·论 著·

三种统计分析方法在基因表达谱数据中的比较研究

史晓雯, 肖纯, 刘芸良, 刘艳
哈尔滨医科大学卫生统计学教研室,黑龙江 哈尔滨 150081

摘要: 目的 比较 SCAD-支持向量机、支持向量机和弹性网三种方法对基因表达谱数据的变量筛选和预测判别能力。 **方法** 根据设置的参数生成不同条件的基因表达谱模拟数据和实际数据,利用 FDR、一致性错误率和 ROC 曲线下面积(AUC 值)从三个方面评价三种方法的变量筛选和预测判别能力。 **结果** 模拟实验显示在差异变量数不变的情况下,随着差异变量间相关系数的增加,三种方法建立模型的变量筛选和预测判别能力均提高;当差异变量间相关系数不变时,随着差异变量数目的增加,SCAD-支持向量机和弹性网方法的变量筛选和预测判别能力均呈下降趋势,而支持向量机呈现提高趋势。 **结论** SCAD-支持向量机不仅改善了支持向量机不能直接进行变量筛选的不足同时提高了模型的精度以及判别的准确性。综合来看 SCAD-支持向量机的变量筛选和预测判别能力更优,处理变量间有高度相关性的基因表达谱数据时可以获得更高的预测精度和更稳定的模型估计。

关键词: SCAD-支持向量机; 弹性网; 一致性错误率; ROC 曲线下面积
中图分类号:R195.1 **文献标识码:**A **文章编号:**1006-3110(2018)02-0155-05 DOI:10.3969/j.issn.1006-3110.2018.02.008

Comparison of three statistical methods based on gene expression profile data

SHI Xiao-wen, XIAO Chun, LIU Yun-liang, LIU Yan
Department of Medical Statistics, Harbin Medical Univeristy, Harbin, Heilongjiang 150081, *China*
Corresponding author: LIU Yan, E-mail: liuyan@ems.hrbmu.edu.cn

Abstract: Objective To compare the variable selection and predictive ability of gene expression profile data among the three methods, including smoothly clipped absolute deviation-support vector machine (SCAD-SVM), support vector machine (SVM) and Elastic Net. **Methods** Different conditions of gene expression profile simulation data and the actual data of colon cancer were generated according to the set of parameters. The false discovery rate (FDR), the consistency error rate and the area under

基金项目:国家自然科学基金(81172741;30972537)
作者简介:史晓雯(1991-),女,黑龙江省齐齐哈尔市人,硕士研究生,研究方向:流行病学与卫生统计学。
通信作者:刘艳,E-mail:liuyan@ ems. hrbmu. edu. cn。

境交互也成为致病的重要因素。为了更有效地检测统计方法的性能,仍需要对目前的 SNPs 数据仿真方法进行改进,主要涉及仿真性能、致病位点设置的多样性、运行时间等方面。

参考文献

[1] Olivier M. A haplotype map of the human genome[J]. Physiol Genomics, 2003, 13(1):3-9.
[2] Frazer KA, Ballinger DG, Cox DR, et al. A second generation human haplotype map of over 3.1 million SNPs[J]. Nature, 2007, 449(7164):851-861.
[3] Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3[J]. Fly, 2012, 6(2):80-92.
[4] 涂欣, 石立松, 汪樊,等. 全基因组关联分析的进展与反思[J]. 生理科学进展, 2010, 41(2):87-94.
[5] 权晟, 张学军. 全基因组关联研究的深度分析策略[J]. 遗传, 2011, 33(2):100-108.
[6] 郝兴杰, 胡林, 张淑君. 全基因组关联分析方法的研究进展[J]. 畜牧兽医学报, 2016, 47(2):213-217.
[7] 郑娟娟, 孙远洁, 李昂,等. 探讨χ^2 检验结合 FDR 筛选致病 SNPs 位点的适用条件[J]. 实用预防医学, 2012, 19(11):1604-1608.
[8] 刘匆提, 李昂, 门志红,等. 惩罚 logistic 回归方法在 SNPs 数据变量筛选研究中的应用[J]. 实用预防医学, 2016, 23(11):1395-1399.
[9] Hendricks AE, Dupuis J, Gupta M, et al. A comparison of gene region simulation methods[J]. PLoS One, 2012, 7(7):e40925.
[10] Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data[J]. Genetics, 2003, 165(4):2213-2233.
[11] Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs[J]. Bioinformatics, 2011, 27(16):2304-2305.
[12] Li J, Chen Y. Generating samples for association studies based on HapMap data[J]. BMC Bioinformatics, 2008, 24:9-44.
[13] Li C, Li M. GWAsimulator: a rapid whole-genome simulation program[J]. Bioinformatics, 2008, 24(1):140-142.
[14] Durrant C, Zondervan KT, Cardon LR, et al. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes[J]. Am J Hum Genet, 2004, 75(1):35-43.
[15] Rosenberg NA, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms[J]. Nat Rev Genet, 2002, 3(5):380-390.
[16] Peng B, Kimmel M. simuPOP: a forward-time population genetics simulation environment[J]. Bioinformatics, 2005, 21(18):3686-3687.
[17] Peng B, Amos CI, Kimmel M. Forward-time simulations of human populations with complex diseases[J]. PloS Genet, 2007, 3(3):e47.
[18] Peng B, Amos CI. Forward-time simulation of realistic samples for genome-wide association studies[J]. BMC Bioinformatics, 2010, 1(11):442.
[19] Wright FA, Huang H, Guan X, et al. Simulating association studies: a data-based resampling method for candidate regions or whole genome scans[J]. Bioinformatics, 2007, 23(19):2581-2588.
[20] 孙远洁, 郑娟娟, 李昂,等. 复杂性疾病 SNPs 数据模拟的实现与效果评价[J]. 实用预防医学, 2013, 20(1):4-8.

收稿日期:2017-03-01

the ROC curve (AUC) were used to evaluate the variable selection and predictive ability of the above-mentioned three methods. **Results** The simulation test showed that the variable selection and predictive ability of the models established by the three methods were improved when the number of differential variables was fixed and the correlation coefficient between differential variables increased. When the correlation coefficient between differential variables was constant and the number of differential variables increased, the variable selection and predictive ability of SCAD-SVM and Elastic Net showed a downward tendency, whereas those of SVM showed an upward tendency. **Conclusions** SCAD-SVM not only improves the deficiency of SVM, which can not make variable selection directly, but also simultaneously promotes the precision and prediction accuracy of the model established. On the whole, SCAD-SVM is superior in the variable selection and predictive ability; moreover, it can get higher prediction precision and more stable model estimate when manipulating the high correlation data between variables of gene expression profile data.

Key words: SCAD-SVM; Elastic Net; consistency error rate; the area under the ROC curve

随着人类基因组测序逐渐完成,大量的基因表达谱数据不断涌现,基因表达谱数据的变量个数远远大于样本例数[1]。同时这些基因之间是彼此关联的,增加了基因表达谱数据分析的复杂性。对基因表达谱数据分析的重要任务之一是根据已知的数据进行变量筛选和建立判别模型,找出有意义的差异变量,对未知的样本进行分类,这对疾病的诊断和预测具有非常重要的意义。支持向量机是效果很好的分类器,在变量个数远远大于样本数的情况下也能很好实现判别。但支持向量机不能直接进行变量筛选,通常需要借助其他算法,例如 SVM-RFE 算法,或主成分分析。而 SCAD-支持向量机则因为加入了惩罚项通过压缩变量系数直接对差异变量进行提取,从而解决支持向量机变量筛选方面不足。SCAD-支持向量机方法从基因表达谱数据的角度在疾病的判别与预测的研究中探索较少。本文将 SCAD-支持向量机作为主要研究方法,应用于不同情况下的基因表达谱数据,并与常用的支持向量机、弹性网方法进行比较,探讨 SCAD-支持向量机在基因表达谱数据中的变量筛选能力与模型预测判别能力。

1 原理与方法

1.1 支持向量机 支持向量机(SVM)是 Cones 于 1995 年提出的[2],它适用于少样本、非线性的解决及高维模式识别。支持向量机是建立在统计学习理论的 VC[3](Vapnik Chervoenkis)维理论和结构风险最小原理基础上的,它的基本思想是通过寻找一个能够满足分类要求的最大间隔超平面,同时在保证分类精度的情况下使该分类面两侧的空白区域达到最大化。支持向量机最初是针对线性可分问题进行研究的,可以实现线性可分数据的最优分类。假设二分类问题中,$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 为样本集。其中,$x_i\in R^P$,$y_i\in\{-1,1\}$,$i=1,2,\cdots,n$;p 为每个样本的维数;n 为样本个数。超平面的方程可记为:$\omega\cdot X+b=0$,其中:ω 为权向量;b 为阈值。对该方程归一化,如果使它满足约束条件 $y_i(\omega\cdot X_i+b)\geq 1(i=1,2,\cdots,n)$,则该超平面可以实现对样本的正确分类,分类间隔为 $2/\|\omega\|$,使 $\|\omega\|$ 最小化才能保证超平面两侧的间隔最大化。对于不可分情况,需要在约束条件中引入一个松弛项 ξ_i,则支持向量机最优超平面的寻找问题可以转换为下式的求解问题:

$$\min_{\omega,b,\xi}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i$$
$$s.t.\ y_i(\omega\cdot X_i+b)\geq 1-\xi_i$$
$$\xi_i\geq 0,i=1,2,\cdots,n$$

ξ_i 表示样本被错误分类的程度;C 为正则参数,用于控制对错误分类样本的惩罚程度。由于支持向量机自身的算法原理,虽然支持向量机是很好的分类器,但是在变量筛选方面却有其不足,不能够直接筛选变量,通常需要借助其他算法才能更好的筛选变量。由于可以假设变量数目远大于样本含量的高维数据通常是线性可分的[4],利用非线性的核函数增加复杂性不必要,因此,在该文章中使用线性可分支持向量机模型。

1.2 SCAD 方法 L1 惩罚方法又称为 LASSO,LASSO 的思想直接来自于 Breiman[5] 在 1993 年提出的非负套索方法(least absolute shrinkage and selectionator operator,LASSO)。LASSO 方法是通过最大化对数似然值的同时,受限制于所有系数的绝对值的和小于等于 λ 来得到系数的估计值。L1 惩罚是通过对系数的绝对值惩罚从而将一些自变量的系数压缩至 0。LASSO 惩罚的形式可定义为:$pen_\lambda(w)=\lambda\|\omega\|_1=\lambda\sum_{i=1}^{n}|\omega_i|$,其中,$\lambda\geq 0$ 为惩罚参数,而 SCAD 则是一个非凸惩罚函数,是由 Fan 和 Li 首次提出[6],是在 LASSO 的基础上提出了数学性质更加优良的 SCAD 罚,是对 L1 惩罚过度压缩系数的惩罚函数项的弥补。SCAD 的惩罚函数的系数 ω_i 被定义为

$$P_{SCAD(\lambda)}(\omega_i)=\begin{cases}\lambda|\omega_i| & if|\omega_i|\leqslant\lambda,\\ -\dfrac{|\omega_i|^2-2a\lambda|\omega_i|+\lambda^2}{2(a-1)} & if<|\omega_i|\leqslant a\lambda,\\ \dfrac{(a+1)\lambda^2}{2} & if|\omega_i|>a\lambda\end{cases}$$

$\omega_i, i=1,2,\cdots,n$ 是定义超平面的系数；$a>0$ 和 $\lambda>0$ 惩罚参数。Fan 和 Li 认为 SCAD 预测准确率在选择不同的调优参数 a 时并不敏感，并且建议值=3.7 可以用于分析[7]。SCAD 惩罚对于拥有小系数的变量的压缩，SCAD 与 L1 惩罚有相同的功能。然而对于大型系数，SCAD 提供了一个固定的惩罚，这是与 L1 惩罚不同的。这减少了估计的偏倚。此外，SCAD 惩罚相比与 L1 惩罚有更好的理论基础。

1.3 SCAD-支持向量机 Hastie 认为支持向量机的最优化问题其实相当于惩罚问题[8]。它是在损失函数后加上了二次惩罚以收缩系数。Hastie[9] 相继提出 SCAD-支持向量机并进行讨论。后来，经常将 SCAD 与支持向量机的结合用于高维数据的处理。

SCAD-支持向量机的惩罚形式为：

$$\min_{\omega,b,\xi}\frac{1}{n}\sum_{i=1}^{n}[1-y_i f(x_i)]+\sum_{i=1}^{p}P_{SCAD(\lambda)}(\omega_i)$$

其中，$\lambda\geqslant 0$ 是惩罚参数。

SCAD-支持向量机可以产生很稀疏的解，从而自动直接的实现变量选择，进一步解决了支持向量机不能直接进行变量筛选的不足。并且在实现自动筛选变量的基础上提高模型的解释性。处理变量间有高度相关性的高维数据时可以获得了更高的预测精度和更稳定的模型估计。

1.4 弹性网 岭回归是 1970 年由 Hoerl 和 Kennard[10] 提出的，它是一种有偏估计，以损失部分信息、降低精度为代价获得回归系数，是对最小二乘估计的改进。岭回归也 L2 惩罚，也叫权重衰减（weight decay）。L2 模型的惩罚形式为

$$pen_\lambda(w)=\lambda\|w\|_2^2=\lambda\sum_{i=1}^{n}\omega_i^2$$

其中，$\lambda\geqslant 0$ 为惩罚参数。参数 λ 同样用来控制压缩量，但只会使系数趋于 0，而不会等于 0。

弹性网（Elastic Net）思想则是 Zou 和他的导师 Trevor Hastie 在 2005 年提出的[11]，其结合了 LASSO 和岭回归的思想，即同时引入两个惩罚项。弹性网将具有相关性的变量放入模型，克服了 LASSO 变量选择的严苛性，结合岭回归提高模型的准确率。弹性网惩罚的形式为 $pen_\lambda(w)=\lambda_1\|w\|_1+\lambda_2\|w\|_2^2$

λ_1 和 λ_2 是分别代表 L1 惩罚参数和 L2 惩罚参数且均为正数。弹性网结合了 L1 和 L2 惩罚项既满足了产生稀疏模型的要求，又允许变量间具有共线性，从而解除了筛选变量个数的限制，同时可以提高模型的预测准确度，可以有效的处理高维数据。

1.5 调整参数 λ 的选择 惩罚参数 λ 的估计方法有交叉验证法、广义交叉验证法、无偏估计的风险分析[22] 三种方法。在实际应用中，交叉验证是最常用的方法[8]。通过估计方程对数似然函数计算惩罚参数 λ，并最终决定进入模型的自变量。本研究采用交叉验证法计算，取对数似然函数的最大值所得的惩罚参数 λ_1 和 λ_2 作为模型的惩罚系数。而支持向量机采用 SVM-RFE 算法采取序列后退进行变量的筛选。

2 结 果

2.1 模拟数据的参数设置 本研究进行模拟实验旨在讨论不同情况下即随着差异基因数目增多（即噪声变量数目降低）与差异变量间相关性增加，SCAD-支持向量机、支持向量机、弹性网三种方法的模型预测判别能力优劣。

模拟实验的设置条件参照了现有的基因表达谱数据，并考量了算法模型预测的特点进行设置。本研究设置了 1 000 个变量和一个二分类表型变量病例对照数据，差异变量数目分别为 10、50、100、150、200（即差异变量占总变量的 1%、5%、10%、15%、20%），差异变量之间的相关系数分别为 0.3、0.5、0.8，样本量为 100（病例组与对照组比例为 1:1），每组参数组合重复次数 100 次，见表 1。

表 1 模拟实验参数设置

参数名称及说明	参数值	参数名称及说明	参数值
模拟实验次数	100	差异变量间相关系数	0.3、0.5、0.8
样本量	100	交叉验证次数	10
总变量数	1 000	L1 正则化系数	[0,5]
差异变量数量	10、50、100、150、200	L2 正则化系数	[0,100]
		迭代次数	1 000

2.2 模拟实验结果 三种方法变量筛选能力的比较采用假发现率 FDR（false discovery rate）：拒绝原假设的个数占所有个数的比例，FDR 越低表示变量筛选能力越好；三种方法模型预测能力的比较，即对三种方法的一致性错误率，其定义为：一致性错误率=预测结果错误的样本数量/样本总数（即一致性错误率越小，方法的模型预测判别能力越好）和 AUC 值（ROC 曲线下面积，AUC 值越大，方法的模型预测判别能力越好）进行比较。三种方法分别使用 R 2.15.3 统计软件中的“penalizedSVM”、“e1071”、“glmnet”函数包实现。其中 SCAD-支持向量机和弹性网均采用交叉验证法计

算惩罚参数,而支持向量机采用 SVM-RFE 算法采取序列后退进行变量的筛选。

表 2 三种方法假发现率 FDR 值的比较结果

相关系数	方法	nsign * = 10	nsign = 50	nsign = 100	nsign = 150	nsign = 200
r * * = 0.3	SCAD-SVM	0.39±0.03	0.41±0.06	0.44±0.04	0.46±0.05	0.50±0.05
	SVM	0.56±0.04	0.72±0.04	0.64±0.02	0.57±0.04	0.54±0.04
	Elastic-Net	0.40±0.03	0.47±0.06	0.50±0.04	0.52±0.05	0.55±0.05
r = 0.5	SCAD-SVM	0.38±0.07	0.44±0.04	0.47±0.04	0.50±0.04	0.53±0.04
	SVM	0.54±0.04	0.53±0.04	0.41±0.04	0.39±0.07	0.36±0.07
	Elastic-Net	0.33±0.07	0.40±0.04	0.43±0.04	0.48±0.04	0.51±0.04
r = 0.8	SCAD-SVM	0.31±0.04	0.32±0.04	0.37±0.02	0.39±0.04	0.41±0.04
	SVM	0.65±0.05	0.59±0.04	0.47±0.06	0.44±0.03	0.42±0.03
	Elastic-Net	0.28±0.04	0.39±0.02	0.47±0.04	0.45±0.04	0.45±0.04

注:* 模拟实验中设置差异变量的个数,* * 差异变量之间的相关系数。

三种方法筛选变量的能力从表 2 可以看出,在差异变量数目不变的情况下,随着差异变量间相关系数的增加,三种方法筛选出的变量假发现率 FDR 一致性错误率均呈现下降趋势;当差异变量间相关系数不变时,随着差异变量数目的增加,假发现率 FDR 表现为 SCAD-支持向量机和弹性网方法呈上升趋势,而支持向量机呈现下降趋势;当差异变量数目和差异变量间相关系数相同时,SCAD-支持向量机的假发现率 FDR 均少于弹性网和支持向量机。

三种方法模型预测能力从表 2 可以看出,在差异变量数目不变的情况下,随着差异变量间相关系数的增加,三种方法建立的模型的一致性错误率均呈现下降趋势;AUC 值均呈现上升趋势。当差异变量间相关系数不变时,随着差异变量数目的增加,一致性错误率表现为 SCAD-支持向量机和弹性网方法呈上升趋势,而支持向量机呈现下降趋势;AUC 值表现为 SCAD-支持向量机和弹性网方法均呈下降趋势,而支持向量机呈现上升趋势。当差异变量数目和差异变量间相关系数相同时,SCAD-支持向量机的一致性错误率均少于弹性网和支持向量机;当差异变量间的相关系数固定,差异变量数目较少(0<nsign≤100)时,三种方法的一致性错误率表现为支持向量机>弹性网>SCAD-支持向量机;AUC 值表现为 SCAD-支持向量机>弹性网>支持向量机。差异变量数较多(100<nsign≤200)时,三种方法的一致性错误率表现为弹性网>支持向量机>SCAD-支持向量机;AUC 值表现为 SCAD-支持向量机>支持向量机>弹性网,见图 1、图 2。

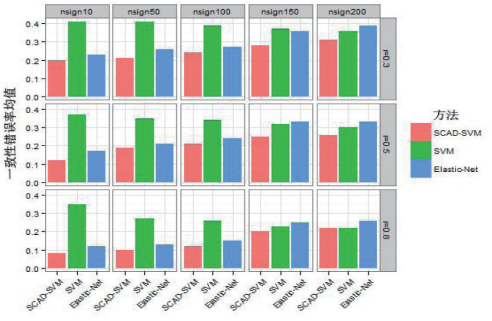

图 1 三种方法一致性错误率的比较结果

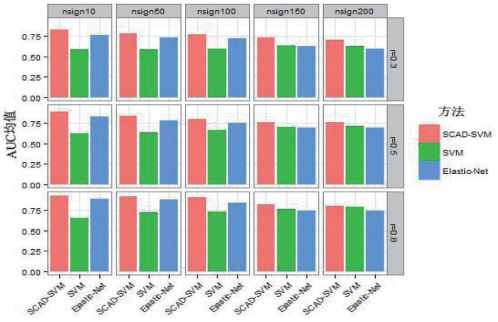

图 2 三种方法 AUC 值的比较结果

表 3 三种方法一致性错误率、AUC 值的比较结果

相关系数	方法	指标	nsign * = 10	nsign = 50	nsign = 100	nsign = 150	nsign = 200
r * * = 0.3	SCAD-SVM	错误率	0.20±0.06	0.21±0.04	0.24±0.08	0.28±0.04	0.31±0.05
		AUC 值	0.83±0.06	0.78±0.05	0.77±0.10	0.73±0.07	0.70±0.06
	SVM	错误率	0.41±0.03	0.41±0.04	0.39±0.07	0.37±0.06	0.36±0.04
		AUC 值	0.59±0.03	0.59±0.05	0.60±0.06	0.64±0.03	0.66±0.06
	Elastic-Net	错误率	0.23±0.05	0.26±0.04	0.27±0.06	0.36±0.05	0.39±0.03
		AUC 值	0.76±0.05	0.73±0.04	0.72±0.05	0.63±0.03	0.60±0.04
r = 0.5	SCAD-SVM	错误率	0.12±0.06	0.19±0.05	0.21±0.06	0.25±0.07	0.26±0.04
		AUC 值	0.89±0.04	0.84±0.05	0.80±0.06	0.76±0.06	0.74±0.07
	SVM	错误率	0.37±0.05	0.35±0.04	0.34±0.05	0.32±0.04	0.30±0.05
		AUC 值	0.63±0.04	0.64±0.04	0.66±0.05	0.70±0.05	0.71±0.05
	Elastic-Net	错误率	0.17±0.04	0.21±0.04	0.24±0.04	0.33±0.04	0.33±0.07
		AUC 值	0.83±0.07	0.78±0.04	0.75±0.04	0.69±0.04	0.67±0.07
r = 0.8	SCAD-SVM	错误率	0.08±0.03	0.10±0.04	0.12±0.07	0.20±0.05	0.22±0.07

续表 3

相关系数	方法	指标	nsign * = 10	nsign = 50	nsign = 100	nsign = 150	nsign = 200
		AUC 值	0.93±0.06	0.92±0.07	0.91±0.08	0.82±0.03	0.78±0.05
	SVM	错误率	0.35±0.05	0.27±0.06	0.26±0.04	0.23±0.04	0.22±0.04
		AUC 值	0.65±0.05	0.72±0.05	0.73±0.05	0.76±0.02	0.77±0.04
	Elastic-Net	错误率	0.12±0.04	0.13±0.04	0.15±0.02	0.25±0.04	0.26±0.04
		AUC 值	0.89±0.04	0.88±0.04	0.84±0.03	0.74±0.06	0.73±0.04

注：* 模拟实验中设置差异变量的个数；* * 差异变量之间的相关系数。

2.3 *实例分析* 本研究从 TCGA 数据库下载 40 例结肠癌患者的基因表达谱数据，以及 22 例对照数据。全基因组表达谱数据一共测得 2 000 个基因的表达值。将 SCAD-支持向量机、支持向量机、弹性网三种方法应用到该数据中，通过 5 折交叉验证法得出一致性错误率和 AUC 值，三种方法的一致性错误率表现为支持向量机>弹性网>SCAD-支持向量机；三种方法的 AUC 值表现为 SCAD-支持向量机>弹性网>支持向量机，见表 4。其中取 SCAD-SVM 方法 AUC 值最大的一次筛选出差异变量个数为 35 个。

表 4 实际数据分析结果比较

方法	一致性错误率	AUC 值
SCAD-SVM	0.12±0.05	0.87±0.04
SVM	0.19±0.02	0.76±0.05
Elastic-Net	0.16±0.05	0.81±0.06

3 讨 论

高维数据分析问题的复杂性并不是传统统计学可以简单解决的。基因表达谱数据的挖掘问题便是其中一类。其典型的数据特征即是自变量个数远远大于样本量，这不仅要求自动的变量选择和特征压缩，往往还要求精确预测判别，对病例诊断起指导作用为病理的研究节省时间和精力。支持向量机是由 Vapnik 提出的机器学习方法。在变量个数远远大于样本数的情况下也能很好实现判别。但支持向量机进行变量筛选时通常需要借助其他算法，例如 SVM-RFE 算法，或主成分分析法。而 SCAD-支持向量机则因为加入了惩罚项通过压缩变量系数直接对差异变量进行提取，从而解决支持向量机变量筛选方面不足。弹性网结合了 L1 和 L2 惩罚的优点也可以较好的处理具有较高共线性变量的高维数据。本文主要探讨 SCAD-支持向量机在差异基因数量不同和相关系数变化的情况下，利用 FDR、一致性错误率和 ROC 曲线下面积（AUC 值）对比 SCAD-支持向量机、支持向量机、弹性网三种方法对基因表达谱数据的变量筛选和预测能力的情况。模拟实验结果显示在差异变量数不变的情况下，随着差异变量间相关系数的增加，三种方法建立的模型的变量筛选和预测判别能力均提高；当差异变量间相关系数不变时，随着致病位点数的增加，SCAD-支持向量机和弹性网方法的变量筛选和预测判别能力均呈下降趋势，而支持向量机呈现上升趋势。综合模拟实验和实际数据的分析来看，SCAD-支持向量机的变量筛选和预测判别能力更优，处理变量间有高度相关性的基因表达谱数据时可以获得了更高的预测精度和更稳定的模型估计。

综合来看本文提出的 SCAD-支持向量机不仅改善了支持向量机不能直接进行变量筛选的不足同时提高了模型的精度以及判别的准确性。将该方法应用于不同样本的基因表达谱数据进行疾病诊断结果的判别预测，可以得到更高的分类预测判别精度，提高疾病诊断的识别正确率及对原始样本数据集进行有效降维，有利于辅助医疗人员快速、准确地对患者进行初步诊断，指导医生根据疾病诊断的关键特征来预测不同患者所处的危险状况，从而给予合理的检查及治疗，实现合理用药，节约医疗资源。

参考文献

[1] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Mach Learn, 2002, 44(3): 438-443.

[2] Vapnik VN. The nature of statistical learning theory[J]. Springer, New York, 1995, 23(5): 22-26.

[3] 张学工.关于统计学习理论与支持向量机[J].自动化学报, 2000, 2(1): 32-42.

[4] Markowetz F, Spang R. Molecular diagnosis: classification, model selection and performance evaluation[J]. Method Inform Med, 2005, 44(3): 438-443.

[5] Tibshirani R. Regression shrinkage and selection via the LASSO[J]. J R Stat Soc B, 1996, 58: 267-288.

[6] 刘匆提, 李昂, 门志红, 等. 惩罚 logistic 回归方法在 SNPs 数据变量筛选研究中的应用[J]. 实用预防医学, 2016, 23(11): 1395-1399.

[7] Fan JQ, Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. J Am Stat Assoc, 2001, 96: 1348-1360.

[8] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining inference and prediction[J]. Springer, New York, 2001, 12(3): 368-376.

[9] Zhang HH, Ahn J, Lin X, et al. Gene selection using support vector machines with non-convex penalty[J]. Bioinformatics, 2006, 22(1): 88-95.

[10] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems [J]. Technometrics, 1970, 12(1): 55-67.

[11] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. J R Stat Soc B, 2005, 67(2): 301-320.

收稿日期：2017-03-22