计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20170077
55(5):986-993,2018

结合近邻传播聚类的选择性集成分类方法

孟 军 张 晶 姜丁菱 何馨宇 李丽双
(大连理工大学计算机科学与技术学院 辽宁大连 116023)
(mengjun@dlut.edu.cn)

Selective Ensemble Classification Integrated with Affinity Propagation Clustering

Meng Jun, Zhang Jing, Jiang Dingling, He Xinyu, and Li Lishuang
(*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116023)

Abstract Mining useful knowledge from gene expression data is a hot research topic and direction of bioinformatics. Gene microarray data are characterized by high dimensionality, small sample size and high redundancy. Therefore, a gene selection method based on the intersection neighborhood rough set is presented to select important genes for the classification of microarray data. First, pathway knowledge is used to preselect genes, and each pathway unit is corresponding to a gene subset. Then the attribute reduction method based on rough set is applied to select important genes without redundancy for classification. Due to the large number of pathway knowledge units, many base classifiers are generated. In order to further improve the diversity among base classifiers and the efficiency of ensemble model, it is necessary to select part of base classifiers. Affinity propagation (AP) clustering needn't to set the number of clusters and the starting points, and it can obtain clusters more quickly and accurately. Therefore, AP clustering algorithm is used to group base classifiers into many clusters with significant diversity among them, then selecting a classifier from each cluster to generate the final ensemble classifier. Experimental results on three Arabidopsis thaliana biotic and abiotic stress response datasets show that the proposed method can improve the accuracy compared with the existing ensemble methods by more than 12%.

Key words selective ensemble; affinity propagation; pathway; intersection neighborhood; gene microarray data

摘 要 从海量的基因微阵列数据中提取出有价值的信息是生物信息学的研究热点.基因微阵列数据具有高维度、小样本和高冗余的特性.因此,提出一种基于相交邻域粗糙集的基因选择方法,挑选出关键基因用于对微阵列数据进行分类.首先利用 pathway 知识进行基因初步选择,每个 pathway 单元对应一个基因子集,然后采用基于粗糙集的属性约简方法筛选出无冗余的关键基因.由于 pathway 知识单元的数量较多,对应生成大量的基分类器,为了进一步提高基分类器之间的差异性和集成的效率,对基分类器进行选择是十分必要的.近邻传播聚类不需要提前设定聚簇数量和起始点并且可以更快速、精确地进行聚类.因此,使用近邻传播聚类方法对基分类器进行分组,产生差异性较大的聚簇,再从每个簇中选择一个分类器构建集成分类器.在拟南芥的生物和非生物胁迫响应相关的微阵列数据集上的实验结果表明:在准确率方面,提出的方法与现有的集成方法相比最多可以提高 12%.

关键词 选择性集成;近邻传播;通路;相交邻域;基因微阵列数据

中图法分类号 TP181

收稿日期:2017-02-21;**修回日期**:2017-05-03

基金项目:国家自然科学基金项目(61472061,61672126)

This work was supported by the National Natural Science Foundation of China (61472061, 61672126).

高通量测序技术的发展,为研究者提供了海量的基因表达数据,从中提取出有价值的信息已经成为生物信息学的研究热点[1-2].

植物在生长过程中经常会受到病虫害和环境因素的影响,如何预测并做好防治工作,对林业、农牧业、环境保护等多方面的发展将起到非常重要的作用.由于基因表达数据具有"高维度"、"小样本"和"高冗余"的特点,采用传统的单分类算法会出现分类稳定性差和准确率偏低等问题,因而对此类数据的分析需要处理能力较强的分类模型.

利用基分类器之间的信息互补性构建的集成分类模型,能够充分发挥各基分类器的优势,因此,具有更稳定、更精确的分类性能[3-4].

以往的集成模型通常只是以基因表达数据这一单一数据源为基础而设计的,由于基因之间具有生物学上的相互作用,融合已有的生物知识到分类模型中能够提高其分类性能.基因本体(GO)知识首先被应用到癌症的预测中,相关的实验表明结合生物知识能够提高预测结果的准确性并且增强其生物学角度上的可解释性与可信性[5].之后,结合通路(pathway)知识的预测模型同样被应用到了癌症的预测中[6].近年来,在疾病分类领域出现了结合超盒原理的 pathway 水平下的分类模型[7].

本文利用 pathway 知识对基因进行初步选择,使得筛选出的每个基因子集中只包含一个 pathway 知识单元中的基因.因为基因微阵列数据的高冗余性,只有极少数的基因与分类相关[8],所以在每个基因子集上采用了基于相交邻域粗糙集的基因选择模型为之后的分类工作选择出重要的、无冗余的基因.

在基分类器数量较多的情况下,会存在一些冗余的分类器,导致整体的差异性较差.为了提高集成分类的性能,对基分类器进行选择是十分必要的.选择性集成方法可以大致分为 4 类:迭代优化法、排名法、聚类法和模式挖掘法.在基于聚类技术的选择性集成中,Lin 等人[9]提出了基于 K-means 聚类和循环序列的动态基分类器选择策略.Zhang 等人[10]提出了基于谱聚类的集成剪枝方法.Krawczyk[11]在加权的 Bagging 集成分类中使用了基于聚类的剪枝方法.

本文采用基于近邻传播(affinity propagation,AP)聚类的选择性集成方法,在聚类的过程中最大化簇内的相似性和簇间的差异性[12].然后从每个聚簇内挑选出有代表性的基分类器进行集成,由于产生的基分类器聚簇之间具有较大的差异性,使得所选择的基分类器之间同样具有较大差异性.

本文的主要贡献有 2 个方面:

1) 在分类模型中引入先验知识,将 pathway 知识应用到对基因的选择中,利于基于相交邻域粗糙集的基因选择模型在不改变分类性能的前提下筛选出重要的基因.

2) 将聚类方法应用到集成分类模型中,采用基于 AP 聚类的选择性集成分类方法.

1　基于相交邻域粗糙集的基因选择

1.1　结合 pathway 的基因初选

KEGG 的 pathway 数据库整合了分子互动网络的知识,包括图解的细胞生化过程,如代谢、细胞周期、信号传递、膜转运等,还包括同系保守的子通路等信息[13].该数据库包含 7 个方面的分子间相互作用和反应网络:

1) 新陈代谢;

2) 遗传信息加工;

3) 环境信息加工;

4) 细胞过程;

5) 生物体系统;

6) 人类疾病;

7) 药物开发[13].

每个 pathway 单元列出一个通路中所包含的所有基因.采用将每个知识单元中所包含的基因与基因微阵列数据中的基因相对应的方式来进行 2 种数据的结合.每个 pathway 单元对应一个基因子集,同时形成一个只包含该基因子集中基因的训练集.由于各基因子集中包含不同的基因,数据分布差异显著,因此基于这些基因子集训练得到的基分类器之间将具有显著的差异性.并且这种知识融合方法还能够降低数据的维数,因此适合于对基因微阵列数据进行处理.目前还有一些基因在相关的 pathway 数据库中没有完善的注释信息.随着生物知识的完善,结合 pathway 的分类方法将会有更好的效果.

定义 1[14]. 基因表达数据决策表的形式化定义为 $GEDT=(U,C\cup D,V,f)$,U 代表论域,是一个非空且有限的对象集合,对应于基因表达数据中的样本;C 代表条件属性集,即所有基因的集合;D 代表决策属性集,即类别属性.要求 C 与 D 均为非空且有限的属性集合.$A=C\cup D$,对于 $\forall a\in A$,f_a:$U\rightarrow V_a$ 是一个全映射,其中 V_a 代表属性 a 的值域,且 $V=\bigcup_{a\in A}V_a$,$f=\bigcup_{a\in A}f_a$.

对于 $GEDT=(U,C\cup D,V,f)$，将其与 pathway 知识结合，可以得到一种如表 1 所示的结合 pathway 知识的信息表，其中 $s_i(i=1,2,\cdots,n)$ 表示基因表达数据中的第 i 个样本，f_{kji} 表示样本 s_i 在 pathway 知识单元 $P_k(k=0,1,\cdots,m)$ 中基因 $p_{kj}(j=1,2,\cdots,|P_k|)$ 上的表达量.

Table 1 Information Table Formed by Combing Microarray Data with Pathway Knowledge

表 1 结合 pathway 知识和微阵列数据的信息表

Sample	P_1			P_2			…	P_m		
	p_{11}	p_{12}	…	p_{21}	p_{22}	…		p_{m1}	p_{m2}	…
s_1	f_{111}	f_{121}	…	f_{211}	f_{221}	…	…	f_{m11}	f_{m21}	…
s_2	f_{112}	f_{122}	…	f_{212}	f_{222}	…	…	f_{m12}	f_{m22}	…
⋮										
s_n	f_{11n}	f_{12n}	…	f_{21n}	f_{22n}	…	…	f_{m1n}	f_{m2n}	…

1.2 基因选择算法

Yao 和 Lin[15] 对 Pawlak 提出的经典粗糙集模型进行了一定的扩展，将原有的等价关系用任意的二元关系替代，提出了广义的粗糙集理论，该理论适用于数值型数据的处理. 广义的粗糙集将原理论中的基本集合由等价类变成了现有理论中的邻域，因此新提出的模型也可以看作是基于邻域的粗糙集模型. 本文使用基于相交邻域的粗糙集模型对基因进行选择. 相交邻域较之于邻域，能够更大程度上减少属性之间的相互影响[16]. 对于基因选择问题，在构建邻域时仅从表达数据本身来说，无法准确获知基因之间的关联性. 因此，相交邻域的定义形式更具合理性.

定义 2[15]. R 是 U 上的一个二元关系，给定对象 $x\in U$，x 的邻域表示为

$$N_R(x)=\{y\mid xRy,y\in U\}. \tag{1}$$

定义 3[16]. 对于属性集合 $B\subseteq C$，给定对象 $x\in U$，x 的相交邻域定义为

$$IN_B(x)=\{y\mid \forall b\in B,|f_b(x)-f_b(y)|\leqslant\delta_b,y\in U\}. \tag{2}$$

定义 4[15]. 对于对象集合 $X\subseteq U$，给定属性集合 $B\subseteq C$，则基于相交邻域，集合 X 的上、下近似集分别定义为

$$\overline{IN}_B(X)=\{x\mid IN_B(x)\cap X\neq\varnothing,x\in U\}, \tag{3}$$

$$\underline{IN}_B(X)=\{x\mid IN_B(x)\subseteq X,x\in U\}. \tag{4}$$

粗糙集理论的基本思想是利用上近似集和下近似集所确切表示的对象集合来近似表示 U 上的一个模糊的对象集 X.

定义 5[15]. 对于对象集合 $X\subseteq U$，基于属性子集 $B\subseteq C$，X 的正域、负域和边界的定义形式分别为

$$POS_B(X)=\underline{IN}_B(X), \tag{5}$$

$$NEG_B(X)=U-\overline{IN}_B(X), \tag{6}$$

$$BN_B(X)=\overline{IN}_B(X)-\underline{IN}_B(X), \tag{7}$$

其中，正域是确定属于集合 X 的所有对象组成的集合，其对象范围与下近似集相同；负域是确定不属于集合 X 的所有对象的集合，即属于论域 U 却不属于上近似集的对象所组成的集合；边界集合是属于上近似集而不属于下近似集的对象所组成的集合，即无法确定是否能归类于集合 X 的对象集合.

决策表上的属性约简是指对条件属性进行的约简，条件属性的约简与决策属性是相对的.

定义 6[17]. 决策表 $GEDT=(U,C\cup D,V,f)$，对于属性集合 $B\subseteq C$，相对于决策属性集 D 的正域定义为

$$POS_B(D)=\bigcup_{X\in U/IND(D)}POS_B(X), \tag{8}$$

其中，$U/IND(D)$ 是由等价关系 $IND(D)$ 导出的所有不同等价类的集合，每个等价类中的样本具有相同的类别标签.

定义 7[17]. 对于条件属性 $c\in B\subseteq C$，如果 $POS_{B-\{c\}}(D)=POS_B(D)$ 成立，则称属性 c 在条件 B 中相对于决策属性 D 是非必要的，则属性 c 是可以约简的；如果条件不成立，称属性 c 在条件属性 B 中相对于决策属性 D 是必要的，则是不可约简的.

定义 8[17]. 对于条件属性集合 $B\subseteq C$，如果 B 中的每一个属性相对于决策属性 D 都是必要的，则称条件属性子集 B 相对于决策属性 D 是独立的.

定义 9[17]. 对于条件属性集合 $B\subseteq C$，集合 $E\subseteq B$，当且仅当 E 相对于决策属性 D 是独立的，且满足公式 $POS_E(D)=POS_B(D)$，则 E 被称为 B 的一个相对于决策属性 D 的约简.

约简中的每一个条件属性相对于决策属性都是必要的，并且约简出的条件属性子集与原始条件属性集合具有同等的分类性能. 将以上的属性约简流程列于算法 1 中.

算法 1. 基于相交邻域的属性约简算法.

输入：决策表 $GEDT=\{U,C\cup D,V,f\}$、待约简的属性子集 B、相交邻域阈值 δ；

输出：B 的相对于决策属性 D 的约简 RED.

① 对于每个对象 $x\in U$，计算其基于属性子集 B 的相交邻域 $IN_B(x)$.

② 将所有对象依据类别属性 $D=\{d\}$ 进行划分，得到基于决策属性的等价类集合 $U/IND(D)$.

③ 计算在属性集合 B 条件下相对于决策属性 D 的正域 $POS_B(D)$.

④ 令 $RED=B$.

⑤ 对于每个对象 $x\in U$,计算其基于属性子集 $RED-\{b\}$ 的相交邻域 $IN_{RED-\{b\}}(x)$,并且计算在属性子集 $RED-\{b\}$ 条件下的相对于决策属性 D 的正域 $POS_{RED-\{b\}}(D)$.

如果 $POS_{RED-\{b\}}(D)=POS_B(D)$,则 $RED=RED-\{b\}$.

⑥ 采用秩和检测对基因进行排序,从后向前依次进行验证.重复步骤⑤直到所有的基因 $b\in B$ 都进行了验证.

2　结合近邻传播聚类的选择性集成

在构建集成分类模型的过程中,会产生多个基分类器,对基分类器进行剪枝能够减少所需的存储空间和计算资源[18-19].另外,Zhou 等人[20-21]通过实验表明选择部分合适的基分类器进行集成,其分类效果优于对全部基分类器进行集成,这也是选择性集成的另外一个优势.

本文使用基于 AP 聚类的选择性集成分类方法,首先将基分类器划分成多个聚簇,然后从每个聚簇中选择作为聚簇中心的基分类器进行集成.

2.1　近邻传播聚类算法

由 Frey 和 Dueck[22]提出的 AP 聚类算法,由于其较好的性能已被成功应用于数据挖掘的许多领域[23].在 AP 聚类算法中,每一个数据点都被视为潜在的聚簇中心,数据点之间不断地进行信息传递直到算法收敛或迭代结束.与传统的聚类方法如 K-means和自组织映射相比,AP 聚类算法有 3 个优势:

1) 不需要事先指定聚簇个数,也不需要初始化聚簇中心点;

2) 聚类结果更加稳定与准确;

3) 在达到同样聚类精确度的条件下所需时间更短.

AP 聚类算法以相似性矩阵 $\boldsymbol{S}$ 作为输入,矩阵中元素 $s(i,j)$ 表示数据点 i 和 j 之间的相似性,数值越大则表明 2 个数据点之间的相似性越大.矩阵对角线上的值 $s(k,k)$ 称作数据点 k 的参考度(preference),该值越大说明该数据点更加适合作为聚簇中心,因此生成的聚簇个数也就越多.一般而言,将所有数据点的 preference 值设置为相同的数值,以保证所有的数据点具有同等的机会成为聚簇中心.

AP 聚类算法在迭代的过程中传递 2 种信息,Responsibility 和 Availability. $r(i,k)$ 代表前者,表示数据点 k 作为数据点 i 的聚簇中心的适应程度;$a(i,k)$ 代表后者,表示数据点 i 选择数据点 k 作为其聚簇中心的倾向程度.

$$r(i,k)=s(i,k)-\max(a(i,j)+s(i,j)),\tag{9}$$

$$a(i,k)=\min(0,r(k,k)+\sum_{j=1}^{N}(\max(0,r(j,k)))),\tag{10}$$

其中,$j\in\{1,2,\cdots,N\}$,$j\neq k$.

AP 聚类可以自动确定聚簇个数,在迭代的过程中如果 $r(k,k)+a(k,k)>0$,就选择数据点 k 作为聚簇中心.迭代结束之后将剩余的数据点分配给距其最近的聚簇中心.

为了提高 AP 聚类算法的稳定性,引入阻尼系数 λ,这样一来,$r(i,k)$ 和 $a(i,k)$ 就受到上一次迭代计算值的约束.改进后的计算为

$$r_i=(1-\lambda)r_i+\lambda r_{i-1},\tag{11}$$

$$a_i=(1-\lambda)a_i+\lambda a_{i-1},\tag{12}$$

其中,r_i 和 a_i 代表第 i 次迭代的结果,r_{i-1} 和 a_{i-1} 代表第 $i-1$ 次的迭代结果.

本文使用 AP 聚类算法对基分类器依据其在剪枝集上的分类结果进行聚类,并且选择作为聚簇中心的基分类器进行集成,以保证被选择的基分类器之间具有较大的差异性.

2.2　相似性矩阵计算

在对基分类器进行聚类时,依据各基分类器在剪枝集上的分类结果计算相似性,并采用基于 *kappa* 系数和准确率的相似性计算公式.

Cohen[24]提出的 *kappa* 系数是对一致性的一种度量,这种评价系数能够消除因随机性带来的误差,本文采用 *kappa* 系数来测量基分类器之间的相似性[25].

定义 10[24]. 依据基分类器的分类结果,*kappa* 系数的计算为

$$kappa=\frac{Pr(a)-Pr(e)}{1-Pr(e)},\tag{13}$$

$$Pr(a)=\frac{C^{tt}}{N}+\frac{C^{ff}}{N},\tag{14}$$

$$Pr(e)=\frac{C^{tt}+C^{tf}}{N}\times\frac{C^{tt}+C^{ft}}{N}+\frac{C^{ft}+C^{ff}}{N}\times\frac{C^{tf}+C^{ff}}{N},\tag{15}$$

其中,N 代表剪枝集中样本数量,对于属于基分类器集合 $\{H\}_{i=1}^{m}$ 中的 2 个基分类器 H_i 和 H_j,C^{tt} 代表 2 个基分类器同时分类正确的样本数量;C^{ff} 是 2 个

基分类器同时分类错误的样本数量；C^{tf} 是 H_i 分类正确 H_j 分类错误的样本的数量；C^{ft} 与 C^{tf} 相反.

kappa 系数的数值范围是 $-1\sim1$. 当 $kappa<0$ 时，表明一致性比由随机性引起的还差；当 $kappa>0$ 时，数值越大表明一致性越好.

定义 11. 基分类器之间基于 *kappa* 系数的相似性定义为

$$D_d(i,j)=\begin{cases}0, & kappa\leqslant 0.\\ kappa, & kappa>0.\end{cases} \tag{16}$$

在对基分类器进行聚类时，除了选择出具有差异性的基分类器集合之外，各个基分类器的分类准确率也是影响集成结果的十分重要的因素. *kappa* 系数只是对基分类器之间分类结果的相似性进行了度量，而在 2 个基分类器同时分类错误的情况下，其结果同样是相似的，所以没有考虑到其分类性能. 因此本文定义了基于 *kappa* 系数和分类准确率的相似性计算公式.

在剪枝集中有 N 个样本的情况下，当且仅当 H_i 和 H_j 对第 k 个样本同时分类正确时，$c_k=1$，其他情况下 $c_k=0$.

定义 12. 基分类器之间基于准确率的相似性定义为

$$D_a(i,j)=\frac{1}{N}\sum_{k=1}^{N}c_k. \tag{17}$$

定义 13. 基分类器之间基于 *kappa* 系数和分类准确率的相似性计算为

$$s(i,j)=\alpha D_d(i,j)+(1-\alpha)D_a(i,j), \tag{18}$$

其中，$\alpha\in[0,1]$，是基于 *kappa* 系数的相似性的权重.

2.3 选择性集成分类流程

在结合近邻传播聚类的选择性集成分类模型（selective ensemble classification integrated with affinity propagation clustering，SECIAPC）中，将 pathway 作为先验知识引入到基因微阵列数据的分类中，之后使用基于相交邻域粗糙集的基因选择模型对每个基因子集进行属性约简. 在基分类器选择阶段，采用基于 AP 聚类的选择性集成方法，集成分类模型流程如图 1 所示.

步骤 1. 基因微阵列数据与 pathway 知识相结合生成多个基因子集.

步骤 2. 采用基于相交邻域粗糙集的基因选择模型在每个基因子集中筛选出重要的基因.

步骤 3. 在每个约简之后的基因子集上训练 SVM 基分类器.

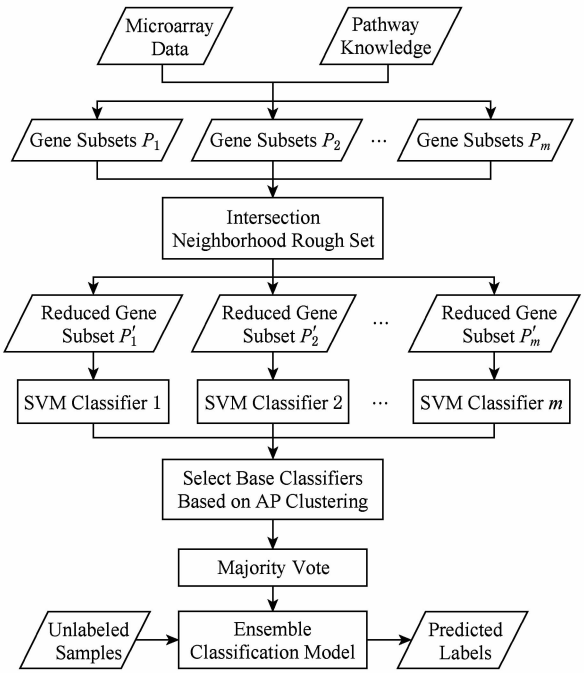

Fig. 1 Flow chart of classification model

图 1 分类模型流程

步骤 4. 全体基分类器对剪枝集进行分类，依据分类结果采用 AP 聚类算法对分类器进行聚类，并选择作为聚簇中心的基分类器进行集成.

步骤 5. 步骤 4 选择出的基分类器对测试集进行分类，使用简单多数投票法对分类结果进行融合.

3 实验结果与分析

3.1 数据集

由于拟南芥的生物学实验数据和基因注释信息相对比较丰富，所以常被用在不同类型的植物胁迫响应的研究中[26]. 本文采用拟南芥胁迫响应数据集和相关的 pathway 知识来对分类模型进行评估，从公开的生物信息网站 GEO① 下载了 Arabidopsis-Drought，Arabidopsis-Oxygen 和 Arabidopsis-TEV 三个数据集，它们分别对应于拟南芥干旱胁迫、氧气胁迫和烟草蚀纹病毒胁迫. 具体信息如表 2 所示，每个数据集都包含实验组和对照组，分别记为 *Class*1 和 *Class*2. 本文从数据库② 下载了拟南芥相关的 pathway 数据，其中包含 121 个 pathway 知识单元，对应形成 121 个基因子集，并且每个基因子集对应生成 1 个基分类器.

① http://www.ncbi.nlm.nih.gov/geo/

② https://www.arabidopsis.org

Table 2　Detail Information of Datasets

表 2　数据集的具体信息

Dataset	#*Gene*	#*Sample*	#*Class*1	#*Class*2
Arabidopsis-Drought	20 728	59	29	30
Arabidopsis-Oxygen	22 810	79	43	36
Arabidopsis-TEV	32 916	53	28	25

3.2　数据预处理及参数设置

为避免因基因表达数据数值范围不一致带来的分类误差,同时也便于邻域阈值的确定,所有数据被标准化到[−1,1].实验中设定相交邻域阈值 δ 的变化范围为 0.05～0.95,步长为 0.1,并且在每次实验中假设各个基因对应的阈值相同.AP 聚类中的阻尼系数 $\lambda=0.1$,基分类器的距离计算公式中权值 α 设置为从 0.1 到 0.9,步长为 0.1.随机选择 60%的样本作为训练样本,20%的样本作为剪枝样本,余下的作为测试样本.实验中采用支持向量机(support vector machine, SVM)作为分类算法生成基分类器,设定 SVM 分类器的核函数为具有较好分类性能和良好适应能力的 RBF 核函数 $K(x,y)=\exp(-\gamma\|x-y\|^2)$.

二分类问题的评价标准基于 4 种基础的指标:真阳性(true positives, *TP*)、假阳性(false positives, *FP*)、真阴性(true negatives, *TN*)和假阴性(false negatives, *FN*).

对分类模型选用准确率(accuracy, *ACC*)、敏感性(sensitivity, *SN*)、特异性(specificity, *SP*)和几何平均数(geometric mean, *G-mean*)四种标准评价其性能.准确率反映了对所有样本的分类准确率;敏感性和特异性分别反映对正、负类样本的分类准确率;几何平均数从全局衡量了对正、负类样本的分类性能.4 种评价标准的定义为

$$ACC=(TP+TN)/(TP+TN+FP+FN), \quad (19)$$

$$SN=TP/(TP+FN), \quad (20)$$

$$SP=TN/(TN+FP), \quad (21)$$

$$G\text{-}mean=\sqrt{SN\times SP}. \quad (22)$$

3.3　结果分析

在本文使用的 3 个数据集上,通过实验表明当 δ 分别取值为 0.45,0.55,0.65 时、α 分别取值为 0.1,0.2,0.8,0.9 时得到较好的分类结果,但是不同数据集上 δ 和 α 对分类结果的影响不同,所以在不同数据集上选取分类性能最好的 δ 和 α 值.为了更好地验证本文提出方法的优势,将该方法与单个的 SVM 分类器和 Random Subspace, Bagging, AdaboostM1 和 Stacking 这 4 种常用的集成分类方法进行比较.利用 Weka 软件对这 4 种方法的实现进行对比实验[27].Random Subspace 的基本原理是抽取不同的属性子集,对应每个属性子集形成一个训练集;Bagging 的原理是不断的有放回的抽取样本子集;AdaboostM1 方法在对样本进行抽样的过程中,对被错误分类的样本赋予更大的权重,来提升对这些样本的分类能力;Stacking 方法一般采用 2 层结构,将第 1 层的分类结果作为下一层的输入.本文提出的 SECIAPC 方法采用结合 pathway 知识的方式生成不同的基因子集,与以上 4 种集成方法不同的是使用 AP 聚类对基分类器进行选择,并对选择出的部分基分类器进行集成.所有的集成分类方法都采用 SVM 分类器进行集成,3 个数据集上分类性能的比较如表 3～5 所示.

Table 3　Classification Results on Arabidopsis-Drought Dataset

表 3　Arabidopsis-Drought 数据集上的分类结果

Classifier	*ACC*/%	*SN*/%	*SP*/%	*G-mean*
SVM	91.67	100.00	83.33	0.912 9
Random Subspace	91.67	100.00	83.33	0.912 9
Bagging	87.50	100.00	75.00	0.866 0
AdaboostM1	95.83	100.00	91.67	0.957 4
Stacking	91.67	100.00	83.33	0.912 9
SECIAPC	93.85	95.00	92.86	0.939 2

Table 4　Classification Results on Arabidopsis-Oxygen Dataset

表 4　Arabidopsis-Oxygen 数据集上的分类结果

Classifier	*ACC*/%	*SN*/%	*SP*/%	*G-mean*
SVM	87.50	100.00	66.67	0.816 5
Random Subspace	87.50	100.00	71.44	0.845 2
Bagging	96.88	100.00	91.67	0.957 4
AdaboostM1	96.88	100.00	91.67	0.957 4
Stacking	87.50	100.00	66.67	0.816 5
SECIAPC	98.33	100.00	96.25	0.981 1

Table 5　Classification Results on Arabidopsis-TEV Dataset

表 5　Arabidopsis-TEV 数据集上的分类结果

Classifier	*ACC*/%	*SN*/%	*SP*/%	*G-mean*
SVM	33.33	100.00	0.00	0.000 0
Random Subspace	33.33	100.00	0.00	0.000 0
Bagging	33.33	100.00	0.00	0.000 0
AdaboostM1	52.38	85.71	35.71	0.553 3
Stacking	33.33	100.00	0.00	0.000 0
SECIAPC	64.17	77.14	46.00	0.595 7

在 Arabidopsis-Drought 数据集上，SECIAPC 方法的准确率和几何平均数与其他方法中性能最好的 AdaboostM1 相当，对负类样本的分类准确率有所提高. 与其他3 种集成方法相比，SECIAPC 方法的性能大幅度提高，特别是在对负类的分类性能上提高了至少 10%. 在 Arabidopsis-Oxygen 数据集上，在准确率和几何平均数方面，SECIAPC 与 AdaboostM1 和 Bagging 这 2 种性能最好的方法相比提高了近 2%. 在 Arabidopsis-TEV 数据集上，其他方法对于正负类样本的分类能力极其不均衡. SECIAPC 方法在准确率上与其他方法比较至少提高了 12%，并且极大地提高了对负类样本的分类性能.

为了验证所选基因的生物学意义，使用 GO 术语① 检查部分被选择基因的功能注释，GO 数据库中运用 Term 的概念来显示基因特性，这些 GO Term 被划分为 3 类：细胞成分(cellular component，CC)、分子功能(molecular function，MF)和生物过程(biological process，BP)，分别对应表中的 3 种类别. 部分被选择的基因的详细注释如表 6 所示. 可以看出，这些基因都与胁迫响应相关.

Table 6 Gene Annotations in GO for Some Selected Genes

表 6 部分所选基因的 GO 基因注释信息

Gene	Category	Annotation
AT3G51910	MF	transcription factor activity, sequence-specific DNA binding
	BP	response to heat
	CC	nucleus
AT3G30775	MF	proline dehydrogenase activity
	BP	defense response to bacterium response to water deprivation
	CC	mitochondrion
AT1G64860	MF	DNA binding/protein binding
	BP	water transport Response to salt stress
	CC	chloroplast
AT1G55020	MF	linoleate 9S-lipoxygenase activity
	BP	defense response
	CC	plastid
AT5G27600	MF	long-chain fatty acid-CoA ligase activity
	BP	response to ozone
	CC	peroxisome

4 结束语

本文采用基于相交邻域粗糙集的基因约简模型作用在由 pathway 产生的信息表上，在引入先验知识的同时进行重要基因的选择. 使用基于 AP 聚类的选择性集成方法，在基分类器选择的过程中将分类和聚类 2 种机器学习技术进行结合，进一步提高了集成分类的性能.

在 3 个拟南芥胁迫响应相关数据集上的实验结果表明，提出的 SECIAPC 方法能够得出较好的分类结果. 同时与单分类器、传统集成方法的分类结果进行对比，其分类性能有很大的提升. 由于基因微阵列数据的高维属性，整个集成分类模型的分类效率有待进一步提高.

① http://www.geneontology.org/

参 考 文 献

[1] Maji P, Garai P. On fuzzy-rough attribute selection: Criteria of max-dependency, max-relevance, min-redundancy, and max-significance [J]. Applied Soft Computing, 2013, 13(9): 3968-3980

[2] Maji P, Paul S. Rough-fuzzy clustering for grouping functionally similar genes from microarray data [J]. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2013, 10(2): 286-299

[3] Landesfeind M, Kaever A, Feussner K, et al. Integrative study of arabidopsis thaliana metabolomic and transcriptomic data with the interactive marvis-graph software [J]. PeerJ, 2014, 2: Article No. e239

[4] Rokach L. Ensemble-based classifiers [J]. Artificial Intelligence Review, 2010, 33(1/2): 1-39

[5] Chen Xi, Wang Lily. Integrating biological knowledge with gene expression profiles for survival prediction of cancer [J]. Journal of Computational Biology, 2009, 16(2): 265-278

[6] Bandyopadhyay N, Kahveci T, Goodison S, et al. Pathway-Based Feature Selection Algorithm for Cancer Microarray Data [J]. Advances in Bioinformatics, 2010, 2009(5235): 103-118.

[7] Gatza M L, Lucas J E, Barry W T, et al. A pathway-based classification of human breast cancer [J]. Proc of the National Academy of Sciences, 2010, 107(15): 6994-6999

[8] Zhang Lijuan, Li Zhoujun. Gene selection for cancer classification in microarray data [J]. Journal of Computer Research and Development, 2009, 46(5): 794-802 (in Chinese)
(张丽娟，李舟军. 微阵列数据癌症分类问题中的基因选择[J]. 计算机研究与发展，2009，46(5)：794-802)

[9] Lin Chen, Chen Wenqiang, Qiu Cheng, et al. LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy [J]. Neurocomputing, 2014, 123: 424-435

[10] Zhang Huaxiang, Cao Linlin. A spectral clustering based ensemble pruning approach [J]. Neurocomputing, 2014, 139: 289-297

[11] Krawczyk B. Forming ensembles of soft one-class classifiers with weighted bagging [J]. New Generation Computing, 2015, 33(4): 449-466

[12] Ding Xiangwu, Guo Tao, Wang Mei, et al. A clustering algorithm for large-scale categorical data and its parallel implementation [J]. Journal of Computer Research and Development, 2016, 53(5): 1063-1071 (in Chinese)
(丁祥武，郭涛，王梅，等. 一种大规模分类数据聚类算法及其并行实现[J]. 计算机研究与发展，2016，53(5)：1063-1071)

[13] Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: New developments in KEGG [J]. Nucleic Acids Research, 2006, 34(Suppl 1): D354-D357

[14] Pawlak Z. Rough sets [J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356

[15] Yao Yiyu, Lin Tsauyoung. Generalization of rough sets using modal logic [J]. Intelligent Automation and Soft Computing, 1996, 2(2): 103-120

[16] Meng Jun, Zhang Jing, Li Rui, et al. Gene selection using rough set based on neighborhood for the analysis of plant stress response [J]. Applied Soft Computing, 2014, 25: 51-63

[17] Pawlak Z. Imprecise categories, approximations and rough sets [G] //Rough Sets. Beijing: Springer, 1991: 9-32

[18] Ni Zhiwei, Zhang Chen, Ni Liping. Haze forecast method of selective ensemble based on glowworm swarm optimization algorithm [J]. Pattern Recognition and Artificial Intelligence, 2016, 29(2): 143-153 (in Chinese)
(倪志伟，张琛，倪丽萍. 基于萤火虫群优化算法的选择性集成雾霾天气预测方法[J]. 模式识别与人工智能，2016，29(2)：143-153)

[19] Tang Chao, Wang Wenjian, Li Wei, et al. Human action recognition algorithm based on selective ensemble rotation forest [J]. Pattern Recognition and Artificial Intelligence, 2016, 29(4): 313-321 (in Chinese)
(唐超，王文剑，李伟，等. 基于选择性集成旋转森林的人体行为识别算法[J]. 模式识别与人工智能，2016，29(4)：313-321)

[20] Zhou Zhihua, Wu Jianxin, Tang Wei. Ensembling neural networks: Many could be better than all [J]. Artificial Intelligence, 2002, 137(1): 239-263

[21] Zhou Zhihua, Tang Wei. Selective ensemble of decision trees [C] //Proc of the 9th Int Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. Berlin: Springer, 2003: 476-483

[22] Frey B J, Dueck D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972-976

[23] Leone M, Weigt M. Clustering by soft-constraint affinity propagation: Applications to gene-expression data [J]. Bioinformatics, 2007, 23(20): 2708-2715

[24] Cohen J. A coefficient of agreement for nominal scales [J]. Educational and Psychological Measurement, 1960, 20(1): 37-46

[25] Ben-David A. Comparison of classification accuracy using Cohen's weighted kappa [J]. Expert Systems with Applications, 2008, 34(2): 825-832

[26] Karim S. Exploring Plant Tolerance to Biotic and Abiotic Stresses [D]. Uppsala, Sweden: Swedish University of Agricultural Sciences, 2007

[27] 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

Meng Jun, born in 1964. PhD, professor. Member of CCF. Her main research interests include data mining, machine learning, and bioinformatics.

Zhang Jing, born in 1992. Master candidate. Her main research interests include data mining and machine learning (1500106500@qq.com).

Jiang Dingling, born in 1994. Master candidate. Her main research interests include gene selection and function prediction (645949890@qq.cm).

He Xinyu, born in 1983. PhD candidate. Her main research interests include text mining for biomedical information extraction (444801898@qq.com).

Li Lishuang, born in 1967. PhD, professor. Her main research interests include text mining, natural language processing and machine translation (lils@dlut.edu.cn).