ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2012/06(01)-0067-11
DOI: 10.3778/j.issn.1673-9418.2012.01.005

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-51616056

植物抗性基因识别中的随机森林分类方法*

郭颖婕[1], 刘晓燕[1], 郭茂祖[1+], 邹 权[2]

1. 哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001

2. 厦门大学 信息科学与技术学院, 福建 厦门 361005

Identification of Plant Resistance Gene with Random Forest*

GUO Yingjie[1], LIU Xiaoyan[1], GUO Maozu[1+], ZOU Quan[2]

1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

2. School of Information Science and Technology, Xiamen University, Xiamen, Fujian 361005, China

+ Corresponding author: E-mail: maozuguo@hit.edu.cn

GUO Yingjie, LIU Xiaoyan, GUO Maozu, et al. Identification of plant resistance gene with random forest. Journal of Frontiers of Computer Science and Technology, 2012, 6(1): 67–77.

Abstract: The traditional homology sequence alignment based approaches usually have high false positive rate and consequently new resistance genes are difficult to be identified. This paper presents a resistance gene identification approach by applying random forest classifier and *K*-Means under-sampling method. In order to solve the aimless problem in gene-mining research, two main contributions are provided. Firstly, it introduces random forest and 188 dimension features to identify resistance genes, accordingly the sample statistic learning approach can efficiently capture the internal characteristic of resistance genes. Secondly, it selects a more representative training subset and reduces the identification errors for solving the serious imbalanced classification during the training process. The experimental results indicate that the approach can efficiently identify the resistance genes, not only precisely classifying the existing experimental verified data, but also obtaining high accuracy on the negative sample dataset.

Key words: random forest; classifier; resistance gene; cluster; under-sampling

---

*The National Natural Science Foundation of China under Grant Nos. 60932008, 61172098, 60871092, 61001013 (国家自然科学基金); the Fundamental Research Funds for the Central Universities of China under Grant No. HIT.ICRST.2010022 (中央高校基本科研业务费专项资金); the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No. 201003446 (高等学校博士学科点专项科研基金).

Received 2011-05, Accepted 2011-07.

摘 要：为了解决传统基于同源序列比对的抗性基因识别方法中假阳性高、无法发现新的抗性基因的问题，提出了一种利用随机森林分类器和 *K*-Means 聚类降采样方法的抗性基因识别算法。针对目前研究工作中挖掘盲目性大的问题，进行两点改进：引入了随机森林分类器和 188 维组合特征来进行抗性基因识别，这种基于样本统计学习的方法能够有效地捕捉抗性基因内在特性；对于训练过程中存在的严重类别不平衡现象，使用基于聚类的降采样方法得到了更具代表性的训练集，进一步降低了识别误差。实验结果表明，该算法可以有效地进行抗性基因的识别工作，能够对现有实验验证数据进行准确的分类，并在反例集上也获得了较高的精度。

关键词：随机森林；分类器；抗性基因；聚类；降采样

文献标识码：A 中图分类号：TP18

1 引言

随着分子生物学与基因组学的快速发展，从分子层面对抗性基因进行识别，对获得抗病优质转基因植物、探究抗病机制都具有十分重要的意义[1]。长久以来，植物通过衍生出多种防御机制来对抗病原物的侵袭。其中一些抗病策略是单纯通过物理或化学的方式来进行阻隔，而更多的则是基于植物与病原物之间基因-基因相互作用来实现[2]。

植物抗性基因(resistance gene, R-gene)在识别特定的病原体无毒基因方面具有重要作用[3]。植物体免疫系统的快速进化导致了很大程度上的基因多样性[4−5]。尽管对抗性基因的认识有限，但近期研究表明，由抗性基因编码的蛋白质在结构上具有模块化的区域，且区域之间会动态地发生一些相互作用来表达其抗性功能。根据这些区域可以将抗性基因大致分为 5 个类别：(1) NBS-LRR，含有核苷酸结合位点(NBS)和富亮氨酸重复(LRR)的胞内受体蛋白基，包括 2 个亚类，TIR-NBS-LRR 和 CC-NBS-LRR；(2) 细胞间的苏氨酸/丝氨酸蛋白激酶(PK)基因；(3) LRR-TM，N 端存在一个胞外 LRR，C 端具有由疏水氨基酸组成的跨膜区的受体蛋白基因；(4) PK-LRR-TM，除含有 LRR-TM 结构外，还具有 PK 结构；(5) SA-CC。此外，还有一些抗性基因无法归属于这 5 个大类[6−7]。

传统的抗性基因识别主要使用基于同源比对的方法[8−9]，认为序列上相似的蛋白质倾向于具有相似的结构和功能。但一些相关研究表明，相似性高的序列也存在表达不同功能的情况；同时，一些相似性低的序列却表达了相同的功能。由此可知，基于同源比对的方法会存在两个问题：一是只能发现与现有抗性基因类似的序列，对未出现过的新基因缺乏识别与挖掘能力；二是假阳性高，会挖掘出大量假抗性基因，对后续的生物验证实验造成困扰。针对以上问题，本文结合机器学习中的随机森林算法，将抗性基因的识别问题转化为一个两类分类问题。

随机森林(random forest)是一种统计学习理论，利用 bootstrap 重抽样方法从原始样本中抽取多个样本，对每个 bootstrap 样本构建决策树，然后将所有决策树中出现最多的投票结果作为最终预测结果。该方法有很多优点：它能够处理高维度的数据，无需进行特征选择；能够在训练过程中检测到属性之间的相互影响；实现简单，易于并行化。因此被成功应用到诸多领域[10−12]。大量理论和实证研究都证明了该算法具有很高的预测准确率，对异常值和噪声数据具有很好的容忍度，且不会出现过拟合现象。因此，选用该算法可以保证最终模型的稳定性，并且最终得到的分类模型具有很好的泛化能力。

2 随机森林及其理论背景

2.1 随机森林定义

随机森林算法是由 Breiman[13]提出的基于决策树分类器的融合算法。其基本思想是将很多的弱分类器集成为一个强分类器。

定义 **1**(随机森林) 随机森林是由多个决策树 $\{h(\boldsymbol{x},\boldsymbol{\theta}_k)\}$ 组成的分类器，其中 $\{\boldsymbol{\theta}_k\}$ 是相互独立且

同分布的随机向量。最终由所有决策树综合决定输入向量 $\boldsymbol{x}$ 的最终类标签。

2.2 随机森林的理论背景

为了构造 k 棵树，需要先产生 k 个随机向量 $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,...,\boldsymbol{\theta}_k$，这些随机向量 $\boldsymbol{\theta}_i$ 是相互独立的，并且是同分布的。随机向量 $\boldsymbol{\theta}_i$ 用于构造决策分类树 $h(\boldsymbol{x},\boldsymbol{\theta}_i)$，简化为 $h_i(\boldsymbol{x})$。

给定 k 个分类器 $h_1(\boldsymbol{x}),h_2(\boldsymbol{x}),...,h_k(\boldsymbol{x})$ 和一组从随机向量 $\boldsymbol{X}$、Y 的分布中随机抽取的训练集，定义边缘函数：

$$mg(\boldsymbol{X},Y)=av_kI(h_k(\boldsymbol{X})=Y)-\max_{j\neq Y}av_kI(h_k(\boldsymbol{X})=j) \quad (1)$$

其中，$I(\bullet)$是示标函数。该边缘函数刻画了对向量 $\boldsymbol{X}$ 正确分类 Y 的平均得票数超过其他任何类平均得票数的程度。可以看出，边际越大分类的置信度就越高。于是分类器的泛化误差为：

$$PE^*=P_{\boldsymbol{X},Y}(mg(\boldsymbol{X},Y)<0) \quad (2)$$

下标 $\boldsymbol{X}$、Y 表明该误差是在 $\boldsymbol{X}$、Y 空间下的。

将上面的结论推广到随机森林，$h_k(\boldsymbol{X})=h(\boldsymbol{X},\boldsymbol{\theta}_k)$。如果森林中树的数目较大，可以根据大数定律和树的结构得到以下定理。

定理 **1** 随着树的数目增加，对所有随机向量 $\boldsymbol{\theta}_i(i=1,2,...,k)$，$PE^*$趋向于：

$$P_{\boldsymbol{X},Y}(p_{\boldsymbol{\theta}}(h(\boldsymbol{X},\boldsymbol{\theta})=Y)-\max_{j\neq Y}p_{\theta}(h(\boldsymbol{X},\boldsymbol{\theta})=j)<0) \quad (3)$$

定理 1 的证明在文献[9]中已经给出，并且表明随机森林不会发生过拟合。这是随机森林的一个重要特点，并且随着树的数目增加，泛化误差 PE^*将趋于上界，这表明随机森林对未知数据有很好的扩展。

定义 **2** 随机森林的边缘函数：

$$mr(\boldsymbol{X},Y)=P_{\theta}(h(\boldsymbol{X},\boldsymbol{\theta})=Y)-\max_{j\neq Y}P_{\boldsymbol{\theta}}(h(\boldsymbol{X},\boldsymbol{\theta})=j) \quad (4)$$

分类器 $\{h(\boldsymbol{X},\boldsymbol{\theta}_k)\}$ 的强度：

$$s=E_{\boldsymbol{X},Y}mr(\boldsymbol{X},Y) \quad (5)$$

假设 $s\geq 0$，根据切比雪夫不等式，由式(3)、(5)可以得到：

$$PE^*\leqslant\frac{\mathrm{var}(mr)}{s^2} \quad (6)$$

不等式(6)要求的 var(mr)具有以下形式：

$$\begin{cases}\mathrm{var}(mr)=\bar{\rho}(E_{\boldsymbol{\theta}}sd(\boldsymbol{\theta}))^2\\ \mathrm{var}(mr)\leqslant\bar{\rho}(E_{\boldsymbol{\theta}}\,\mathrm{var}(\boldsymbol{\theta}))\end{cases} \quad (7)$$

而

$$\begin{cases}E_{\boldsymbol{\theta}}\,\mathrm{var}(\boldsymbol{\theta})\leqslant E_{\boldsymbol{\theta}}(E_{\boldsymbol{X},Y}mg(\boldsymbol{\theta},\boldsymbol{X},Y))^2-s^2\\ E_{\boldsymbol{\theta}}\,\mathrm{var}(\boldsymbol{\theta})\leqslant 1-s^2\end{cases} \quad (8)$$

由式(6)~(8)，可以得到以下结论：

定理 **2** 随机森林的泛化误差上界的定义为：

$$PE^*\leqslant\frac{\bar{\rho}(1-s^2)}{s^2} \quad (9)$$

其中，$\bar{\rho}$ 是相关系数的均值，s 是树的分类强度。

由定理 2 可知，随机森林的泛化误差上界可以根据两个参数推导出来：森林中每棵决策树的分类精度即树的强度 s，以及这些树之间的相互依赖程度 $\bar{\rho}$。当随机森林中各个分类器的相关程度 $\bar{\rho}$ 增大时，泛化误差 PE^*上界就增大；当各个分类器的分类强度增大时，泛化误差 PE^*上界就减小。正确理解这两者之间的相互影响是人们理解随机森林工作原理的基础。

3 设计识别抗性基因的随机森林分类器

3.1 特征选择

抗性基因蛋白一级结构可以看做由 20 种字母(表示 20 个氨基酸)组成的一串字符，因此，如何从这串字符中提取适当的特征值，提高预测精度是本文的工作重点。

目前，氨基酸序列特征提取算法主要分为两类：一类为基于氨基酸组成和位置的特征提取算法，例如氨基酸组成、二肽组成、伪氨基酸组成等；另一类为基于氨基酸理化特性的特征提取算法。单独使用任一方法都不能很好地体现蛋白质的特性。为了较全面地描述所研究的蛋白质，本文特征的选取结合了组成特征与理化特征，这些特征在蛋白质相互作用预测[14]、蛋白质折叠识别[15]以及蛋白质家族分类[16–18]等工作中都有良好表现。

提取的特征包括：氨基酸组成、疏水性、规范范德华体积、极性、极化率、电荷和表面张力等特征。其中，针对理化性质，通过设计一些特征组来描述蛋白质序列的全局信息。例如，C 表示组成，T 表示转换，D 表示分布。C 是具有性质 p 的氨基酸占所在氨基酸序列的百分比，T 是具有性质 p 的氨基酸

其后紧跟具有性质 q 的氨基酸的频率，D 是具有性质 r 的氨基酸第 1 个、第 25%个、第 50%个、第 75%个以及第 100%个分别处于所在氨基酸序列的位置。例如[17]，给定一条假想的蛋白质序列，该序列的字符排列如下所示：AEAAAEAEEAAAAAEAEEEAAEEAEEEAAE。如图 1 可知，丙氨酸个数为 16 个(n_1=16)，谷氨酸个数为 14 个(n_2=14)。对于这两种氨基酸，C 特征组中的值分别为 $n_1\times100.00/(n_1+n_2)$=53.33 与 $n_2\times100.00/(n_1+n_2)$=46.67。该序列中 A 到 E 与 E 到 A 总共转换 15 次，因此 T 特征值为(15/29)×100.00=51.72。下面计算 D 值，对于丙氨酸，其第 1 个、第 25%个、第 50%个、第 75%个以及第 100%个分别位于所在氨基酸序列的第 1 位、第 5 位、第 12 位、第 20 位和第 29 位，因此丙氨酸的 D 特征组中各分量分别为 1/30×100.00=3.33，5/30×100.00=16.67，12/30×100.00=40.0，20/30×100.00=66.6，29/30×100.00=96.67。同理可知谷氨酸各值为 6.67,26.67,60.0,76.67,100.0。在研究中，根据氨基酸是否具有某类性质可以将其分为三类。以疏水性为例，可以分为亲水、疏水以及中性三类。若选用上述三种描述方式只考虑某一理化性质，则其特征向量的维数为 21 维，C 类 3 维、T 类 3 维以及 D 类 15 维。考虑 8 种理化性质，便得出用于表征每条蛋白质序列的 188 维特征。表 1 是番茄中 Asc1 抗性基因蛋白序列的部分计算结果。此外，需要加入 1 维用于表征其类别的特征，其属性值为 1 表示该蛋白为 R 基因，否则不是 R 基因。

至此，共获得 189 维特征。对初始训练集(333 正例+10 807 反例)进行从序列到向量的转化。

3.2 基于 *K*-Means 聚类的降采样方法

虽然随机森林算法具有很多优点，诸如不需要预处理，不会发生过拟合，较少的参数调整等。但是实际问题中的数据不平衡问题却会对分类器性能造成很大影响。所谓数据不平衡，即类与类之间样本数相差很大，不平衡的程度越大，发生过拟合的可能性就越大。

生物信息学研究中，大多数问题的正例来自于实验验证，获取成本高；而反例的获取不需要实验验证，获取成本低。因此经常发生数据不平衡问题[19]。此类问题已经引起许多学者的关注，并提出了很多解决方法。最早主要是使用随机方法来改变训练集

Sequence	A E A A A E A E E A A A A A E A E E E A A E E A E E E A A E
Sequence index	1 5 10 15 20 25 30
A/E Transitions	| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
Index of A	*1* 2 3 *4* 5 6 7 *8* 9 10 11 *12* 13 14 15 *16*
Index of B	*1* 2 *3* 4 5 6 *7* 8 9 *10* 11 12 13 *14*

Fig.1 The sequence of a hypothetical protein to illustrate derivation of the feature vector of a protein

图 **1** 假想蛋白序列的特征导出示意图

Table 1 Characteristic descriptions of Asc1

表 **1** 抗性基因 **Asc1** 的部分特征描述

理化性质	特征描述值										
氨基酸组成	6.17	1.30	6.17	4.87	9.74	3.90	2.28	8.44	6.49	12.66	2.28
	2.60	1.95	1.30	4.87	6.17	3.90	7.14	2.60	5.20		
疏水性	26.30	29.55	44.15	14.98	19.86	26.71	0.65	20.78	39.93	80.20	100.00
	1.95	31.17	50.32	67.86	99.35	0.32	25.97	53.57	75.32	96.10	
电荷	49.35	22.08	28.57	20.19	24.43	11.73	0.32	27.27	52.59	75.00	96.10
	2.60	29.87	48.38	66.23	98.05	0.65	21.43	48.05	80.84	100.00	

的分布，如随机过采样(over-sampling)和随机降采样(under-sampling)。其中随机过采样通过复制小类样本来实现样本集的平衡。该类方法增加了样本集中样本数目，当数据不平衡程度严重，并且涉及多类别的数据不平衡问题时，计算开销的增长将导致程序性能的大幅下降。与此同时，由于是对小类样本的简单复制，导致分类界面过分靠近小类样本，引发过拟合[20]问题，从而降低分类器泛化性能。而随机降采样，则是通过对大类样本进行随机抽取来实现样本集的平衡，虽然不会发生过拟合问题，但随机抽取会使样本集丢失一些潜在信息。

本文使用的实验数据中，具有抗性的正例数据与非抗性的反例数据存在严重的不平衡性。因此，如何构建平衡训练集也是本文的重要工作。为了避免传统降采样的不足，本文采用基于聚类的降采样算法对反例集进行削减。由于聚类算法[21]中所得到的聚类满足类内间距小、类间间距大的特性，该算法可以在降低反例数量的同时保留尽可能多的反例信息，构建具有代表性的反例集。

基于 *K*-Means 聚类的降采样方法，以 *k*(正例样本的个数)为聚类中心数目，对反例数据集进行 *K*-Means 聚类，提取出 *k* 个聚类中心样本。以这 *k* 个样本作为新的反例集，与原正例集共同组成一个新的平衡训练样本集(如图 2 和算法 1 所示)。

K-Means 是经典的聚类算法，具有简单、快速等特点。其时间复杂度为 $O(nkt)$, n 是所有对象个数，k 是设定类的个数，t 则为迭代次数。

算法 1 基于 *K*-Means 聚类的降采样算法

输入：正例样本集合 S，反例样本集合 $B(|S|<<|B|)$。

输出：平衡训练集 $U(S+|B'|$，其中$|B'|=|S|)$。

1. 计算每个样本集合的样本数，正例样本集合 S 的样本数记为 k，反例样本集合 B 的样本数记为 n;

2. 随机重排 B，从集合 B 的 n 个样本中随机选择 k 个样本作为初始聚类中心;

3. Repeat

4. 计算每个样本到聚类中心的距离(欧氏距离);

5. 选择最近邻的聚类中心，并加入该聚类中心所在的聚类;

6. 根据新产生的聚类，计算该聚类新的聚类中心;

7. Until 每个聚类不再发生变化;

8. 取出最终的 k 个聚类中心记为集合 B';

9. 将步骤 2 中获得的聚类中心集合 B'以及小类样本集合 S 进行合并，作为平衡训练集 U 并输出。

3.3 随机森林决策

随机森林算法中，类标签是由所有决策树的分类结果综合而成，本文使用投票的方式来决定类标签。对测试样例 $\boldsymbol{x}$，预测类标签 c_p，可以得到

$$c_p = \arg\max_c \left(\frac{1}{N} \sum_{i=1}^{N} I\left(\frac{n_{h_i,c}}{n_{h_i}} \right) \right) \tag{10}$$

其中，N 是随机森林中决策树的数目；$I(\bullet)$是示标函数；$n_{h_i,c}$是树 h_i 对类 C 的分类结果；n_{h_i} 是树 h_i 的叶节点个数。抗性基因蛋白识别是一个二类分类问题，因此所有决策树都是二叉树。在预测过程中，将每

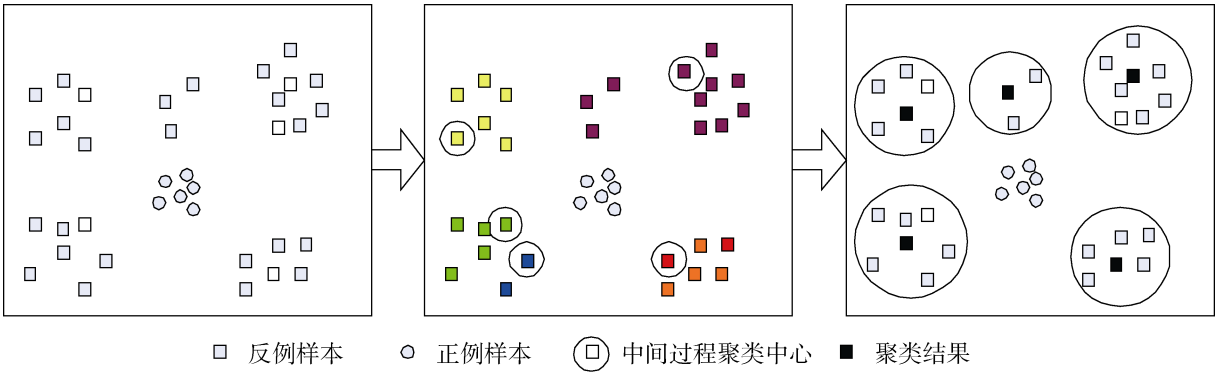

Fig.2 Processes of clustering and under-sampling of unbalance data

图 2 不平衡数据的聚类及降采样过程

个样本在每棵树上的分类结果作为选票，最终投票结果中，得票数超过预先给定阈值的类别为对该样本的最终分类结果。

算法 **2** 基于随机森林的抗性基因识别算法

输入：初始蛋白质训练集 U。

输出：识别抗性基因的随机森林分类器模型。

1. 序列向量化，将训练集 U 中蛋白质序列使用特征提取算法转化为特征向量，得到训练集 U';

2. 使用算法 1，将 U' 转化为平衡训练集 U'';

3. Repeat

4. 使用 bagging 算法，从平衡训练集 U'' 中抽出 $|U''|$ 个样本作为建立第 i 棵树的训练集 S_i，调用 CreateTree 过程，建立决策树;

5. CreateTree 过程：

5.1 如果 S_i 有 m_{all} 维特征，则随机地选取 m_{try} 维特征，并且在这 m_{try} 维特征中选取分类效果最好的一个特征 α 作为该节点的分裂属性，在整个随机森林的构造过程中 m_{try} 是一个常数，本文取 m_{try}= int(lb m_{all}+1);

5.2 根据特征把节点分为两个分支，再分别调用 CreateTree 过程，构造各个分支，直至这棵树能准确分类当前所有训练样例，或所有 G_m 中属性都已被使用;

6. 决策树建立后不进行剪枝;

7. Until 建立了 k 棵树(本文 k=10);

8. 输出训练得到的模型文件。

4 实验设计与实验结果

4.1 实验设计

4.1.1 正负样本的收集

首先，将 http://www.prgdb.org[22]数据库中经实验验证的 73 条植物抗性基因的蛋白序列作为正例。这些序列分别来自 22 个物种，与 31 种病原物发生相互作用(表 2)。另外，由于与抗性基因具有高相似性，文献中提及的抗性基因同源类似物，也被加入正例集中。由此获得包含 333 条蛋白质序列的正例集。

其次，在反例集构建方面，已有研究表明植物抗性基因与其蛋白序列所包含的结构域具有密切关系。因此使用蛋白质家族数据库(Pfam)作为反例集数据来源。Pfam 数据库中共包含 11 912 个蛋白质家族，剔除正例所在的家族，从余下的 11 804 个家族中各随机选取一条蛋白序列，删除其中长度小于 60 个氨基酸(不可能具有抗性)的序列，将最终的 10 807 条蛋白质序列作为反例集。

4.1.2 评价标准

通常采用精度指标来衡量分类器的分类性能，主要包括训练精度和泛化精度：

$$\text{训练精度}=\frac{\text{正确分类的训练样本数}}{\text{训练样本总数}}$$

$$\text{泛化精度}=\frac{\text{正确分类的测试样本数}}{\text{测试样本总数}}$$

然而，这些指标并不能客观地描述类别不平衡条件下分类器的性能。例如，针对由 10 个正例样本和 90 个反例样本组成的数据集，如果一个不具有任何判别效果的分类器将全部样例判定为反例，其训练精度和泛化精度仍可以达到 90%。为此，本文使用生物信息学中预测(分类)问题常用的评价标准，即敏感性(sensitivity, SN)、特异性(specificity, SP)和几何平均数(geometric mean, Gm)：

$$SN=\frac{TP}{TP+FN} \tag{11}$$

$$SP=\frac{TN}{TN+FP} \tag{12}$$

$$Gm=\sqrt{SN\times SP} \tag{13}$$

其中，TP 表示正确预测到的正例个数；TN 表示正确预测到的反例个数；FP 表示把反例预测成为正例的个数；FN 表示把正例预测为反例的个数。

4.2 实验结果和分析

在收集到的训练数据集中，类别不平衡问题十分突出。本文分别采用随机降采样和聚类降采样两种方法对反例数据集进行处理，降采样之后使用随机森林算法构建分类器。为客观地评价两种采样策略，随机森林算法的参数设置相同(决策树数目为 10，每次选择属性个数为 8)。在实验中，采用十折交叉验证(10-fold cross validation)对分类效果进行验证，并统计其 SN、SP 与 Gm 值。表 3 给出了基于两种不同降采样方法的分类器性能。

Table 2 Plant resistance genes identified with indication of donor species, related disease and pathogen

表 **2** 已验证的植物抗性基因物种、相关疾病与病原体信息

基因名	所在物种	疾病	病原体
EFR	Arabidopsis thaliana	Eliciting bacteria	Bacteria with flagellum
ER-Erecta	Arabidopsis thaliana	Bacterial wilt	Ralstonia solanacearum
FLS2	Arabidopsis thaliana	Eliciting bacteria	Bacteria with flagellum
HRT	Arabidopsis thaliana	Turnip crinkle virus	Turnip crinkle virus
PEPR1	Arabidopsis thaliana	Damping off	Pythium
RCY1	Arabidopsis thaliana	Cucumber mosaic virus	Cucumber mosaic virus
RFO1	Arabidopsis thaliana	Fusarium wilt	Fusarium oxysporum
RPM1	Arabidopsis thaliana	Bacterial blight	Pseudomonas syringae
RPP13nd	Arabidopsis thaliana	Downy mildew	Hyaloperonospora parasitica
RPP4	Arabidopsis thaliana	Downy mildew	Peronospora parasitica
RPP5	Arabidopsis thaliana	Downy mildew	Hyaloperonospora parasitica
RPP8	Arabidopsis thaliana	Downy mildew	Hyaloperonospora parasitica
RPS2	Arabidopsis thaliana	Bacterial blight	Pseudomonas syringae
RPS4	Arabidopsis thaliana	Bacterial blight	Pseudomonas syringae
RPS5	Arabidopsis thaliana	Bacterial blight	Pseudomonas syringae
RPW8.1	Arabidopsis thaliana	Powdery mildew	Golovinomyces cichoracearum
RPW8.2	Arabidopsis thaliana	Powdery mildew	Golovinomyces cichoracearum
RSS1	Arabidopsis thaliana	Bacterial wilt	Ralstonia solanacearum
RTM1	Arabidopsis thaliana	Synergistic disease syndromes	Tobacco etch virus
RTM2	Arabidopsis thaliana	Synergistic disease syndromes	Tobacco etch virus
Hs1	Beta procumbens	Beet cyst nematode	Heterodera schachtii
Bs3	Capsicum annuum	Bacterial spot	Xanthomonas campestris pv.
Bs3-E	Capsicum annuum	Bacterial spot	Xanthomonas campestris pv.
Bs2	Capsicum chacoense	Bacterial spot	Xanthomonas campestris pv.
At1	Cucumis melo	Cucurbit downy mildew	Pseudoperonospora cubensis
At2	Cucumis melo	Cucurbit downy mildew	Pseudoperonospora cubensis
Rmd-c	Glycine max	Powdery mildew	Microsphaera sparsa
Rps1-k-1	Glycine max	Phytophthora root	Phytophthora sojae
Rps1-k-2	Glycine max	Phytophthora root	Phytophthora sojae
MLA10	Hordeum vulgare	Powdery mildew(barley)	Blumeria graminis
Mlo	Hordeum vulgare	Powdery mildew(barley)	Blumeria graminis
RPG1	Hordeum vulgare	Stem rust	Puccinia graminis
Dm-3	Lactica sativa	Downy mildew	Bremia lactucae
L6	Linum usitatissimum	Flax rust	Melampsora lini
M	Linum usitatissimum	Flax rust	Melampsora lini
P2	Linum usitatissimum	Flax rust	Melampsora lini
N	Nicotiana glutinosa	Tobacco mosaic virus	Tobacco mosaic virus
Pi33	Oryza sativa	Rice blast disease	Magnaporthe grisea
Xal	Oryza sativa	Bacterial blight	Xanthomonas oryzae
Xa21	Oryza sativa Indica group	Bacterial blight	Xanthomonas oryzae
Pi-ta	Oryza sativa Japonica group	Rice blast disease	Magnaporthe grisea
PGIP	Phaseolus vulgaris	Eliciting fungus	Fungus producing polygalaturonases
Rx2	Solanum acaule	Latent mosaic	Potato virus X
Rpi-blb1	Solanum bulbocastanum	Late blight tomato	Phytophthora infestans
Rpi-blb2	Solanum bulbocastanum	Late blight tomato	Phytophthora infestans
R1	Solanum demissum	Late blight tomato	Phytophthora infestans
Cf4	Solanum habrochaites	Leaf mould	Passalora fulva

续表 **2**

基因名	所在物种	疾病	病原体
Cf4A	Solanum habrochaites	Leaf mould	Passalora fulva
Asc1	Solanum lycopersicum	Alternaria alternate	Alternaria alternate
Bs4	Solanum lycopersicum	Bacterial spot	Xanthomonas campestris
Hero	Solanum lycopersicum	Yellow potato cyst nematode	Globodera
I2	Solanum lycopersicum	Fusarium wilt	Fusarium oxysporum
LeEIX1	Solanum lycopersicum	Eliciting fungus	Fungal ethylene-inducing xylanase
LeEIX2	Solanum lycopersicum	Eliciting fungus	Fungal ethylene-inducing xylanase
Mil.2	Solanum lycopersicum	Root-knot nematode	Meloidogyne Paratrichodorus minor
Sw5	Solanum lycopersicum	Tomato spotted wilt	Tomato spotted wilt virus
Tm2	Solanum lycopersicum	Tobacco mosaic virus	Tobacco mosaic virus
Tm2a	Solanum lycopersicum	Tobacco mosaic virus	Tobacco mosaic virus
Ve1	Solanum lycopersicum	Verticillium wilt potato	Verticillium
Ve2	Solanum lycopersicum	Verticillium wilt potato	Verticillium
Cf5	Solanum lycopersicum var.Cerasiforme	Leaf mould	Passalora fulva
Cf2	Solanum pimpinellifolium	Leaf mould	Passalora fulva
Prf	Solanum pimpinellifolium	Bacterial speck	Pseudomonas syringae
Pto	Solanum pimpinellifolium	Bacterial speck	Pseudomonas syringae
Cf9	Solanum pimpinellifolium	Leaf mould	Passalora fulva
Cf9B	Solanum pimpinellifolium	Leaf mould	Passalora fulva
Gpa2	Solanum tuberosum	Yellow potato cyst nematode	Globodera
Gro1.4	Solanum tuberosum	Late blight potato	Phytophthora infestans
R3a	Solanum tuberosum	Late blight tomato	Phytophthora infestans
Rx	Solanum tuberosum	Latent mosaic	Potato virus X
RY1	Solanum tuberosum subsp andigena	Potato virus Y	Potato virus Y
Hm1	Zea mays	Leaf spot	Bipolaris zeicola
Hm2	Zea mays	Leaf spot	Bipolaris zeicola

Table 3 Performance comparison of two under-sampling methods using random forest

表 **3** 使用两种降采样方法的随机森林分类器性能比较

采样方法	正例样本数	反例样本数	分类精度/(%)		
			SN	*SP*	*Gm*
随机采样(十折交叉验证)	333	333	79.43	81.86	80.64
K-Means 采样(十折交叉验证)	333	333	95.61	98.15	96.87

表 3 中，两种方法所使用的训练集中均含有 666 条序列，包括正例 333 条，反例 333 条。其中，随机采样方法中反例样本由随机降采样获得，而 *K*-Means 样本选择方法中反例样本则由本文算法 1 获得。从分类性能来看，与随机降采样方法相比，聚类降采样方法能够大幅提升分类器分类能力。这表明，采用合理的样本降采样方法，对于解决样本不平衡分类问题是十分重要的。

此外，作为挖掘抗性基因的模型，对于已有生物验证抗性基因的识别能力不容忽视。因此，构建了包含 73 条生物验证的正例序列以及 10 474 条未被选入训练集的反例序列的测试集。由于正例与反例过于悬殊，会因为 *SN* 过小而无法对两种模型进行有效评判。所以改用 *TPrate*、*TNrate* 以及 *ACC* 评价预测水平，实验结果见表 4。

$$TPrate = \frac{TP}{TP + FP} \tag{14}$$

$$TNrate = \frac{TN}{TN + FN} \tag{15}$$

Table 4 Performance comparison of two classifiers on test dataset

表 **4** 两个分类器在测试集上的性能比较

采样方法	正例样本数	反例样本数	分类精度/(%)		
			TPrate		*ACC*
随机采样	73	10 474	正例	85.89	80.47
			反例	80.43	
K-Means 采样	73	10 474	正例	100.00	88.63
			反例	88.57	

$$ACC = \frac{TP + FN}{TP + FP + TN + FN} \tag{16}$$

由表 4 可以看出，采用本文方法获得的模型可以将现有已验证的 73 条抗性基因全部正确分类(表 5)。同时，在反例预测方面，也获得较高的分类精度。实验数据证明，所建立的基于聚类降采样和随机森林的抗性基因识别分类器是合理、有效的。

Table 5 Classification results of resistance genes

表 **5** 实验已验证抗性基因分类结果

物种名称	*R*-gene 个数	结果
Arabidopsis thaliana	20	correct
Beta procumbens	1	correct
Capsicum annuum	2	correct
Capsicum chacoense	1	correct
Cucumis melo	2	correct
Glycine max	3	correct
Hordeum vulgare	3	correct
Lactica sativa	1	correct
Linum usitatissimum	3	correct
Nicotiana glutinosa	1	correct
Oryza sativa	4	correct
Phaseolus vulgaris	1	correct
Solanum acaule	1	correct
Solanum bulbocastanum	2	correct
Solanum demissum	1	correct
Solanum habrochaites	2	correct
Solanum lycopersicum	13	correct
Solanum pimpinellifolium	5	correct
Solanum tuberosum	5	correct
Zea mays	2	correct

5 结束语

针对现有抗性基因识别过程中出现的假阳性高、难以发现新基因的问题，本文提出了一种基于随机森林的识别抗性基因的新方法。

将机器学习算法引入到抗性基因识别问题中，探索了一条识别抗性基因的新思路。主要创新工作体现在：

(1) 使用 188 维特征来表征抗性基因序列，从基因的一级序列中挖掘出尽可能多的有效表达抗性的信息。

(2) 针对存在的类别不平衡问题，采用聚类降采样方式，在降低数据量的同时尽可能保留原始数据中所包含的信息，很大程度上提高了分类器的性能。

(3) 采用随机森林算法，随机森林算法自身的优良特性保证了分类器的泛化能力。

实验结果表明，上述方法具有较高的准确性和较好的泛化能力，可以用于对特定物种的抗性基因的识别。

抗性基因的识别问题是一个重要的课题，其二级结构特征对于保证抗性功能具有重要的作用。本文所建立的模型仅考虑序列的组成信息及理化特征，对于如何合理地提取二级结构特征还需进一步研究。此外，在高维特征的前提下，如何找到更加有效表达抗性的特征子集，及其对该方法的影响将是未来的工作。

References:

[1] Liu Jinling, Liu Xionglun, Dai Liangying, et al. Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants[J]. Journal of Genetics and Genomics, 2007, 34(9): 765−776.

[2] Flor H H. Current status of the gene-for-gene concept[J]. Annual Review of Phytopathology, 1971, 9: 275−296.

[3] Ellis J, Dodds P, Pryor T. The generation of plant disease resistance gene specificities[J]. Trends in Plant Science, 2000, 5(9): 373−379.

[4] Chisholm S T, Coaker G, Day B, et al. Hostmicrobe interactions: shaping the evolution of the plant immune response[J]. Cell, 2006, 124(4): 803–814.

[5] Means T K, Golenbock D T, Fenton M J. The biology of toll-like receptors[J]. Cytokine & Growth Factor Reviews, 2000, 11(3): 219–232.

[6] Buschges R, Hollricher K, Panstruga R, et al. The barley Mlo gene: a novel control element of plant pathogen resistance[J]. Cell, 1997, 88(5): 695–705.

[7] Brandwagt B F, Mesbah L A, Takken F L, et al. A longevity assurance gene homolog of tomato medicates resistance to alternaria alternate f.sp. lycopersici toxins and fumonisin B[J]. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(9): 4961–4966.

[8] Chelkowski J, Koczyk G. Resistance gene analogues of arabidopsis thaliana recognition by structure[J]. Journal of Applied Genetics, 2003, 44(3): 311–321.

[9] Koczyk G, Chelkowshki J. An assessment of the resistance gene analogues of Oryza sativa ssp japonica: their presence and structure[J]. Cellular Molecular Biology Letters, 2003, 8(4): 963–972.

[10] Wang Xiaoting, Ding Xiaoqing, Fang Chi. Accurate localization of facial feature points based on random forest classifier[J]. Journal of Tsinghua University: Science and Technology, 2009, 49(4): 543–546.

[11] Li Jiangeng, Gao Zhikun, Ruan Xiaogang. Random forest-based gene pathway analysis of gastric cancer microarray data[J]. Journal of Biology, 2010, 27(2): 1–4.

[12] Lin Chengde, Peng Guolan. Application of random forest on selecting evaluation index system for enterprise credit assessment[J]. Journal of Xiamen University: Natural Science, 2007, 46(2): 199–203.

[13] Breiman L. Random forests[J]. Maching Learning, 2001, 45(1): 5–32.

[14] Bock J R, Gough D A. Predicting protein-protein interactions from primary structure[J]. Bioinformatics, 2001, 17(5): 455–460.

[15] Chris D, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks[J]. Bioinformatics, 2001, 17(4): 349–358.

[16] Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines[J]. Bioinformatics, 2002, 18(1): 147–159.

[17] Lin H H, Han L Y, Cai C Z, et al. Prediction of transporter family from protein sequence by support vector machine approach[J]. Proteins: Structure, Function and Bioinformatics, 2006, 62(1): 218–231.

[18] Cai C Z, Han L Y, Ji Z L, et al. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence[J]. Nucleic Acids Research, 2003, 31(13): 3692–3697.

[19] Zou Quan, Guo Maozu, Liu Yang, 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.

[20] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.

[21] Show J Y, Yue S L. Cluster-based under-sampling approaches for imbalanced data distribution[J]. Expert Systems with Application, 2009, 36(3): 5718–5727.

[22] Sanseverino W, Roma G, Simone M D, et al. PRGdb: a bioinformatics platform for plant resistance gene analysis[J]. Nucleic Acids Research, 2010, 38(Database): 814–821.

附中文参考文献：

[10] 王晓婷, 丁晓青, 方驰. 基于随机森林的人脸关键点精确定位方法[J]. 清华大学学报: 自然科学版, 2009, 49(4): 543–546.

[11] 李建更, 高志坤, 阮晓钢. 基于随机森林的胃癌微阵列数据基因通路分析[J]. 生物学杂志, 2010, 27(2): 1–4.

[12] 林成德, 彭国兰. 随机森林在企业信用评估指标体系确定中的应用[J]. 厦门大学学报: 自然科学版, 2007, 46(2): 199–203.

[19] 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.

GUO Yingjie was born in 1987. She is a master candidate at Harbin Institute of Technology. Her research interests include bioinformatics and machine learning, etc.
郭颖婕(1987—)，女，浙江宁波人，哈尔滨工业大学硕士研究生，主要研究领域为生物信息学，机器学习等。

LIU Xiaoyan was born in 1963. She is an associate professor at Harbin Institute of Technology. Her research interests include bioinformatics, engineering database and artificial intelligence, etc.
刘晓燕(1963—)，女，山东泰安人，哈尔滨工业大学副研究员，主要研究领域为生物信息学，工程数据库，人工智能等。

GUO Maozu was born in 1966. He is a professor and Ph.D. supervisor at Harbin Institute of Technology. His research interests include machine learning, computational biology and image understanding, etc.
郭茂祖(1966—)，男，山东夏津人，哈尔滨工业大学教授、博士生导师，主要研究领域为机器学习，计算生物学，图像理解等。

ZOU Quan was born in 1982. He is an assistant professor and master supervisor at Xiamen University. His research interests include bioinformatics and data mining, etc.
邹权(1982—)，男，黑龙江佳木斯人，厦门大学助理教授、硕士生导师，主要研究领域为生物信息学，数据挖掘等。