第 57 卷　第 5 期
2021 年 9 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 57, No. 5
Sept., 2021

DOI:10.13232/j.cnki.jnju.2021.05.006

基于机器学习的自发性早产生物标记物发现

李苓玉[1],刘治平[1,2*]

(1.山东大学控制科学与工程学院,济南,250061;2.山东大学智能医学工程研究中心,济南,250061)

摘　要:近年来,基于基因表达微阵列数据的生物标记物示性基因的识别在生物信息学领域备受关注.自发性早产(Spontaneous Preterm Birth,SPTB)生物标记物的成功鉴定有利于降低孕妇早产的风险,具有重要的研究价值.提出一种从公开基因表达数据中识别SPTB生物标记物的方法.首先,从公开数据库下载SPTB的基因表达数据,运用支持向量机-递归特征消除(Support Vector Machine-Recursive Feature Elimination,SVM-RFE)进行基因特征选择,并与其他机器学习与特征选择方法(AdaBoost-RFE,Neural Network-RFE,Radom Forest-RFE和K-Nearest Neighbor-RFE)进行比较,利用准确性、精确度、灵敏度、特异度、F-测度和AUC等指标,对分类效果进行评价.然后,将SVM-RFE排名靠前的基因与其他方法排名靠前的基因取交集,以此作为识别出的SPTB生物标记物.接着,通过聚类分析、相关性分析和功能富集分析对识别的生物标记物进行初步的鉴定.最后,构建SVM分类器,在独立数据集上对所识别的生物标记物进行验证.结果表明,提出的机器学习方法对于SPTB生物标记物的发现是有效的.该方法能在孕妇产前无创检测患有SPTB的可能,减少对人工鉴别的依赖,降低孕妇早产风险.

关键词:生物标记物,自发性早产,机器学习,特征选择,生物信息学

中图分类号:Q811.4　　**文献标志码:**A

Discovery of spontaneous preterm birth biomarkers based on machine learning

Li Lingyu[1], Liu Zhiping[1,2*]

(1. School of Control Science and Engineering, Shandong University, Ji′nan, 250061, China;
2. Center for Intelligent Medicine, Shandong University, Ji′nan, 250061, China)

Abstract: In recent years, the identification of descriptive genes of biomarkers based on gene expression microarray data has attracted much attention in the field of bioinformatics. The successful identification of spontaneous preterm birth (SPTB) biomarkers is conducive to reducing the risk of preterm birth in pregnant women and has important research value. In this paper, we propose a method for identifying biomarkers of SPTB from publically available gene expression data. First, we download SPTB gene expression data from public databases, use SVM-RFE (Support Vector Machine-Recursive Feature Elimination) for gene feature selection, and compares it with other machine learning and feature selection methods, namely AdaBoost-RFE, Neural Network-RFE, Radom Forest-RFE and K-Nearest Neighbor-RFE. With the help of accuracy, precision, sensitivity, specificity, F-measure and AUC, the classification performances are evaluated. Then, the top-ranked genes of SVM-RFE are intersected with the top-ranked genes of the other four methods as the identified SPTB biomarkers, which are sequentially justified by cluster analysis, correlation analysis and functional enrichment analysis. Finally, an SVM classifier is constructed to verify the identified biomarkers on an independent dataset. The results show that machine learning methods are effective for SPTB biomarkers discovery. The proposed method can realize the possibility of SPTB non-

基金项目:国家重点研发计划(2020YFA0712402),国家自然科学基金(61973190)

收稿日期:2021－05－22

*通讯联系人,E-mail:zpliu@sdu.edu.cn

invasively before women's pregnancy, reduce the dependence on artificial identification and the risk of premature delivery of pregnant women.

Key words: biomarkers, spontaneous preterm birth, machine learning, feature selection, bioinformatics

早产是指妊娠满28周至37周期间分娩[1]. 作为医学领域中的一种极为重要、异常复杂而又十分常见的妊娠并发症,早产每年导致310万新生儿死亡[2],早产儿的患后遗症率、发病率和死亡率均显著高于足月产儿,已成为全世界妇产科领域的一项重大挑战[3]. 在医学上,约1/3的早产由孕妇或胎儿意外造成,其他2/3均为自发性早产(Spontaneous Preterm Birth, SPTB),包括未足月自发性分娩(Spontaneous Preterm Labor, SPTL)和未足月胎膜早破(Preterm Premature Rupture of the Membranes, PPROM)[4]. SPTB的病因及其发病机制的研究已成为围产医学领域中亟待解决的难题[5],目前尚无有效的诊断方法检测SPTB或提前使用干预方法来延长分娩过程并将妊娠延长至足月[6]. SPTB生物分子标记物的鉴定对SPTB的早期检测有重要意义[7]. 近年来,人们越来越重视基因表达微阵列数据在早期诊断中的应用[8]. 然而,微阵列通常只用少量样品来测量大量基因,所以在选择特征基因作为生物标记物方面存在着巨大的挑战[9]. 如何使用机器学习方法识别基因表达数据中的特征基因作为分类表型的生物标记物,已成为一个十分重要的研究课题[10].

本文的主要贡献:

(1)将医学中的生物标记物发现问题转化为机器学习中的特征选择问题. 基于机器学习方法发现SPTB生物标记物,从特征选择的角度挑选潜在的与SPTB相关的基因.

(2)提出一种稳健的生物标记物鉴定方法. 将支持向量机-递归特征消除(Support Vector Machine - Recursive Feature Elimination, SVM-RFE)排名靠前的特征基因与其他方法(AB-RFE, NN-RFE, RF-RFE和KNN-RFE)排名靠前的特征基因进行整合作为SPTB生物标记物,通过聚类分析、相关性分析、功能富集分析和外部独立数据集验证对识别的生物标记物进行鉴定.

1 相关工作

早期早产的研究以假说为导向[11],通过传统生物学及化学试验的方法对其进行深入研究. 这些方法实施起来既复杂又繁琐,研究过程犹如大海捞针,取得的成效也微乎其微. 后来早产的研究依赖于对照试验[12]. 首先,分别选取一定数量的早产临产(Preterm with Labor, PL)和早产未临产(Preterm No Labor, PNL)孕妇作为研究组,再分别选取一定数量的足月临产(Term with Labor, TL)和足月未临产(Term No Labor, TNL)孕妇作为对照组. 然后,应诊断的需求从孕妇体内切取、钳取或穿刺等取出子宫肌层组织,采用实时荧光定量PCR(qRT-PCR)等方法检测各组子宫肌层中某些分子的表达水平. 最后,对检测结果进行统计学分析. 该方法需要进行临床取样,实验操作复杂,不仅存在高成本、低发现的问题,对病人的身体和胎儿安全也造成了威胁[13].

当今世界,生物医学数据进入到了多维度大数据时代,人们希望能从数据中挖掘出更多的知识[14]. 微阵列和RNA-seq等高通量组学技术为发现SPTB生物标记物提供了大量的数据[15]. 现代的研究以数据为导向,根据基因的分子特征进行特征选择,可以快速简便地将基因组中的遗传差异与表现型联系起来,比传统的方法更有效[16]. 该方法旨在挖掘组学数据,程序实现简单,可实现无创伤诊断.

近年来,人们越来越关注微阵列技术在早产儿诊断中的应用[8]. 在生物信息和机器学习领域,基于微阵列数据选择某些指示性基因作为生物标记物已成为研究的热门[17],基因医学已经逐渐成为现实. 医学中生物标记物发现问题等价于机器学习中的特征选择问题,因此,本研究希望通过特征选择方法确定与SPTB相关的最佳基因子集,并以此来区分不同表型状态的SPTB.

2　数据与方法

2.1　数　据　SPTB基因表达数据来自NCBI GEO数据库(GSE59491，GSE73685). GSE59491数据集包含165名无症状孕妇的326个样本，包括98个SPTB样本和228个足月产样本[18]. 对数据预进行处理之后，每个样本包含24478个基因. GSE73685数据集包含来自八个组织的183个样本，经过数据预处理，选取与SPTB和足月产相关的17个样本，每个样本包含20909个基因[19]. 表1列出了数据集的详细信息，括号内的数字表示样本量.

表1　实验使用的两个数据集

Table 1　Two datasets used in experiments

数据集	样本个数	基因个数	样本类型	样本表型	参考文献
GSE59491	326	24478	母体全血	足月产(228)/早产(98)	[18]
GSE73685	17	20909	母体全血	足月产(12)/早产(5)	[19]

2.2　SVM-RFE方法　1995年Cortes and Vapnik[20]提出软边距的非线性SVM(Support Vector Machine)并将其应用于手写字符识别问题，为SVM在各领域的应用提供了参考. 作为二分类模型的一种，SVM是定义在特征空间上使得两类间隔最大化的线性分类器，被广泛用于模式识别、数据挖掘与机器学习等领域.

考虑二分类问题：假设来自n个样本的观测值(X_i, y_i)，$i=1,2,\cdots,n$独立同分布，其中$X_i=(x_{i1}, x_{i2}, \cdots, x_{ip})^T$为第$i$个样本的一个$p$维的基因表达数据，$y_i$是取值为+1或−1的标签变量. 当数据线性可分时，SVM寻找最优的分类超平面$\boldsymbol{\omega}^T\cdot\boldsymbol{X}_i+b=0$将两类的样本完全分开，需要求解的优化问题为：

$$\min\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{n}\xi_i$$
$$s.t.\quad y_i\left(\boldsymbol{\omega}^T\cdot\boldsymbol{X}_i+b\right)\geqslant 1-\xi_i, i=1,2,\cdots,n$$
$$\xi_i\geqslant 0, i=1,2,\cdots,n$$

其中，$\boldsymbol{\omega}=(\omega_1,\omega_2,\cdots,\omega_p)^T$代表超平面的法向量(特征的权重向量)，$\xi_i$为松弛变量，$b$代表偏置，$C$为惩罚参数，用于在最小化误差和最大化间隔之间进行平衡.

原始问题可以转化为如下的对偶问题：

$$\min\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(\boldsymbol{X}_i^T\boldsymbol{X}_j\right)-\sum_{i=1}^{n}\alpha_i$$
$$s.t.\quad \sum_{i=1}^{n}\alpha_i y_i=0, i=1,2,\cdots,n$$
$$0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,n$$

其中，α_i为拉格朗日乘子.

由于基因表达微阵列数据训练样本的特征信息量巨大，使用SVM建立预测模型容易因维度高而产生过拟合现象，致使其分类精度不高、泛化能力不强；同时，并不是所有的特征属性都能对预测模型起正面作用，因此对训练数据进行特征选择至关重要[21]. SVM-RFE算法最早由Guyon et al[22]提出并被应用于癌症的分类和预测等研究工作，它可以通过降低数据结构的复杂性将显著的特征变量识别为新的测试实例来解决这一问题. SVM-RFE根据SVM在训练时生成的权向量$\boldsymbol{\omega}$来构造特征排序系数，每次迭代去掉一个排序系数最小的特征，同时保留具有显著影响的特征，最终得到所有特征属性的递减排序. 具体的实现过程如算法1所示，其算法复杂度为$O(d_l N_s)$，即输入向量的维度乘以支持向量的个数[23].

算法1　SVM-RFE算法

输入：训练样本$(\boldsymbol{X}_i, y_i)$, $y_i\in\{0,1\}$, $i=1,2,\cdots,n$.

输出：特征排序集R.

Step1. 初始化原始特征集$S=\{1,2,\cdots,p\}$，特征排序集$R=[\ ]$;

Step2. 循环以下过程直至$S=[\ ]$.

①获取带候选特征集的训练样本$\boldsymbol{X}_i=\boldsymbol{X}_i(:,S)$;

②根据$\min\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(\boldsymbol{X}_i^T\cdot\boldsymbol{X}_j\right)-\sum_{i=1}^{n}\alpha_i$训练SVM分类器;

③计算权重$\boldsymbol{\omega}=\sum_{i=1}^{n}\alpha_i y_i\boldsymbol{X}_i$;

④计算排序准则得分$c_i=\omega_i^2$;

⑤找出排序得分最小(最小权重)的特征$p=\arg\min_i c_i$;

⑥更新特征排序集$R=[p,R]$;

⑦在S中去除此特征$S=S/p$.

2.3　算法复杂度分析　进一步分析比较SVM-RFE算法和其他四种机器学习算法的复杂度，结果如表2所示.

表 2 SVM-RFE算法和其他四种机器学习算法的复杂度

Table 2 The complexity of SVM-RFE and other four machine learning algorithms

算法	复杂度	符号解释
SVM-RFE	$O(d_l N_s)$	d_l:输入向量的维度;N_s:支持向量的个数
AdaBoost-RFE (AB-RFE)	$O(Npn)$	N:数据集中的数据点个数;p:数据集中的特征个数;n:模型中使用的估计器的个数
Neural Network-RFE (NN-RFE)	$O(Tf)$	T:弱学习器的个数;f:弱学习器的运行时间
Random Forest-RFE (RF-RFE)	$O(pN\lg N)$	N:数据集中的样本数;p:数据集中的特征个数
K-Nearest Neighbor-RFE (KNN-RFE)	$O(Np)$	N:数据集中的样本数;p:数据集中的特征个数

2.4 分类评价指标 为了评价SVM-RFE的性能,利用机器学习分类中常用的指标进行评价,主要包括:混淆矩阵、准确率、精确率、灵敏度、特异度、F-测度和AUC[24]. 对于二分类问题,SVM分类器的输出为预测样本为正样本的概率,取值范围为$[0,1]$,通过设置阈值进行分类,概率高于阈值的分为正样本,概率低于阈值的分为负样本. 因此,分类结果有四种情况,常以混淆矩阵的形式给出,如表3所示.

表 3 混淆矩阵

Table 3 Confusion matrix

		预测分类	
		正样本	负样本
实际分类	正样本	TP (True Positive)	TN (True Negative)
	负样本	FP (False Positive)	FN (False Negative)

表3列举的四种情况,一方面对应如下四个比例:表示正样本中被预测对的比例的真阳性率(又称灵敏度)、表示负样本中被预测对的比例的真阴性率(又称特异度)、表示负样本中被预测错的比例的假阳性率以及表示正样本中被预测错的比例的假阴性率;另一方面诱导出其他评价指标:准确率(又称精度)、召回率(又称真正率)、表示所预测为正样本中被预测对的比例的精确率、兼顾精确率和召回率的F-测度. F-测度是精确率和召回率的加权调和平均,其值越高说明模型的性能越好. 各评价指标的计算公式如表4所示.

受试者工作特征曲线(Receiver Operating Characteristic curve, ROC curve)以假阳性率$\left(FPR/\left(1-Sp\right)\right)$为横坐标,以真阳性率$\left(TPR/Sn\right)$为纵坐标,通过设定$[0,1]$的连续分类阈值,可以反映模型在选取不同分类阈值时灵敏度和特异性度的趋势走向,曲线越靠近左上角表示分类效果越好. 曲线下面积(Area Under Curve,AUC)是指ROC曲线下的面积,其取值范围为$[0.5,1]$,越接近1表示模型准确性越高,即分类效果越好.

表 4 评价指标及计算公式

Table 4 Evaluation index and calculation formulas

评价指标	符号	公式
真阳性率/灵敏度	TPR/Sn	$\frac{TP}{TP+FN}$
真阴性率/特异度	TNR/Sp	$\frac{TN}{TN+FP}$
假阳性率	$FPR/(1-Sp)$	$\frac{FP}{FP+TN}$
假阴性率	FNR	$\frac{FN}{FN+TP}$
准确率/精度	ACC	$\frac{TN+TP}{TN+TP+FP+FN}$
精确率	Pre	$\frac{TP}{TP+FP}$
召回率/真正率	$Recall$	$\frac{TP}{TP+FN}$
F-测度	F-measure	$\frac{2\times Pre\times Recall}{Pre+Recall}$

3 结果与讨论

3.1 特征选择结果 使用经验贝叶斯(Empirical Bayes, EB)方法对GSE59491数据集中的基因进行差异表达分析,得到694个具有显著性差异($P.\mathrm{adj}<0.05$)的基因,并将其作为候选集. 应

用SVM-RFE算法剔除冗余特征和不相关者获得特征属性排序，即基因重要性的排序文件．选用SVM-RFE排名前50位的基因进行后续分析，同时，与其他四种机器学习与特征选择方法（AB-RFE，NN-RFE，RF-RFE和KNN-RFE）的效果进行比较．

从数据集中随机选70%样本组成训练集，剩下的30%样本作为训练集，分别使用SVM模型进行拟合．对于训练集，使用tune.svm方法，取样进行十折交叉验证，寻找最佳的gamma值和cost值．此外，核函数选择径向基核函数，以最低程度避免过拟合．在该过程中对选自不同模型的RFE的特征子集获得SVM分类器在不同参数下的正确率（如图1所示），选取精确度最高值对应的参数作为最优参数（精确度相同的选取其中一组参数即可）．

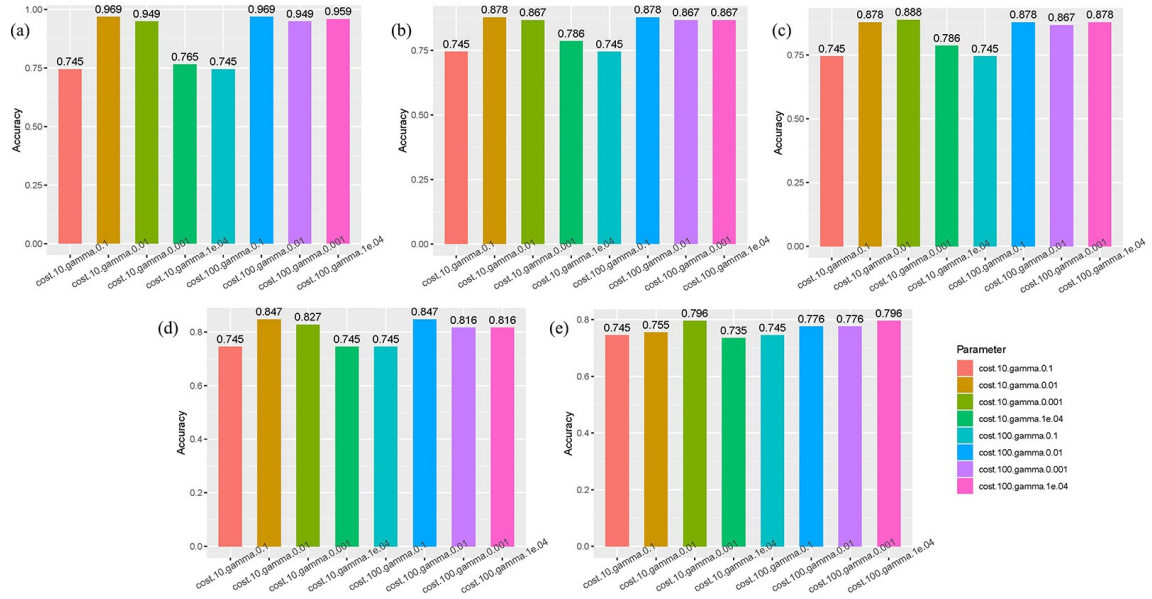

图1　不同参数下模型的精度：(a) SVM-RFE；(b) AB-RFE；(c) NN-RFE；(d) RF-RFE；(e) KNN-RFE

Fig.1　Accuracy of models with different parameters：(a) SVM-RFE，(b) AB-RFE，(c) NN-RFE，(d) RF-RFE，(e) KNN-RFE

在测试集数据上使用最佳参数构建SVM模型进行分类，并与上述四种机器学习与特征选择方法的测试效果进行比较，所得ROC曲线如图2a所示，相关参数评估结果如图2b所示．其中，图2a的图例注释和图2b的坐标横轴均依*AUC*对机器学习与特征选择方法降序排列．

由图可见，无论是*AUC*，还是准确率、精确率、灵敏度、特异度和*F*-测度，SVM-RFE方法在测试数据集上的分类效果都明显优于其他四种方法．特别地，五种机器学习与特征选择方法的*AUC*均高于0.800，其中SVM-RFE方法的*AUC*最高（*AUC*=0.998），KNN-RFE方法的*AUC*最低（*AUC*=0.810）．

3.2　生物标记物鉴定　将SVM-RFE，AB-RFE，NN-RFE，RF-RFE和KNN-RFE得到的排名前50的基因，先两两取交集，再两两取并集，发现有54个基因在五种机器学习与特征选择方法中均排名靠前，同时也都是EB找出来的基因．对于两者重叠部分的54个基因（PLEC，RUVBL2，PCDHGA12，MRPL51，VNN1，GLYR1，SAP30L，VAMP2，SGSH，FNBP1L，TARS，KAT5，PRKAG1，MCM2，PVALB，MIR3117，SLC12A4，POLR3K，ZC3H3，LOC101927699，SUOX，CASC20，LINC01268，PLA2G4C，ST7，SPRTN，SEPSECS-AS1，CDKN2A-DT，LINC-01427，TOE1，CSTF1，OR51A7，FAM50B，ZNF284，SHROOM4，ZNF605，RSL1D1，LRRC-41，C1orf123，ASRGL1，GOLGA7，ACADVL，

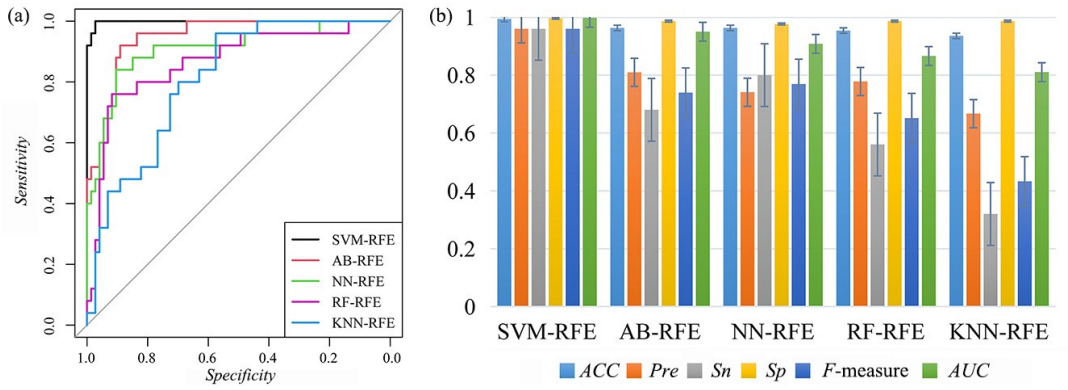

图 2 (a)不同 RFE 方法的 ROC 曲线;(b)不同 RFE 方法的分类效果

Fig.2 (a) ROC curves of different RFE methods, (b) classification effects of different RFE methods

ADAMTS9-AS2, NANOG, LOC84214, RAB26, FADS2, FASTKD5, MIR601, ATP8A2, SYT-15, AQP12B, MFSD4A, MAP3K9)在统计上明显更有意义,生物学方面也更重要,可以作为SPTB生物标记物用于后续分析.

首先,对鉴定的54个生物标记物绘制基因表达热图(如图3所示). 绘制热图的同时进行了聚类,由图可见,热图整体从左到右可以分成两大类,正负样本比例约为1∶3,与数据集的实际分类结果相符,证明了所用方法的有效性.

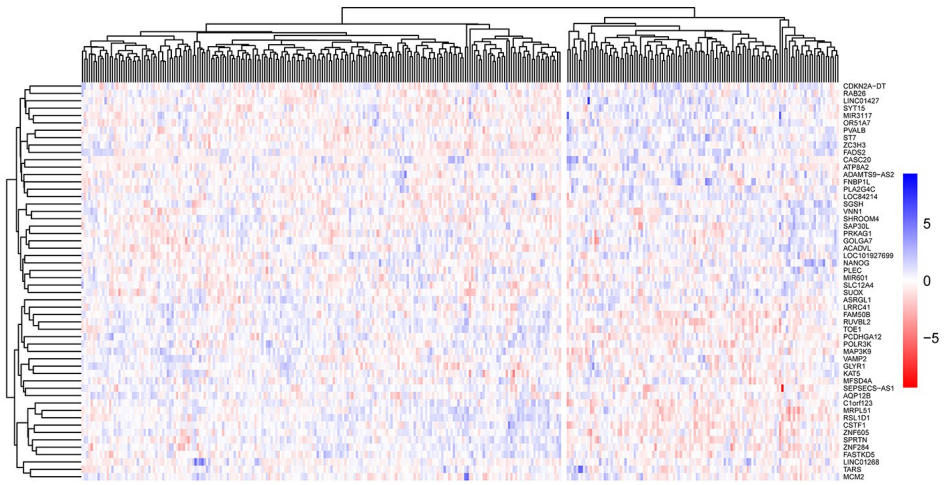

图 3 54个生物标记物的热图

Fig. 3 Heatmap of the 54 biomarkers

计算鉴定54个生物标记物之间的相关性. 挑选相关系数的绝对值较大的基因对,绘制其网络关系图,如图4所示,其中红色代表上调基因,绿色代表下调基因. 由图可见,49个节点由206条边相互连接,每个节点代表一个基因. 这表明被识别的生物标记物不是孤立存在的,而是形成了一个相互连通的网络,更具有生物学意义.

最后,对鉴定的54个生物标记物做生物过程(Biological Process, BP)通路(Gene Ontology, GO)的功能富集分析[25],如图5所示. 由图可见,富集的功能主要与质膜转运有关,这表明识别出的生物标记物具有显著的生物学功能.

3.3 独立数据集验证 对GSE73685数据集包含的20909个基因与上述54个生物标记物取交集,发现46个基因为共有基因. 应用这46个基因表达数据构建SVM分类器,在GSE59491数据集上进行训练,在GSE73685数据集上进行验证,分类结果如图6所示. *AUC*值达到0.883,该结果进一步说明所识别的生物标记物是有效的.

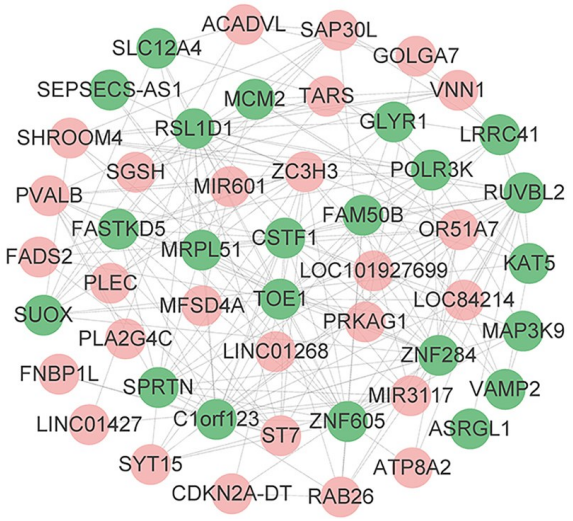

图4　生物标记物相关性的网络图

Fig. 4　Network of correlation in biomarkers

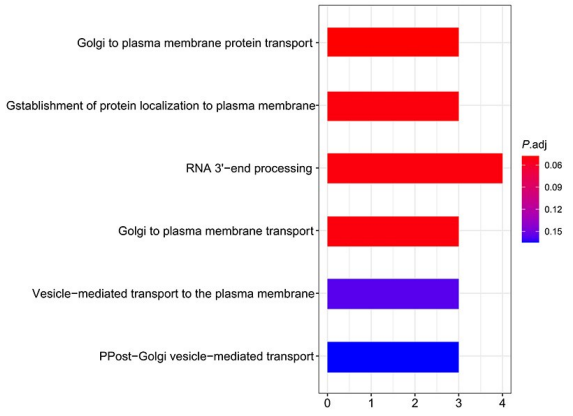

图5　功能富集分析的可视化

Fig. 5　Visualization of functional enrichment analysis

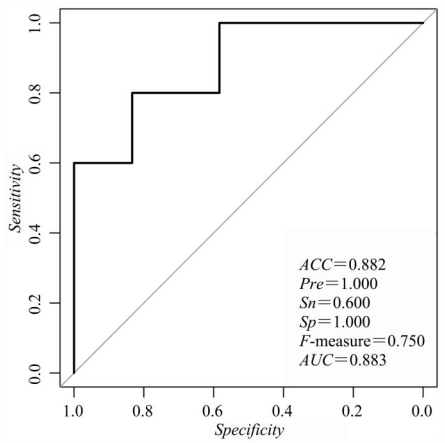

图6　独立数据集验证结果

Fig. 6　Verification results on independent dataset

4　结　论

本文运用机器学习方法挖掘基因表达数据中所蕴含的重要信息，有效识别SPTB生物标记物．首先得到差异表达基因，结合机器学习与特征选择算法进行特征选择．基于训练数据集，使用SVM分类器，通过交叉验证选取最优参数，应用敏感性、特异性、准确性、F-测度和AUC等指标评估模型的分类性能，证明了本文提出的基于机器学习方法的可行性．然后对五种机器学习方法得到的特征基因取交集，将其所得作为识别出的SPTB生物标记物，该方法挑选的生物标记物具有稳健性．最后通过对分类器的学习与训练，在独立数据集上进行验证（即外部验证），表明了所选生物标记物的有效性．

致　谢　感谢实验室的其他成员对本研究的帮助．

参考文献

[1] 中华医学会妇产科学分会产科学组．早产的临床诊断与治疗推荐指南（草案）．中华妇产科杂志，2007，42(7)：498－500.

[2] 冉雨鑫，尹楠林，漆洪波．早产发病机制的新进展．实用妇产科杂志，2019，35(7)：481－483.

[3] Aung M T，Yu Y F，Ferguson K K，et al. Prediction and associations of preterm birth and its subtypes with eicosanoid enzymatic pathways and inflammatory markers. Scientific Reports，2019，9(1)：17049.

[4] Vora B，Wang A L，Kosti I，et al. Meta-analysis of maternal and fetal transcriptomic data elucidates the role of adaptive and innate immunity in preterm birth. Frontiers in Immunology，2018(9)：993.

[5] Zhang G，Feenstra B，Bacelis J，et al. Genetic associations with gestational duration and spontaneous preterm birth. New England Journal of Medicine，2017，377(12)：1156－1167.

[6] Fettweis J M，Serrano M G，Brooks J P，et al. The vaginal microbiome and preterm birth. Nature Medicine，2019，25(6)：1012－1021.

[7] Li L Y，Liu Z P. Biomarker discovery for predicting

spontaneous preterm birth from gene expression data by regularized logistic regression. Computational and Structural Biotechnology Journal, 2020(18): 3434－3446.

[8] Chien C W, Lo Y S, Wu H Y, et al. Transcriptomic and proteomic profiling of human mesenchymal stem cell derived from umbilical cord in the study of preterm birth. Proteomics Clinical Applications, 2020, 14(1): 1900024.

[9] O'Brien C M. Statistical learning with sparsity: The lasso and generalizations. International Statistical Review, 2016, 84(1): 156－157.

[10] Ge Y S, He Z X, Xiang Y Q, et al. The identification of key genes in nasopharyngeal carcinoma by bioinformatics analysis of high - throughput data. Molecular Biology Reports, 2019, 46(3): 2829－2840.

[11] 杨孜, 郭艳军. 早期早产的研究进展. 实用妇产科杂志, 2005, 21(11): 652－654.

[12] 张利宏, 陈诚, 王琳等. 分娩相关基因在早产和足月产子宫肌层的差异表达及其意义. 第三军医大学学报, 2010, 32(10): 1083－1086. (Zhang L H, Chen C, Wang L, et al. Differential expressions and clinical significance of labour-associated genes in preterm labour and term labour myometrium. Acta Academiae Medicinae Militaris Tertiae, 2010, 32(10): 1083－1086.)

[13] 杨慧霞, 郭战坤. 早产的研究进展(一). 中华医学信息导报, 2007, 22(24): 20.

[14] 谢晟堃. 生物医学大数据的现状与展望. 科技风, 2017(20): 20.

[15] Liu Z P. Identifying network - based biomarkers of complex diseases from high - throughput data. Biomarkers in Medicine, 2016, 10(6): 633－650.

[16] 张国庆, 李亦学, 王泽峰等. 生物医学大数据发展的新挑战与趋势. 中国科学院院刊, 2018, 33(8): 853－860. (Zhang G Q, Li Y X, Wang Z F, et al. New challenges and trends in bio - med big data. Bulletin of the Chinese Academy of Sciences, 2018, 33(8): 853－860.)

[17] Benoist G. Prediction of preterm delivery in symptomatic women (preterm labor). Journal de Gynecologie, Obstetrique et Biologie de la Reproduction, 2016, 45(10): 1346－1363.

[18] Heng Y J, Pennell C E, Mcdonald S W, et al. Maternal whole blood gene expression at 18 and 28 weeks of gestation associated with spontaneous preterm birth in asymptomatic women. PLoS One, 2016, 11(6): e0155191.

[19] Bukowski R, Sadovsky Y, Goodarzi H, et al. Onset of human preterm and term birth is related to unique inflammatory transcriptome profiles at the maternal fetal interface. PeerJ, 2017(5): e3685.

[20] Cortes C, Vapnik V. Support - vector networks. Machine Learning, 1995, 20(3): 273－297.

[21] 朱诗生, 汪昕蓉, 毛礼厅等. 肿瘤类疾病的过度与错误医疗检查控制机制与模型的研究. 计算机应用研究, 2019, 36(5): 1428－1432. (Zhu S S, Wang X R, Mao L T, et al. Study on evaluation mechanism of excessive treatment and misdiagnosis of tumor diseases. Application Research of Computers, 2019, 36(5): 1428－1432.)

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

[23] Burges C J C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121－167.

[24] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1997, 30(7): 1145－1159.

[25] 尤元海, 张建中. 基因表达谱芯片的数据挖掘. 中国生物工程杂志, 2009, 29(10): 87－91. (You Y H, Zhang J Z. Data mining from microarray gene expression profile. China Biotechnology, 2009, 29(10): 87－91.)

(责任编辑 杨可盛)