPred、MiPred、microPred 和 HuntMi 进行对比实验,结果显示 CLUSTER-RS 在识别中敏感性不及 microPred 但优于其他模型,特异性为六者最优,而且从整体性能指标准确性和马修兹系数可以看出,CLUSTER-RS 比其他算法具有优势。结果表明,CLUSTER-RS 取得了较好的识别效果,在敏感性和特异性上实现了很好的平衡,即在性能指标平衡方面优于对比方法。

关键词:microRNA 识别;分类能力;特征聚类;随机子空间;相关性
中图分类号:TP301.6 **文献标志码**:A

microRNA identification method based on feature clustering and random subspace

RUI Zhiliang[1*], ZHU Yuquan[1], GENG Xia[1], CHEN Geng[2]
(1. *School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*;
2. *School of Technology, Nanjing Audit University, Nanjing Jiangsu* 210029, *China*)

Abstract: As sensitivity and specificity of current microRNA identification methods are not ideal or imbalanced because of emphasizing new features but ignoring weak classification ability and redundancy of features. An ensemble algorithm based on feature clustering and random subspace method was proposed, named CLUSTER-RS. After eliminating some features with weak classification ability using information ratio, the algorithm utilized information entropy to measure feature relevance and grouped the features into clusters. Then it selected the same number of features randomly from each cluster to compose a feature set, which was used to train base classifiers for constituting the final identification model. By tuning parameter and selecting base classifiers to optimize the algorithm, experimental comparison of CLUSTER-RS and five classic microRNA identification methods (Triplet-SVM, miPred, MiPred, microPred, HuntMi) was conducted using latest microRNA dataset. CLUSTER-RS was only inferior to microPred in sensitivity and performed best in specificity, and also had advantage in accuracy and Matthew correlation coefficient. Experiments show that, CLUSTER-RS algorithm achieves good performance and is superior to the rivals in the aspect of balance between sensitivity and specificity.

Key words: microRNA identification; classification ability; feature clustering; random subspace; relevance

0 引言

microRNA(miRNA)是一类长度约为 23 nt(Nucleotide,核苷酸)的内源性非编码单链 RNA。miRNA 的识别和相应靶基因预测是研究 miRNA 生物学功能和作用机制的基础,精确地识别 miRNA 可以让生物研究者更好地分析基因调控网络[1]、理解转录后基因过程、指导药物研发等。通过计算识别方法对可能的 miRNA 序列进行识别,可以不受表达时间、表达水平、表达组织特异性的限制,很好地解决了传统实验方法投入高、周期长的问题。miRNA 计算识别方法的研究提高了识别效率,推动了 miRNA 研究更快地发展,在生物大数据背景下具有重大意义。

目前,对 miRNA 计算识别的研究已经取得了较好的发展,从头预测方法[2]得到广泛关注。一般过程是对 miRNA 前体(pre-miRNA)提取二级结构特征,结合序列特征[3],建立分类模型,从而对未知序列进行判别。Xue 等[4]分析 pre-miRNA 和伪发夹序列的局部连续子结构的分布时,发现了两者的显著区别,采用三联体局部结构-序列特征描述样本,然后使用支持向量机(Support Vector Machine, SVM)建立分类模型软件 Triplet-SVM,用于预测未知序列是否为 miRNA。Ng 等[5]分别从序列的核苷酸比例、热力学特性、拓扑结构和发卡折叠方面提取特征描述样本,同样用 SVM 构建了分类器

收稿日期:2014-09-11;**修回日期**:2014-11-06。 **基金项目**:国家自然科学基金资助项目(71271117);江苏省科技型企业技术创新资金资助项目(BC2012201);江苏省六大人才高峰项目(2013-WLW-005)。

作者简介:芮志良(1990 -),男,江苏南京人,硕士研究生,主要研究方向:数据挖掘、生物信息学; 朱玉全(1966 -),男,江苏常州人,教授,博士,主要研究方向:模式识别、数据挖掘、云计算; 耿霞(1978 -),女,山西汾阳人,讲师,博士研究生,主要研究方向:数据挖掘、生物信息学; 陈耿(1965 -),男,江苏无锡人,教授,博士,主要研究方向:数据挖掘。

miPred。Jiang 等[6]在文献[4]的基础上引入了二级结构折叠最小自由能(Minimum Free Energy,MFE)和随机化检验 p-值两个特征,并使用随机森林构建了集成分类器 MiPred。Batuwita 等[7]在文献[5]的特征集的基础上引入了 19 个新的特征来描述样本,新引入的特征包括结构熵、结构焓等热力学特征和碱基对距离、茎平均碱基配对数目以及自由能相关特征,然后采用过滤法选取了 21 个信息增益高的特征也采用 SVM 构建了分类器 microPred。Gudys 等[8]选择了文献[7]中使用的 21 个特征与文献[4]中信息增益最高的 4 个特征,同时引入其他 7 个特征组合作为特征集合,通过评估多个分类算法的性能后选择了敏感性与特异性最平衡的随机森林构建分类器 HuntMi。文献[4-6]中分类模型仅使用原始特征集而没有考虑特征分类能力;文献[7-8]虽然考虑选取分类能力强的特征作为特征集,但没有考虑特征之间的相关性可能导致的冗余问题。有用的信息太少或者无用的信息太多都会对识别模型的性能产生坏的影响,冗余信息的存在也会误导分类模型的建立,从而导致 miRNA 识别效果不好。为此,本文提出一种将特征聚类和随机子空间相结合的 miRNA 识别方法,该方法通过剔除部分特征后聚类,再从每个簇中随机选取特征的方式,解决了特征集中存在分类能力弱和特征冗余问题,使构建的分类模型具有较好的识别性能。在相同测试集上与五个经典方法进行对比,实验证明本文提出的方法具有一定的优势。

1　相关知识介绍

1.1　特征分类能力度量

信息增益率是衡量特征分类能力的经典指标。设数据集 D 为训练数据集,用 $C_i(i=1,-1)$ 分别表示真实 pre-miRNA 和伪发夹结构序列。令 D_P 表示正例样本集合,D_N 表示反例样本集合,并用 $|D|$、$|D_P|$、$|D_N|$ 分别表示样本总数、正例样本总数、反例样本总数。则数据集 D 的信息熵可用式(1)计算:

$$H(D) = -\sum_i P(C_i)\,\mathrm{lb}\,P(C_i) \tag{1}$$

其中:$P(C_i)$ 表示任意样本属于类别 C_i 的概率,若 $i=1$,则 $P(C_i)=|D_P|/|D|$,反之 $P(C_i)=|D_N|/|D|$。按特征 f 对数据集 D 进行划分,设特征 f 根据训练数据的观测有 n 个不同的值 $\{v_1,v_2,\cdots,v_n\}$,可用特征 f 将 D 划分为 n 个子集 $\{D_1,D_2,\cdots,D_n\}$,其中 $D_j(j\in\{1,2,\cdots,n\})$ 包含数据集 D 在特征 f 上取值为 v_j 的样本。则根据特征 f 划分样本的信息熵用式(2)计算;由此可以通过式(3)计算得到特征 f 的互信息,即信息增益。

$$H_f(D) = -\sum_{j=1}^{n}\frac{|D_j|}{|D|}\times H(D_j) \tag{2}$$

$$Gain(f) = H(D) - H_f(D) \tag{3}$$

由于信息增益存在偏向于选择取值较多的属性的问题,本文将采用信息增益率作为衡量属性分类能力强弱的指标,如式(4)所示。假设存在特征 A 和特征 B,若 $GainRatio(A)>GainRatio(B)$,则认为选用特征 A 的分类结果比 B 好,即 B 的分类能力较弱,因此倾向于剔除特征 B。

$$GainRatio(f) = \frac{Gain(f)}{-\sum_{j=1}^{n}|D_j|/|D|\,\mathrm{lb}(|D_j|/|D|)} \tag{4}$$

1.2　特征相关性度量

特征相关性度量分为两类:一类基于线性关联,如 Pearson 积矩相关系数、线性相关系数等;一类基于熵,如信息增益、不一致度等。本文使用基于熵的对称不确定性评价特征间的非线性相关度。用 $P(x_i)$ 表示特征 X 取第 i 个值的概率,$P(x_i|y_j)$ 表示特征 Y 取值为 y_j 时特征 X 取值为 x_i 的概率。X 的信息熵由式(2)计算;已知特征 Y 后 X 的条件信息熵 $H(X|Y)$ 的计算方法如式(5)所示;特征 X 与 Y 之间的互信息 $IG(X|Y)$ 如式(6)所示。

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\,\mathrm{lb}\,P(x_i|y_j) \tag{5}$$

$$IG(X|Y) = H(X) - H(X|Y) \tag{6}$$

定义对称不确定性 $SU(X,Y)$ 用于衡量特征 X、Y 之间的相关性,如式(7)所示。由此得到特征相关性度量 $SU(X|Y)\in[0,1]$,当取值为 0 时特征 X、Y 互相独立,当取值为 1 时代表特征 X、Y 完全相关。

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \tag{7}$$

2　基于特征聚类和随机子空间的 miRNA 识别

2.1　算法基本思路

miRNA 识别问题归根到底是二分类问题。一般情况下,集成分类方法的泛化能力高于其中的任何一个基分类器,而且基分类器的差别越大,集成后的效果越好。本文核心思想是构造差异性大且精确度高的基分类器,从而得到较好的集成分类模型。本文提出一种基于特征聚类和随机子空间的 miRNA 识别算法,该算法首先通过特征预处理得到候选特征集;再根据特征的相关性对候选特征集聚类,从每个簇中随机选取等量特征组成特征集用于构建基分类器;最后通过多数投票法判断未知序列是否为 miRNA。

本文初始特征集包含了文献[3-8]中所有特征,为防止弱分类能力特征过多而影响识别性能,在构建分类模型之前先剔除一部分分类能力弱的特征,同时能节省计算时间和存储空间开销。对于初始特征集合 $IF=\{f_1,f_2,\cdots,f_M\}$,计算各特征的信息增益率 $GR=\{r_1,r_2,\cdots,r_M\}$,设定信息增益率阈值 λ,剔除 $r_i<\lambda(i\in\{1,2,\cdots,M\})$ 的特征 f_i 得到候选特征集 $CF=\{f_i|\ GainRatio(f_i)\geqslant\lambda,\ i=1,2,\cdots,M\}$。

为了解决特征冗余的问题,本文提出利用特征之间相关性对候选特征集聚类与随机子空间相结合的集成分类方法。由 1.2 节可知,两个特征之间相关性越强,则对称不确定性 SU 越大,距离越小。设 CF 的大小为 N,定义特征之间的距离矩阵 $\boldsymbol{Dist}(f_i,f_j)=1-SU(f_i,f_j)$,其中 $i,j\in\{1,2,\cdots,N\}$。由于 $SU(f_i,f_j)=SU(f_j,f_i)$,因此该矩阵为对称矩阵,其定义如式(8)所示:

$$\boldsymbol{Dist}=\begin{bmatrix}0 & 1-SU(f_2,f_1) & \cdots & 1-SU(f_N,f_1)\\ 1-SU(f_2,f_1) & 0 & \cdots & 1-SU(f_N,f_2)\\ \vdots & \vdots & & \vdots\\ 1-SU(f_N,f_1) & 1-SU(f_N,f_2) & \cdots & 0\end{bmatrix} \tag{8}$$

对候选特征集 CF 使用凝聚的最短距离(Single-Linkage, SL)层次聚类算法,初始化为每个特征为一类,通过 ***Dist*** 逐步合并同时更新矩阵,直到满足任意两个簇内特征间最小距离大于阈值为止。令距离阈值 $d = 1 - (EX + r * SD)$,其中 EX 和 SD 分别为信息增益率最大的特征与其他特征之间对称不确定性的平均值和标准差,r 为调节参数。通过 r 调整距离阈值 d,得到 k 个内部相关性强的特征簇 $T_1, T_2, \cdots, T_k$。然后利用随机子空间方法从每个特征簇中随机选择 m 个特征组合作为特征集 $S = \bigcup_{i=1}^{k} Rand(T_i, m)$,其中 $Rand(T_i, m)$ 表示从集合 T_i 中随机选取 m 个元素。通过特征集 S 在原始数据上投影得到的数据集训练基分类器,最后使用分类器集成的方法预测给定样本所属的类别。这样组合而成的特征集不仅降低了冗余性而且具有代表性,随机性则保证了差异性,因此可以得到精确度高且差异性大的基分类器。

2.2 算法描述

综上所述,本文提出一种基于特征聚类的随机子空间集成算法 CLUSTER-RS,具体描述如算法 1 所示。

算法 1 CLUSTER-RS。

输入 训练集 D(正据集 D_{Pos},反例集 D_{Neg}),测试样本 X,分类学习算法 L,每个簇随机选取特征的个数 m,信息增益率阈值 λ,特征间距离阈值 d。

输出 待测样本的预测类别。

1) 采用 MDL(Minimum Description Length)方法对训练集数据离散化,结果以 D' 表示;
2) 采用式(1) ~ (4) 计算 D' 中各特征的信息增益率 r,剔除 $r < \lambda$ 的特征,得到候选特征集,记为 F;
3) 对特征集 F 使用式(2)、(5) ~ (8) 计算,得到特征距离矩阵 ***Dist***;
4) 根据 ***Dist*** 使用 SL 层次聚类算法对特征聚类,终止条件为任意两个簇内特征间最小距离 $> d$,得到 k 个特征簇,记为 F_i $(i \in \{1,2,\cdots,k\})$;
5) 划分 D_{Neg},每份数量为 $2^{-1} \times |D_{Pos}|$,分别与 D_{Pos} 中随机选取 $2^{-1} \times |D_{Pos}|$ 个正例组合,得到数据子集记为 D_p $(p = 1,2,\cdots, 2 \times |D_{Neg}| / |D_{Pos}|)$,设定基分类器数目为 $q = 2 \times |D_{Neg}| / |D_{Pos}| - 1$;
6) for $p = 1$ to q do
7) 初始化基分类器特征集 $f(p) = \varnothing$;
8) for $i = 1$ to k do
9) 从 F_i 中随机选取 m 个特征,加入 $f(p)$ 中;
10) end
11) 将 $f(p)$ 在 D_p 上进行投影得到子空间上的训练数据集记为 D_p'';
12) 由分类学习算法 L 和数据集 D_p'' 训练得到基分类器 C_p;
13) end
14) 初始化正反类别投票数 $Y_P = 0, Y_N = 0$;
15) for $p = 1$ to q do
16) 使用基分类器 C_p 预测 X 的类别,若为正类,则 $Y_P = Y_P + 1$,否则 $Y_N = Y_N + 1$;
17) end
18) 用多数投票法确定待测样本 X 的类别:若 $Y_P > Y_N$,预测为正类,否则预测为反类。

3 实验及分析

3.1 评价标准

本文使用生物信息学分类问题中常用的四个评价指标,分别为:敏感性(SEnsitivity, SE)、特异性(SPecificity, SP)、准确性(Accuracy, Acc)和马修兹系数(Matthew correlation coefficient, Mcc),如式(9) ~ (12)所示。其中:T_P 表示真正例个数,F_P 表示伪正例个数,T_N 表示真负例个数,F_N 表示伪负例个数。

$$SE = T_P/(T_P + F_N) \times 100\% \quad (9)$$

$$SP = T_N/(T_N + F_P) \times 100\% \quad (10)$$

$$Acc = (T_P + T_N)/(T_P + T_N + F_P + F_N) \times 100\% \quad (11)$$

$$Mcc = (T_P \times T_N - F_P \times F_N)[(T_P + F_P) \times (T_N + F_N) \times (T_P + F_N) \times (T_N + F_P)]^{-1/2} \times 100\% \quad (12)$$

3.2 数据集与特征集

本文训练集中的正例来自于 miRBase19[9] 中的人类 pre-miRNA 序列。为了防止过拟合,移除其中的冗余序列,对 1 600条序列使用 DNACLUST[10] 将序列聚类,然后从每个类中随机选择 1 条序列。实验中调整相似度参数并计算的时间开销和冗余情况,5 次实验平均结果如表 1 所示。考虑时间开销和去除冗余的效果,本文最终选择相似度参数为 80% 的实验结果获得 1 377 条非冗余的 pre-miRNA 作为最终的正例集。参考 UCSC refSeq 的基因注释,从蛋白质编码区没有选择性剪切事件的序列中选取得到 8 494 条与真实 pre-miRNA 相似的伪发夹结构序列作为训练集中的反例。此外,采用统一的标准,在反例集上采用了与正例集移除冗余序列相同的操作,发现聚类结果只有一个类,因此反例集中并不存在冗余序列。

表 1 不同相似度参数下 DNACLUST 的实验结果

相似度/%	时间开销/s	相似度/%	时间开销/s
99	3.2	80	54.2
90	12.6	70	192.4

本文以 miRBase20 和 21 两个版本中新增的人类 pre-miRNA 共 291 条序列作为测试集的正例,而以文献[3]通过反例样本选择方法得到的 1 446 条伪人类 pre-miRNA 序列作为测试集的反例。数据集详情如表 2 所示。

表 2 数据集详细信息

数据集	正例数	反例数
训练集	1 377	8 494
测试集	291	1 446

另外,本文从序列特征、结构特征、序列-结构特征三个方面,选取 149 个特征作为初始特征集合。其中:64 个来源于文献[3],32 个来源于文献[4],29 个来源于文献[5],2 个来源于文献[6],19 个来源于文献[7],3 个来源于文献[8]。

3.3 实验分析与讨论

首先调整信息增益率阈值选取合适数量的特征,本文设置为 0.01,得到信息增益率较大的 78 个特征。设置距离阈值为 0.04,得到 6 个簇,按簇的大小降序排序分别为 21,16,13,11,9,8。通过 3 种基分类器(SVM, C4.5, Naive Bayes)和 5 折交叉验证的方式对比分类性能选择最优分类器与 m 的取值,分别取 $m = 1,2,3,4,5,6,7,8$ 测试分类模型的性能,选择其中分类性能最好的基分类器和参数 m 构建最终分类模型。本文实验均在 Weka[11] 环境下进行,基分类器数量为 11。最后使用测试集分别在 Triplet-SVM、miPred、MiPred、microPred、HuntMi 与 CLUSTER-RS 上进行测试比对。通过实验测试基分类器和参数 m 取值对分类性能的影响,实验结果如表 3 所示。

表 3　不同基分类器与参数 m 下模型的性能

m	SE/%			SP/%		
	SVM	C4.5	Naive Bayes	SVM	C4.5	Naive Bayes
1	72.67	77.02	66.17	70.92	76.18	67.04
2	78.92	79.63	69.23	77.34	78.81	71.69
3	84.05	82.40	77.43	85.84	79.54	77.08
4	87.36	86.72	82.65	89.21	82.60	82.50
5	91.33	90.45	85.30	88.56	88.26	84.72
6	92.79	87.75	88.25	91.04	85.58	87.66
7	88.40	83.96	83.62	87.62	82.10	82.91
8	85.62	79.22	80.13	83.51	78.92	79.85

表 3 给出了在不同基分类器和不同参数 m 下在训练集上 5 折交叉验证的结果,由于用于训练的数据在每个基分类器上是类别平衡的,但是用于验证的 1/5 的数据是不平衡的,因此评价指标 Acc 近似等于 SP,而 Mcc 在这种情况下不能真实反映分类情况的[12],故表 3 只给出敏感性指标(SE)和特异性指标(SP)。从理论上说,m 越大则每个基分类器从每个特征簇中获得的信息越多,分类性能越好,但是从表中可以看出,当 m = 6(SVM、Naive Bayes)或者 m = 5(C4.5)时性能达到最优后有不同程度的下降,这是由于每个特征簇中的特征高度相关,m 值的增大得不到更多的信息,反而使基分类器的差异性变小,从而导致集成性能下降。实验结果表明在训练集上当 m = 6 且使用 SVM 作为基分类器时效果最佳。

文献[4]的作者在国际上最先开展 microRNA 发夹的真伪辨别工作,开发了基于 Perl 语言的经典程序 Triplet-SVM;文献[5]提出了很多新颖且有效的分类特征并实现了 miRNA 分类程序 miPred;而文献[6-8]均在二者的基础上作出了不同程度的改进。由于本文的研究正是基于上述研究成果,为验证本文方法的性能,使用全部训练集和 SVM 作为基分类器构建分类模型,并通过测试集在 CLUSTER-RS 上重复 10 次实验取平均值作为结果与在五个经典方法上的实验结果进行对比,如表 4 所示。

表 4　CLUSTER-RS 与五个经典方法在测试集上实验对比

算法	SE/%	SP/%	Acc/%	Mcc/%
Triplet-SVM	69.42	83.26	80.94	45.10
miPred	82.82	89.53	86.87	61.55
MiPred	85.57	82.43	82.96	55.99
microPred	91.07	79.11	81.12	55.89
HuntMi	89.35	91.08	90.79	72.05
CLUSTER-RS	89.69	91.56	91.25	73.19

从表 4 中可以看到,CLUSTER-RS 在 miRNA 识别中敏感性(SE)不及 microPred 但优于其他模型,特异性(SP)为六者最优,而且从整体性能指标 Acc 和 Mcc 可以看出,CLUSTER-RS 比其他算法具有优势,这说明本文提出的方法在敏感性和特异性上实现了很好的平衡。

CLUSTER-RS 之所以识别性能优异,主要有以下原因:1)通过特征预处理后保留了分类能力较强的特征,改善了基分类器精确性;2)聚类后从每个簇中选取的特征组成特征集在合理降低特征冗余度、减少开销的同时进一步提高基分类器的性能;3)具有随机子空间特性,因为在不同的聚类簇中随机选取特征,保证了基分类器的差异性;4)充分利用了反例,数据集的划分后组合也提高了基分类器的差异性。

4　结语

近年来,对 microRNA 的识别研究已成为当前生命科学和生物信息学领域最前沿的方向之一。对于目前计算识别性能不佳的现状,本文提出一种将特征聚类和随机子空间方法结合的识别算法。该算法考虑并解决了特征分类能力弱和特征冗余问题,通过在相同的测试集上与经典模型实验对比,结果表明本文方法在 microRNA 识别方面具有一定的优势。需要指出的是,本文方法涉及到一些需要人工设定的参数,如信息增益率阈值、特征间距离阈值等,在以后的工作中将研究这些参数选定的方法,使提出的方法更加具有通用性。

参考文献:

[1] SU N. Bioinformatical analysis of gene regulatory network consisting of transcription factor and microRNA [D]. Beijing: Peking University, 2013: 1-11. (苏乃芳. 转录因子和 microRNA 组成的基因调控网络的生物信息学分析[D]. 北京:北京大学, 2013: 1-11.)

[2] HU L, HUANG Y, WANG Q, *et al*. Benchmark comparison of ab initio microRNA identification methods and software [J]. Genetics and Molecular Research, 2012, 11(4): 4525-4538.

[3] WEI L, LIAO M, GAO Y, *et al*. Improved and promising identification of human microRNAs by incorporating a high-quality negative set [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014, 11(1): 192-201.

[4] 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.

[5] NG K L, MISHRA S K. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures [J]. Bioinformatics, 2007, 23(11): 1321-1330.

[6] 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(Web Server issue): W339-W344.

[7] BATUWITA R, PALADE V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction [J]. Bioinformatics, 2009, 25(8): 989-995.

[8] GUDYS A, SZCZESNIAK M W, SIKORA M, *et al*. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification [J]. BMC Bioinformatics, 2013, 14: 83.

[9] KOZOMARA A, GRIFFITHS-JONES S. miRBase: integrating microRNA annotation and deep-sequencing data [J]. Nucleic Acids Resarch, 2011, 39(Database issue): D152-D157.

[10] GHODSI M, LIU B, POP M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes [J]. BMC Bioinformatics, 2011, 12: 271.

[11] HALL M, FRANK E, HOLMES G, *et al*. The WEKA data mining software: an update [J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18.

[12] ZOU Q, GUO M, LIU Y, *et al*. A classification method for class-imbalanced data and its application on bioinformatics [J]. Journal of Computer Research and Development, 2010, 47(8): 1407-1414. (邹权, 郭茂祖, 刘扬, 等. 类别不平衡的分类方法及在生物信息学中的应用[J]. 计算机研究与发展, 2010, 47(8): 1407-1414.)