一种改进的 microRNA 预测模型集成方法

董红斌 石 丽 李 涛
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

摘 要 现有的 microRNA 预测方法往往存在数据集类不平衡和适用物种单一的问题。针对以上问题,所做主要工作如下:1)提出基于序列熵的分层采样算法,该算法可在保持样本总体分布的基础上,采样生成正样本和负样本数量平衡的训练集;2)提出基于信噪比和相关性的特征选择,用于缩小训练集规模,以达到提高训练速度的目的;3)提出 DS-GA 算法,用于缩短 SVM 分类器参数的优化时间,达到减少过拟合的目的;4)结合集成学习的思想,经采样、特征选择、分类器参数优化 3 个步骤,建立了一种物种间通用的 microRNA 预测模型。实验表明,该模型有效解决了类不平衡问题,且不局限于单一物种,对混合物种的测试集预测取得了较好效果。

关键词 microRNA,预测,采样,特征选择,类不平衡
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.012

Improved Ensemble Method on MicroRNA Prediction Model

DONG Hong-bin SHI Li LI Tao
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract The existing microRNA prediction methods often present the problems of imbalance data set class and single applicable species. In order to solve the above problems, the main work is as follows. Firstly, a hierarchical sampling algorithm based on sequence entropy was proposed, which can generate a training set enhancing balance positive and negative samples based on the overall distribution of the samples. Secondly, a feature selection algorithm based on signal-to-noise ratio and correlation was designed to reduce the scale of training set and achieve the purpose of improving training speed. Thirdly, the DS-GA was proposed to shorten the optimization time of SVM classifier parameters and avoid the over-fitting problem. At last, based on the idea of ensemble learning, a common microRNA prediction model was established by sampling, feature selection and classifier parameter optimization. Experiments show that the model solves the problem of imbalance effectively, it is not limited to a single species and achieves better results for the hybrid species test set prediction.

Keywords MicroRNA, Prediction, Sampling, Feature selection, Imbalance class

1 引言

MicroRNA 是一类长度约为 20～23 个核苷酸的单链非编码 RNA,具有高保守性和内源性,可在转录后水平调控基因表达[1]。研究表明,microRNA 可以调节约 50%的蛋白质编码[2],并参与多种生理过程和多种病理过程[3]。单个 microRNA 能够广泛调节生物学功能,如 DNA 修复、炎症和细胞凋亡等,这些都是肿瘤发生的基础[4]。MicroRNA 的表达谱已经进入癌症诊所作为诊断和预后的生物标志物,以评估癌症患者的肿瘤起始、进展和对复发的反应[5]。因此,利用机器学习手段发现新的 microRNA 成为了研究热点。

Lu 等人[6]使用最小自由能和 P 值等特征,设计了一种基于 RF 的方法,用于使用混合特征对 microRNA 前体及伪 microRNA 前体进行分类。Bentwich 等人[7]在热力学稳定性和结构特征的基础上提出了一个评优函数,该函数涵盖了 pre-microRNA 中所有发夹结构的特性,提出了结合微阵列分析和序列定向克隆的生物信息学预测方法。Ng 等人[8]采用高斯径向基函数作为相似度测量标准来判断 29 种固有的发夹折叠属性的相似性,并且基于它们的二核苷酸序列、发夹折叠特征和非线性的热力学以及拓扑学来识别 pre-microRNAs。Huang 等人对多年生黑麦草进行研究,第一次通过计算方法检测到 33 个潜在的 microRNA 标靶[9]。在 2016 年,他们评估了用于 pre-microRNAs 分类的特征的可用性,采用基于序列保守性和茎环结构的计算基因组同源性比较法,对代表 81 个家族的 199 个潜在的 microRNA 进行预测[10]。Zhao 等人[11]则进一步研究了 pre-microRNAs 的结构序列特征,指出 RNA 二级结构的最小自由能与双螺旋结构中碱基配对和不配对具有同样重要的意义。Huang 等人[12]提取了 26 种病毒

到稿日期:2017-05-10 返修日期:2017-06-28 本文受国家自然科学基金项目(61472095)资助。
董红斌(1963－),男,教授,博士生导师,CCF 会员,主要研究方向为演化计算、数据挖掘与机器学习,E-mail:donghongbin@hrbeu.edu.cn(通信作者);石 丽(1990－),女,硕士生,主要研究方向为智能信息处理;李 涛(1990－),男,博士生,主要研究方向为机器学习、智能信息处理。

的 microRNA 序列和二级结构的 54 个特征，建立支持向量机和随机森林的 ViralmiR 模型，其平均精度高于历史最高的 83%。Khalifa 等人[13]使用 8 种特征选择方法和 7 种不同的植物物种作为正样本，证明了特征选择对于进一步预测 microRNA 是至关重要的。

Yousef 等人基于序列特征，用数字特征描述 pre-microRNA，然后用机器学习的方法建立 microRNA 识别模型[14]，并利用特征选择方法构建了针对 microRNA 预测的一分类模型[15]。文献[16]利用带发夹的茎环结构对 microRNA 预测的重要影响，将识别 microRNA 成熟体转换成了识别 microRNA 前体，并设置多个相关特征进行计算和比较；然后通过特征选择建立了一种广义的 MotifmiRNAPred 模型，用于预测植物中未被发现的 microRNA。

Zhong 等人[17]通过 AdaBoost 算法，使用 SVM 分类器得到分类精度较高的 MirID 模型，并在 12 个物种数据集上进行了实验。结果表明，该模型在拟南芥、线虫、EB 病毒、果蝇等数据集中达到了 100%的准确率。Cevik 等人[18]针对不同物种间共享的 microRNA 分析问题（即 microRNA 具有相同的名称和功能，但序列不同，属于不同的物种）提出了 CCA（Canonical Correlation Analysis）方法，用于分析不同物种共享的 microRNA 的关系；2015 年，他们解决了 CCA 方法对噪声和异常值敏感的问题，根据 microRNA 预测时序列驱动特征的特点，Cevik 等提出了集成的 ECCA（Ensemble Canonical Correlation Analysis）方法，实验结果表明应用于一对物种的 ECCA 的冗余指数与它们的遗传距离具有相关性。可见，跨物种的 microRNA 预测具有一定的意义，通过建立通用模型，利用已知物种发现的 microRNA 即可预测其他物种的新 microRNA。

目前，大多研究将重点集中在针对 microRNA 数据集的特征选择方法和针对特定物种的分类模型的建立方面。但是，研究人员无法针对每个物种都建立这样的模型。因此，研究 microRNA 预测模型并建立具有一定通用性的模型，是本文研究的重点。

2 一种改进的 microRNA 预测模型集成方法

在预测新的 microRNA 的研究中，样本的选取往往不确定，且非 microRNA 样本的数量远大于 microRNA 样本的数量。样本序列能提取到较多的特征，动物和植物序列提取到的特征虽有明显区别，但也不乏共性。本文注重类不平衡问题，并关注新特征的发现，意在寻找更多能标识碱基序列数据集的特征；同时注重数据集规模约减，根据特征选择算法选择具有代表性的特征子集，以缩短模型中基分类器的训练时间并提高分类的准确率。microRNA 的预测模型集成方法的预测流程如图 1 所示。

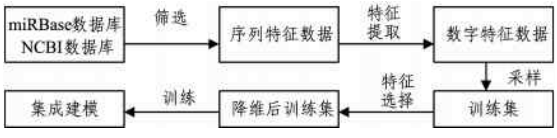

图 1 microRNA 预测流程图

Fig. 1 Prediction flow chart of microRNA

文中针对类不平衡问题而提出的采样算法对大量的数字特征样本进行采样，构成负样本集，但训练集的属性和样本数量依然很大，直接训练分类模型会使训练耗时过长且分类错误率高。鉴于此，提出基于信噪比和相关性的特征选择算法，利用经过特征选择处理后的训练集训练预测模型。

在训练预测模型过程中，选择支持向量机作为分类器。为防止分类器过拟合，同时提高准确率，提出基于 DS-GA（Deterministic Selection Genetic Algorithm）的支持向量机参数优化方法。该方法采用确定策略选择保留至下一代的个体，在下一轮对个体适应度值进行评估时，被确定选择的个体的适应度值沿用父代的适应度值。通过对以分类器正确率为适应度值的遗传算法的参数优化过程加以改进，缩短了算法的运行时间，达到了快速优化支持向量机的参数的目的。

本文依照建模流程，经 microRNA 的特征提取，使用本文提出的 3 个算法构建训练速度快、正确率高的支持向量机作为基分类器，采用集成学习的思想将若干这样的基分类器集成为本文模型。集成建模方法的具体步骤如下：

1）使用基于序列熵的分层采样算法生成若干训练集。本文用于训练模型的生物训练集包括动物类、植物类和病毒类。

2）针对步骤 1）产生的训练集，使用基于信噪比和相关性的特征选择算法。分析特征选择的结果，确定建模时使用的属性，并删除训练集中未被选择的属性。

3）针对步骤 2）处理后的训练集，使用基于 DS-GA 的 SVM 参数优化算法进行处理。将寻优后的参数设置为 SVM 的参数，并用该类型的训练集训练支持向量机。

4）依照步骤 3），可训练若干个基分类器，采用集成学习的思想建立本文的 microRNA 预测模型，将以少数服从多数的投票方式集成各个基分类器的分类结果作为本文模型最终的输出。

2.1 基于序列熵的分层采样算法

研究发现，生物序列数据集的序列亦具有熵的性质。生物进化理论认为，物种是可变的，且父代性状通过遗传传递给子代，而性状保持或者性状突变都是由基因决定的。遗传是一个不可逆过程，基因的遗传变异引发了遗传群体熵的变化，参与遗传过程的分子本身具有一定的熵。熵可被理解为遗传群体中性状状态的多样度或其概率分布的均匀度。各种碱基分子序列本身亦具有熵值，这在其采样构建数据集的过程中有着重要作用。microRNA 识别预测的本质是一个二分类问题，microRNA 的序列较非 microRNA 的序列少得多。采用本文提出的采样算法可构建高质量的负样本集，使分类器对非 microRNA 序列进行全面的学习，有助于预测过程发现新的 microRNA。通过式（1）定义序列熵：

$$SeqQ=\frac{-\sum_{i<j}P_{ij}\cdot\log_2 P_{ij}}{L}+\frac{dS}{L} \tag{1}$$

其中，P_{ij} 为碱基 i 与碱基 j 互补的概率；dS 为 UNAfold 软件计算的结构熵；L 为序列长度。

本文所提出的分层采样算法用下标 h 表示层号，关于第 h 层的记号如下：

1）单元总数 N_h；

2）样本单元数 n_h；

3）利用式(2)定义采样比：

$$f_h=\frac{n_h}{N_h} \tag{2}$$

具体算法如下：

输入：初始数据集 Set=$\{X_1,X_2,\cdots,X_n\}$，n 为自然数

输出：训练集 Train=$\{X_i,X_k,\cdots,X_l\}$，且 $X_i,X_k,X_l\in$ Set

Step 1　对 Set 的 class 属性列计数，统计每类的样本数目。假设 class 为两类，统计结果记为 P，N。

Step 2　对 Set 的 Q 属性列进行分层统计，统计结果记为 count。

Step 3　计算采样比，并根据采样比计算采样数，记为 num；若 N>P，则对于每个分层，num=count * f_h。

Step 4　按 num 数目随机采样，结果记为 Train。

Step 5　计算总体方差和样本方差，若后者过大，则转入 Step 3；否则，停止。

假设待采样的非 microRNA 样本有 M 个，microRNA 样本有 N 个。将样本按熵值分成 12 层，每层的采样比例为 N/M。经采样后，非 microRNA 的数目为 N。算法不仅通过采样缩小了非 microRNA 数据集的规模，而且使 microRNA 样本和非 microRNA 样本的数目达到平衡。但是，由于计算中的舍入会造成一定的误差，有的数据集在采样后的数目可能小于 N。定义采样调整规则为：按比例在熵值分布最多的 3 个分组中进行采样，采样数目为 $M-N$。

2.2　基于信噪比和相关性的特征选择算法

特征选择是高维数据分类中的关键环节，将采样后的数据集构成训练集，虽然类别间的样本数目达到了平衡，但训练分类器非常耗时。鉴于此，在建立 miroRNA 预测模型的过程中增加属性筛选环节。

信噪比方法可以去噪声，它关注特征对分类的重要程度，进而选出优秀的特征代表[20]。

信噪比的计算如式(3)所示：

$$SNR(a_i)=\frac{|\mu_+(a_i)-\mu_-(a_i)|}{\delta_+(a_i)+\delta_-(a_i)} \tag{3}$$

其中，$\mu_+(a_i)$为第 i 个属性 a_i 在正类的均值；$\mu_-(a_i)$为第 i 个属性 a_i 在负类的均值；$\delta_+(a_i)$为第 i 个属性 a_i 在正类中的标准差；$\delta_-(a_i)$为第 i 个属性 a_i 在负类中的标准差。

虽然通过信噪比可以筛选出与类别相关程度大的属性，但是该步骤没有考虑特征之间的相关性。图 2 给出了样本中冗余属性值的一个示例。图 2 中 A==U 表示碱基 A 与碱基 U 配对的数量，AU%表示碱基 A 与碱基 U 相邻出现的次数。可以看出，这两个属性存在明显的线性关系，这种现象在经信噪比筛选后的集合中不是特例。由此，本文将引入能过滤掉冗余属性的方法。

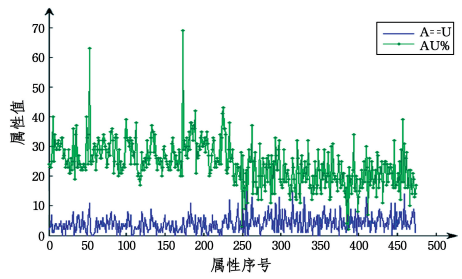

图 2　示例：样本中的冗余属性值曲线

Fig. 2　Example: redundancy attribute value curve in sample

使用皮尔逊相关系数(Pearson Correlation Coefficient)来衡量属性间的相似度。皮尔逊相关系数的显著缺点是对除了线性关系外的其他关系均不敏感。当 P_{XY} 大于 0.6 时，说明属性 X 和属性 Y 属于强相关范畴，也是本文重点考虑的范畴。属性 X 和属性 Y 的皮尔逊相关系数的计算公式如下：

$$P_{XY}=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2}\sqrt{\sum_i(y_i-\bar{y})^2}} \tag{4}$$

其中，x_i 为 X 的第 i 个属性值；y_i 为 Y 的第 i 个属性；$\bar{x}$为 X 的均值；$\bar{y}$为 Y 的均值。

本文研究的数据集属于高维大样本数据集，若采用传统的特征选择方法，运算时间过长，所选择的特征子集的分类精度不能满足需求。针对此问题，本节提出一种适用于高维数据集的 filter 型特征选择方法。根据信噪比特征对分类的相关性进行排序，然后考虑属性间的相关性，进行特征选择。

具体算法如下：

输入：数据集合 S=$\{X_1,X_2,X_3,\cdots,X_n\}$，n 为自然数；其属性集合 A=$\{A_1,A_2,\cdots,A_n\}$，n 为自然数

输出：属性子集 Attribute

Step 1　对集合中的各属性进行标准化。

Step 2　根据式(3)计算 A_i 的 SNR_i，$A_i\in$ A。

Step 3　对 SNR 进行降序排列，取 SNR 较大的前 λ 个属性，记为 tempSet=$\{A_i,A_j,\cdots,A_k\}$，tempSet⊆SetA。

Step 4　根据式(4)计算 x 与 y 的 P_{xy}，x，y∈tempSet，结果记为相似矩阵 G。

Step 5　初始化集合 Attribute=∅。

Step 6　随机选择 x∈tempSet，将 x 加入集合 Attribute 中并将 x 从 tempSet 中删除。

Step 7　随机选择 y∈tempSet，评价 y 与集合 Attribute 中的属性的相似度。若存在 G_{xy} 大于 0.6，则将 y 从 tempSet 中删除；否则，将 y 加入 Attribute 中并将 y 从 tempSet 中删除。

Step 8　若 tempSet≠∅，则转入 Step 6，否则转入 Step 9。

Step 9　停止。

假设属性集 *SetA* 中有 N 个属性，样本数目为 M，则计算信噪比并对其进行排序的时间复杂度为 $O(N*M)+O(N\log N)$。该算法首先筛选出信噪比大的前 λ 个属性作为待搜索的特征子集，并采用随机搜索相关性小的属性构成最终的特征子集。前者选择信噪比大的属性，缩小了搜索规模，从而减小了时间复杂度。后者相关性矩阵计算的时间复杂度为 $O(\lambda^2)$。随机搜索的时间复杂度为 O(1)。经过该过程，可以明确哪些特征在模型的预测过程中对识别起关键作用。未来，对同类生物信息学数据集进行特征提取时，即可减少提取特征的数目，缩短此步骤花费的时间。

2.3　基于 DS-GA 的 SVM 优化算法

在对属性维数高、样本规模小的数据集进行分类的问题中，支持向量机分类器表现出许多特有的优势；但其参数的选择影响着支持向量机的分类精度和泛化能力，未经参数优化的模型的泛化能力往往较差。使用支持向量机作为分类器进行建模时，参数选择往往凭借研究人员的经验，或利用进化算法、网格搜索方法进行最优参数的选择。遗传算法亦是使用频率较高的方法，它具有随机性及全局搜索能力。

本文对遗传算法实行确定选择策略，提出了基于 DS-GA (Deterministic Selection-Genetic Algorithm)的 SVM 参数优化方法。其本质仍是基于生物进化的思想搜索一定范围内的个体，并对其实行优胜劣汰的策略。其核心是采用确定选择(Deterministic Selection)机制，即从父代种群中选择一定量的个体直接进入子代，并省略这一部分在子代中的适应度值评估步骤。本文定义确定选择比例 β，若 $\beta=0$，则退化为随机选择的遗传算法；若 $\beta=1$，则全部遗传至下一代，失去了搜索的能力。因此，β 的取值范围为 $\beta\in(0,1)$。

按 β 值首先对种群中的个体进行选择操作，复制选出部分的个体信息进入下一代。这时，父代种群个体无变化。然后对父代种群个体进行交叉和变异操作，随机选择 $P(t)*(1-\beta)$ 个个体进入下一代，并计算其适应度值。子代个体由上述两部分组成，当再次进行遗传操作时，来自父代的个体适应度值继承自上一代。

定义 DS-GA 的编码方法如下：本文将 SVM 的核函数参数和错误惩罚因子编码成为染色体，使用 32 位二进制编码。种群大小为 40，惩罚参数 $c\in[2^{-8},2^{8}]$，核函数参数 $\delta\in[2^{-8},2^{8}]$。适应度值取决于个体解码后，采用该参数的支持向量机经十折交叉验证后得出的分类正确率。迭代次数达到 100 时，算法终止。

本文采用的 DS-GA 将确定选择比 β 定为 0.5，将父代个体按适应度值进行排序，将适应度值靠前的 20 个个体作为子代的一部分。在父代种群的个体中，采用单点交叉和变异操作产生新个体，其中交叉概率为 0.9，变异概率为 0.1。随机选择 50%的新个体作为子代种群的另一部分。生成新一代种群后，种群中一部分个体的适应度值是可以通过父代确定的。流程如图 3 所示。

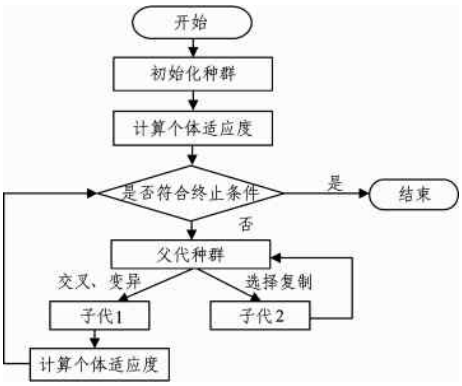

图 3 DS-GA 的流程图

Fig. 3 Flow chart of DS-GA

3 实验结果与分析

3.1 实验环境

本文主要使用的编程环境是 Matlab 和 Eclipse，操作系统为 Windows 7；所使用的电脑配置为 Intel Core(TM)2 Quad CPU，主频 2.66 GHz，内存 4 GB。在数据处理的过程中，使用 ViennaRNA Package 2.0，UNAfold 等软件预测序列的二级结构，计算最小自由能、热力学特性等属性。另外，还用到了 weka 和 libsvm，所涉及的参数均为 weka 默认参数。

3.2 数据集

实验选择的数据集分为两部分，一部分是生物信息数据库的碱基序列数据，来自 miRBase(http://www.mirbase.org/)和 NCBI(https://www.ncbi.nlm.nih.gov/pubmed/)，如表 1 所列；另一部分是公共的机器学习数据集，来自 UCI (http://archive.ics.uci.edu/ml/)和 mldata(http:// mldata.org/ repository/data)，如表 2 所列。

表 1 数据集 I

Table 1 Data set I

序号	名称	正样本数量	负样本数量
1	animal	7054	207393
2	arabidopsis	232	29253
3	human	1406	86324
4	plant	2183	90134
5	virus	237	839

表 2 数据集 II

Table 2 Data set II

序号	名称	属性数量	正样本数量	负样本数量
1	DLBCL	7130	58	19
2	Lung	7130	87	10
3	Ovarian	15155	162	91
4	Prostate	12601	52	50
5	Sonar	60	97	111

生物组训练集的正样本来自 miRBase 数据库的第 19 版。测试集的正样本由该数据库的第 20 版和第 21 版新增样本组成，包含人、家蚕、秀丽隐杆线虫、猕猴、大鼠、拟南芥、水稻、大豆等共 31500 条样本；负样本选择长度与 microRNA 长度绝对值之差小于 100nt、归一化后最小自由能小于 0.05、茎环结构配对高于 0.15 的非 microRNA 序列。选择 6 个公共数据集对本文所提算法的效果进行评估。其中，5 个为 2 分类数据集，4 个是具有高维小样本特点的集合，1 个是声呐信号数据集；还有 1 个是 UCI 数据库中的 wine 数据集。

在基于机器学习的 microRNA 的预测识别过程中，首要问题是通过特征提取将样本的碱基序列信息进行特征量化，以将其转化为可以被分类器学习的数字特征。使用 RNAfold 来预测所有的 RNA 序列的二级结构。本文提取了 191 个特征作为初始特征集。其中包括：序列特征(包括 C+G 含量、序列长度、碱基配对数和 microRNA 序列中的三碱基，如 UCC 和 GAU 等)的出现次数占序列总长度的比例；32 个三联体方式表示的序列-结构特征，分别由 3 个连续碱基的二级结构符号(共 $2^3=8$ 种)和与其对应的一级结构中的第一个碱基(A，U，G，C 4 种中的一种)组成；由二级结构预测得到的结构特征，如茎环数、最小自由能(FREE_ENERGY)、序列的二级结构熵、结构焓以及热力学特征等。

3.3 实验设计及结果分析

首先，针对上述 3 种算法进行有效性验证实验和对比实验；其次，为验证本文提出的建模方法的有效性，设计了 microRNA 预测模型分类能力的实验及本文模型与其他 3 种经典模型的对比实验。

实验正确率 ACC 的计算公式如下：

$$ACC=\frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

其中，TP 为正样本中被正确预测的样本数；FP 为负样本中被错误预测的样本数；TN 为负样本中被正确预测的样本数；FN 为正样本中被错误预测的样本数。

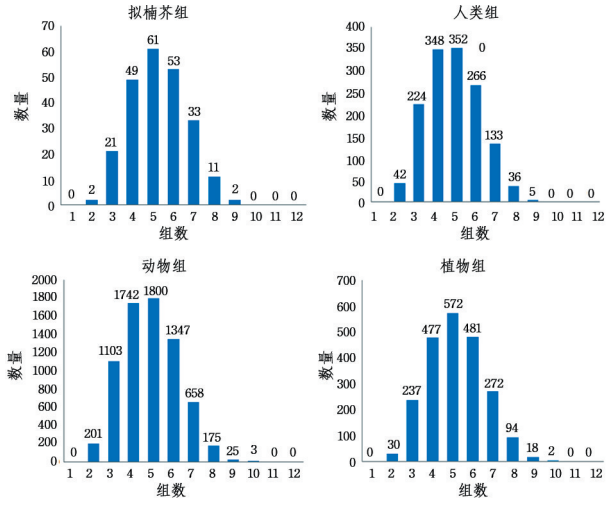

图 4　各组数据集的各层采样数

Fig. 4　Sampling number of each layer of each dataset

利用基于序列熵的分层采样算法生成平衡训练集。其中，按照碱基序列的序列熵值分布范围将其分组，这里的组即为层的具体化。序列熵处于 0 到 1 之间的组，其分类间隔为 0.1。鉴于序列熵值小于 0 和大于 1 的样本寥寥无几，将它们划分为 2 个组。因此，本文将其共划分为 12 组，范围分别为 $(-\infty,0]$，$(0,0.1]$，$(0.1,0.2]$，$(0.2,0.3]$，$(0.3,0.4]$，$(0.4,0.5]$，$(0.5,0.6]$，$(0.6,0.7]$，$(0.7,0.8]$，$(0.9,1.0]$，$(1.0,+\infty)$。对拟南芥组、人类组、动物组和植物组数据进行采样后，各层的采样数如图 4 所示。

使用该采样算法产生训练集，最终获得了动物组、植物组、拟南芥组、人类组各 50 个训练集及病毒组的 10 个训练集。这些训练集均为类别数目平衡的训练集。

为了验证基于信噪比和相关性的特征选择算法的有效性，针对动物组、拟南芥组、人类组、植物组、病毒组分别进行实验，该特征选择算法分别选择了 27，25，27，27，26 个特征作为特征子集。比较 5 组数据集使用基于信噪比和相关性的特征选择算法和未使用特征选择算法的 SVM，Voted Perceptron，Naive Bayes，ZeroR 分类器的十折交叉验证正确率，如表 3 所列。其中，5 组数据集的特征选择算法所选择的特征如表 4 所列。可以看出，mfe2，mfe4，A＝U，G＝U 等 16 个特征在 5 组数据集中均被筛选出；序列-结构特征只有动物组、拟南芥组和病毒组被选出。在这些特征子集中，热力学特征、茎环数量、碱基配对等特征对分类的影响较大。

表 3　不同分类器的分类正确率/%

Table 3　Lassification accuracy of different classifiers/%

Dataset	Classification Accuracy(Unselected/selected)							
	SVM		Naive Bayes		Voted-Perceptron		ZeroR	
animal	78.64	94.87	70.88	85.05	92.27	92.30	49.97	49.97
arabidopsis	55.60	88.57	82.53	88.75	86.20	94.82	49.57	49.57
human	71.69	83.93	88.90	92.27	89.18	91.64	49.86	49.86
plant	63.39	82.29	73.29	83.45	84.83	88.75	49.93	49.93
virus	80.17	84.18	98.94	99.36	59.07	71.42	49.36	49.36

表 4　5 组训练集的特征选择结果

Table 4　Feature selection results of five training sets

序号	名称	属性数量	属性名
1	animal	27	C…，U…，mfe1，mfe4，mfe2，mfe3，A＝U，G＝U，gc＊1.0/len，gu＊1.0/len，gc＊100/len，gu＊100/len，bpStems，dH，dS，dHL，dSL，TmL，efe，Diversity，EAFE，Diff，tri1，tri2，tri3，tri4，loops
2	arabidopsis	25	U(.(，mfe1，mfe4，mfe2，A＝U，G＝C，gc＊100/tot，gc＊1.0/len，gu＊1.0/len，gc＊100/len，gu＊100/len，bpStems，auStems，Tm，dHL，dSL，efe，Diversity，EAFE，Diff，tri1，tri2，tri3，tri4，loops
3	human	27	mfe1，mfe4，mfe2，mfe3，A＝U，G＝U，gc＊1.0/len，gu＊1.0/len，gc＊100/len，gu＊100/len，bpStems，gcStems，dH，dS，dHL，dSL，TmL，efe，freq，Diversity，EAFE，Diff，tri1，tri2，tri3，tri4，loops
4	plant	27	mfe1，mfe4，mfe2，mfe3，A＝U，G＝C，G＝U，gc＊1.0/len，gu＊1.0/len，gc＊100/len，gu＊100/len，bpStems，dH，dS，Tm，dHL，dSL，efe，freq，Diversity，EAFE，Diff，tri1，tri2，tri3，tri4，loops
5	virus	26	C(.(，C…，G(((，mfe1，mfe4，mfe2，mfe3，A＝U，G＝U，gu＊1.0/len，gu＊100/len，bpStems，dH，dS，dHL，dSL，TmL，efe，freq，Diversity，EAFE，Diff，tri2，tri3，tri4，loops

表 5　SVM 的分类正确率/%

Table 5　Classification accuracy rate of SVM/%

Dataset	Relief	Cfs Subset	Consistency Subset	本文算法
DLBCL	75.32	75.32	75.32	77.92
Lung	89.58	89.58	89.58	89.58
Ovarian	65.61	100	99.60	96.04
Prostate	50.98	89.58	68.62	73.29
Sonar	81.25	77.88	/	78.36

本文设计的对比实验选择了 5 组公共数据集，用到的特征选择方法有 ReliefF，Cfs Subset，Consistency Subset，选择的分类器有 SVM，Naive Bayes，J48。使用 SVM 作为分类器，十折交叉验证的分类正确率如表 5 所列。在该分类器上，本文算法在 Ovarian 数据集上的正确率较其他数据集高，但是在同类算法中，该结果只比 ReliefF 算法的分类效果好。

使用 Naive Bayes 作为分类器，进行十折交叉验证的分类

正确率如表 6 所列。其中,除 Lung 数据集外,使用 Relief 算法的特征选择的分类正确率偏低。针对 Ovarian 数据集和 Lung 数据集,Cfs Subset 特征选择后的分类正确率达到了 100%。

表 6 Naive Bayes 的分类正确率/%

Table 6 Classification accuracy rate of Naive Bayes/%

Dataset	Relief	Cfs Subset	Consistency Subset	本文算法
DLBCL	85.71	96.10	93.50	90.90
Lung	96.91	100	96.87	98.95
Ovarian	60.47	100	99.20	96.04
Prostate	90.19	100	84.31	93.13
Sonar	67.30	67.78	/	78.36

使用 J48 作为分类器,进行十折交叉验证的分类正确率如表 7 所列。

表 7 J48 的分类正确率/%

Table 7 Classification accuracy rate of J48/%

Dataset	Relief	Cfs Subset	Consistency Subset	本文算法
DLBCL	72.72	79.22	90.90	88.31
Lung	89.58	98.95	98.95	98.95
Ovarian	62.05	96.04	98.81	84.18
Prostate	93.13	98.95	95.10	93.13
Sonar	76.92	78.36	/	81.73

SVM 分类器在特征选择后的表现良好,考虑使用 SVM 作为 microRNA 模型训练的基分类器。但是,SVM 进行分类预测时需要对参数进行调节。本文使用 DS-GA 对 SVM 进行参数优化。

在 wine 数据集上,比较 PSO 优化、遗传算法优化、基于 DS-GA 优化这 3 种方法的效果,如表 8 所列。

表 8 wine 数据集上不同参数优化方法的对比

Table 8 Comparison of different parameters optimization method on wine dataset

	粒子群算法	遗传算法	DS-GA
运行时间/s	45.08	21.39	11.26
交叉验证正确率/%	82.0225	98.8764	98.9764
测试集正确率/%	80.8989	97.7528	98.9764
c	49.648	93.0048	80.2035
δ	13.284	4.5484	3.051

为了验证 DS-GA 优化算法在生物组数据集上的有效性,分别用特征选择后的动物组、植物组等 5 类数据作为 SVM 的训练集,使用 DS-GA 为 SVM 寻找较优的参数。以动物组为例,种群大小设定为 40,所得适应度曲线如图 5 所示。

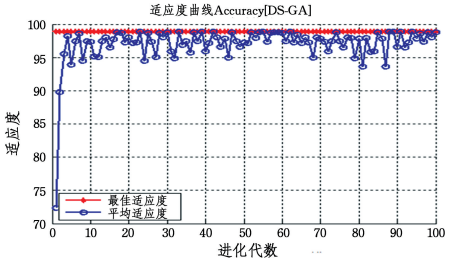

图 5 DS-GA 适应度曲线

Fig. 5 Fitness curve of DS-GA algorithm

图 5 中种群进化 100 代后,选择参数 c 为 83.0618,选择参数 δ 为 3.875,这时的交叉验证正确率为 98.8764%。不同方法搜索到的参数值及各指标的对比情况如表 8 所列。可以看出,PSO 算法的运行时间最长,本文算法约为遗传算法运行时间的一半;同时,本文算法得出的交叉验证的正确率和测试集的正确率均是最高的。

针对生物组数据集,得到一组 SVM 参数及分类正确率,如表 9 所列。实验结果表明,经 DS-GA 优化后,样本集规模小的数据集的十折交叉验证分类正确率较高。与表 3 中未经参数优化的分类正确率相比,各数据集的交叉验证分类正确率均有明显提高。

表 9 SVM 最佳参数及分类正确率

Table 9 Optimal parameters and classification accuracy of SVM

序号	名称	c	g	Accuracy/%
1	animal	40.9325	4.0212	95.6903
2	arabidopsis	9.4223	3.1386	99.7840
3	human	0.2500	0.0625	99.3243
4	plant	1.0000	0.1896	100
5	virus	0.069332	0.24681	100

对 microRNA 预测进行建模时,如果按物种分类进行模型训练,如人类数据集、拟南芥数据集等,它们所包含的正样本较少;同时,地球上的物种数以万亿计,这样为每个物种建模预测的工作量不亚于直接进行生物实验方法的工作量;还有像病毒类这样的样本,更是无法按病毒种类划分数据集。因此,本文意在构建一个能针对大类别(如动物、植物)的 microRNA 预测模型。

本文研究的 microRNA 数据和非 microRNA 数据数量差距悬殊,不平衡性大,直接训练分类器的分类效果差。因此,针对传统集成 SVM 分类 microRNA 模型做了相应的改进。之前统一的方法是人为选择使数据集平衡或用多个这样的数据集进行集成学习,达到解决数据不平衡的目的,但是往往数据集规模较小或会丢失较多的负样本集信息。本文提出基于序列熵的分层采样算法,为模型提供可靠的训练集,并进行基于信噪比和相关性的特征选择操作,减少数据集维度,减轻训练集负担。本文还对所选分类器 SVM 进行了参数优化操作,使核函数功能在建模过程中得到最大化发挥。最后,对若干个这样的分类器的分类结果进行集成,采用少数服从多数的投票机制将结果作为本文 microRNA 预测模型的最终分类正确率。

本文用于建模的训练集属性采用 29 个特征(U(.(,C…,U…,mfe1,mfe4,mfe3,mfe3,A=U,G=U,gu * 100/len,bpStems,dH,dS,dHL,dSL,TmL,efe,Diversity,EAFE,Diff,tri1,tri2,tri3,tri4,loops,Tm,C(.(,G(((,freq)选择经上述算法训练的 5 个基分类器,其中 2 个来自植物样本训练,2 个来自动物样本训练,1 个来自病毒样本训练。采用集成学习思想,将基分类器的分类结果以投票的方式进行整合,从而得到最终的 microRNA 预测模型,使用相同的测试数据集在 MicroPred,HuntMi,MiPred 和本文模型上进行多次实验,不同模型的对比及得到的分类正确率如表 10 所列。其中,除了 MiPred 模型外,均使用 SVM 分类器。microPred 和 MiPred 模型均是针对人类数据集提出的模型,从实验结果可看出其

不具有通用性。而 HuntMi 模型虽然能针对大类物种建模，但当物种跨度大，如使用病毒类数据训练建模，却用植物类数据进行测试时，模型分类正确率较低。HuntMi 模型具有一定的通用性。本文模型使用了混合物种的测试集，表现最佳。

表 10　不同预测模型的对比

Table 10　Comparison of different prediction models

	microPred	HuntMi	MiPred	本文模型
训练集	human	Animal/plant/virus	human	animal、plant、virus
初始特征数	48	34	35	191
分类器	SVM	SVM	RF	SVM
通用性	无	限定范围物种	无	有
正确率/%	88.13	94.80/91.16/69.83	88.90	95.86

结束语　本文针对 microRNA 预测方法存在数据集类不平衡和适用物种单一的问题，研究了生物遗传过程中分子具有熵的性质，并将其引入 microRNA 预测任务中，经采样、特征选择、SVM 参数优化等步骤建立了物种间通用的 microRNA预测模型。基于序列熵的分层采样算法，能够解决 microRNA 数据集的类不平衡问题。实验表明，该算法可在保持样本总体分布的基础上，采样生成正样本和负样本数量平衡的数据集。基于信噪比和相关性的特征选择算法，能缩小训练集规模、提高训练速度。在不同规模数据集上的实验验证了所提算法的有效性。基于 DS-GA 的 SVM 参数优化算法，能缩短 SVM 分类器参数的优化时间。实验表明，DS-GA 在不损失正确率的前提下，达到了提高训练速度的目的。

结合所提算法，采用集成学习思想建立了通用的 microRNA 预测模型。通过混合物种的测试机对模型的分类能力进行了检验，并与经典的 microRNA 预测算法或模型进行了比较。实验表明，本文模型不局限于单物种，对混合物种的测试集预测能取得较好的效果。

参 考 文 献

[1] ERSON-BENSAN A E. Introduction to microRNAs in biological systems[J]. Methods in Molecular Biology, 2014, 1107(1107): 1.

[2] SAÇAR M D, ALLMER J. Current limitations for computational analysis of microRNA in cancer[J]. Pakistan Journal of Clinical and Biomedical Research, 2013, 1(2): 3-5.

[3] 刘长征，余佳. microRNA 鉴定与功能分析技术[M]. 北京：化学工业出版社.

[4] HATA A, KASHIMA R. Dysregulation of microRNA biogenesis machinery in cancer[J]. Critical Reviews in Biochemistry and Molecular Biology, 2016, 51(3): 1-14.

[5] REDDY K B. MicroRNA (miRNA) in cancer[J]. Cancer Cell International, 2015, 15(1): 1-6.

[6] JIANG P, WU H, LU Z H, 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): 339-444.

[7] BENTWICH I, AVNIEL A, KAROV Y, et al. Identification of hundreds of conserved and nonconserved human microRNAs [J]. Nature Genetics, 2005, 37(7): 766-870.

[8] 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.

[9] HUANG Y, ZOU Q, SUN X H, et al. Computational identification of microRNAs and their targets in perennial Ryegrass (Lolium perenne)[J]. Applied Biochemistry and Biotechnology, 2014, 173(4): 1011-1122.

[10] HUANG Y, CHENG J H, LUO F N, et al. Genome-wide identification and characterization of microRNA genes and their targets in large yellow croaker (Larimichthys crocea)[J]. Gene, 2016, 576(1): 261-267.

[11] ZHAO D, WANG Y, LUO D, et al. PMirP: a pre-microRNA prediction method based on structure-sequence hybrid features [J]. Artificial Intelligence in Medicine, 2010, 49(2): 127-132.

[12] 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.

[13] KHALIFA W, YOUSEF M, DEMIRCI M D S, et al. The impact of feature selection on one and two-class classification performance for plant microRNAs[J]. Peerj, 2016, 4(2): e2135.

[14] YOUSEF M, ALLMER J, KHALIFA W. Sequence Motif-Based One-Class Classifiers Can Achieve Comparable Accuracy to Two-Class Learners for Plant microRNA Detection[J]. Journal of Biomedical Science & Engineering, 2015, 8(10): 684-694.

[15] YOUSEF M, ALLMER J, KHALIFA W. Feature Selection for MicroRNA Target Prediction-Comparison of One-Class Feature Selection Methodologies [C] // International Conference on Bioinformatics Models, Methods and Algorithms. 2016.

[16] YOUSEF M, ALLMER J, KHALIFA W. Accurate Plant MicroRNA Prediction Can Be Achieved Using Sequence Motif Features[J]. Journal of Intelligent Learning Systems & Applications, 2016, 8(1): 9-22.

[17] ZHONG L, WANG J T L, WEN D, et al. Effective Classification of MicroRNA Precursors Using Feature Mining and AdaBoost Algorithms[J]. Omics A Journal of Integrative Biology, 2013, 17(9): 486-493.

[18] CEVIK N, SAKAR C O, KURSUN O. Analysis of Relations Between Shared miRNAs of Different Species Using Canonical Correlation Analysis[C] // International Conference on Applied Informatics and Health and Life Sciences. 2013: 1980-1989.

[19] CEVIK N, SAKAR C O, KURSUN O. Analysis of shared miRNAs of different species using ensemble CCA and genetic distance[J]. Computers in Biology & Medicine, 2015, 64: 261-267.

[20] XU J C, LI T, SUN L, et al. Feature selection method based on signal-to-noise ratio and neighborhood rough set [J]. Journal of Data Acquisition and Processing, 2015, 30(5): 973-981. (in Chinese)
徐久成，李涛，孙林，等. 基于信噪比与邻域粗糙集的特征基选择方法[J]. 数据采集与处理，2015，30(5)：973-981.