基于随机复杂度约束的高维特征自动选择算法

刘 峤,王 娟,陈 伟,秦志光
(电子科技大学计算机科学与工程学院,四川成都 610054)

摘 要: 高维特征选择问题是机器学习研究领域的公开问题,当前流行的 1-范数约束正则化解决方案存在的主要问题是缺乏特征组选能力和特征选择能力受样本容量限制.本文从随机复杂度理论的模型冗余度最优下界推导得出了一种易于求解的基于零 - 范数约束的特征选择算法模型.该算法不仅可证优化,而且具备自动特征选择能力,克服了 1-范数约束方法的主要缺点,算法不依赖于对数据真实生成模型的参数假设,具有广泛的适用性.仿真实验表明该算法在常规数据建模任务中的性能表现与 1 - 范数约束方法相当,在真实基因数据集上的测试结果进一步验证了该算法在高维特征空间的性能优于近期发表的一些主要算法.

关键词: 机器学习; 生物信息学; 特征选择; 正则化方法; 高维
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2011) 02-0370-05

An Automatic Feature Selection Algorithm for High Dimensional Data Based on the Stochastic Complexity Regularization

LIU Qiao, WANG Juan, CHEN Wei, QIN Zhi-guang
(*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan* 610054, *China*)

Abstract: Feature selection for high-dimensional sparse feature space is an open issue for machine learning research, prevalent 1-norm regularization approaches share some theoretical drawbacks, such as lack the ability to select out grouped features, and can not select more features than the sample size. This paper considers the sparse modeling problem from the stochastic complexity theory perspective, and derive an easy computable model from its Minimax bound approximation. The proposed approach is proved to be optimized, and can perform automatic feature selection similar to its 1-norm penalized alternatives, but overcome their drawbacks. Furthermore, it does not rely on any parametric assumptions about the true data-generating mechanism, which makes it broadly applicable. Various simulations performed with both synthetic and real biological data show that the proposed approach performs similarly to the popular 1-norm penalized counterparts in ordinary experimental setups, and outperforms the other methods in robustness and predictive accuracy for extremely sparse problems.

Key words: machine learning; bioinformatics; feature selection; regularization; high dimensional

1 引言

高维特征选择问题也称为稀疏建模问题,是当前机器学习研究领域的三大公开理论问题之一,主要研究目标是为了解决现有的特征建模方法在高维特征空间失效的问题[1].主要的理论困难包括:大样本假设不成立导致多数经典统计学方法失效;数据的特征稀疏性致使从经验数据出发反向建模成为病态问题(ill-posed inverse problem);因特征维数过高而导致的维数灾难问题;以及最优子集搜索问题为 NP-hard 问题等[2,3].

该领域近年来一个最重要的研究方向是以套索算法(LASSO)为代表的基于 1-范数约束的正则化特征选择方法(简称 L1 方法).由于 LASSO 的可证优化和它所具备的自动特征选择能力,使得 L1 方法被视为高维特征选择问题最有希望的解决方案[4].然而研究表明 L1 方法自身存在着严重的理论局限性,其中最主要的问题是 L1 模型优化方程的凸性不严格,当特征高度相关时不能保证优化解的唯一性.L1 方法的另一个缺点是选出的特征集合的势无法超越样本容量限制,这显然与特征建模的基本原则相悖[5].

与 L1 方法的研究进程遭遇理论瓶颈形成对比的是,统计学家从理论上证明了基于零-范数约束的正则化方法(简称 L0 方法)在稀疏建模问题中的性能优于 L1 方法[6].本文的研究动机即据此构造出一种新的 L0

收稿日期:2009-11-17;修回日期:2010-03-01
基金项目:国家 863 高技术研究发展计划(No.2006AA01Z411);四川省科技支撑计划(No.08ZC1543)

方法模型,使之既具备 L1 方法的可证优化特性和自动特征选择能力,同时又克服了它的理论局限.

通过借鉴随机复杂度理论有关模型复杂度编码的基本原理,本文推导出了一个可证优化的随机复杂度测度,称为随机复杂度最优判据(Stochastic Complexity Optimization Criterion,SCOC),并据此建立了基于随机复杂度约束的特征选择方法模型.仿真实验和真实基因数据集上的实验数据表明,SCOC 不仅具备 L1 方法的自动特征选择能力,而且建模能力不受特征维度和样本容量的限制,相对于已有方法在高维特征空间具备明显的性能优势.

2　相关工作

在特征建模研究中有两个隐含的指导原则:(1)是以模型拟合优度(或预测精度)来评价模型的好坏;(2)是节俭原则(Principle of Parsimony,也称为奥坎姆剃刀原则),即在从同一数据出发得到的多个性能相当的模型中,简单模型的泛化能力也许是最好的[7].

然而这两个原则是相互矛盾的,在实际应用中往往越复杂的模型对数据的拟合越好[8].因此通常将特征选择问题定义为一个模型拟合优度与模型复杂度之间的零和博弈问题,目标是寻求使得最坏情况下的预测风险极小化的最优解,即 Minimax 意义上的最优(极大风险极小化).

LASSO 算法就是利用了 1-范数约束方程的凸性,从而将 NP-hard 的最优子集搜索问题转化为能够从数学上精确求解的凸优化问题,并借助 1-范数约束条件对参数的缩水(Shrinkage)效应实现了对自动特征选择.然而 LASSO 的理论缺陷限制了它的实用性,为此学术界做了大量努力,重要成果包括快速求解 LASSO 的最小角度回归方法(Least Angle Regression,LARS)[9];通过改进 LASSO 模型的凸性来提高特征组选能力的弹性网络方法(Elastic Net,ENet)[5];以及为改善 LASSO 的参数估计性能而提出的 Adaptive LASSO 方法和 Dantzig Selector 方法等[10,11].尽管这些研究成果部分弥补了 LASSO 的局限性,然而实验表明改进只是局部的[12].

近期的研究表明 L1 方法在稀疏特征空间中的决策风险相对于 L0 方法渐进地趋于无穷大[6],这一结论为突破 L1 方法框架的束缚提供了新的理论依据.然而 L0 方法并非新生事物,经典统计建模方法如 AIC 和 BIC 等就属于该方法范畴,且已被证明不适用于稀疏建模任务[7].随机复杂度理论揭示出 AIC 等传统 L0 方法在高维特征空间失效的原因在于其模型复杂度约束低于模型冗余度最优下界[13].本文的工作即以此为基础,从模型随机复杂度的最优下界出发推导出 SCOC 模型.

为验证 SCOC 模型在稀疏建模问题中的实际性能,本文选择了两组新近发表的实验数据作为参照,分别来自 Fan 等人提出的 FAIR(Features Annealed Independence Rules)方法[14],和 Greenshtein 等人提出的 EB(Empirical Bayes)方法[15].理由是两者均采用了相同的公开基因数据集,便于客观比较实验结果,且二者的实验结果代表了目前在该领域的最好水平.选择基因芯片数据进行算法性能测试的另一个原因是该数据在稀疏建模研究中具有典型的代表性(样本容量远低于特征维度),与一般的低维 UCI 基准数据集相比更能够反映出特征选择算法在性能上的优缺点.

3　算法细节

3.1　SCOC 的数学模型

首先定义数据生成模型的符号表示如下:

$$\boldsymbol{Y}_{n\times 1}=\boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1}+\varepsilon_{n\times 1} \tag{1}$$

其中响应向量 $\boldsymbol{Y}$ 包含 n 维样本观测值,$\boldsymbol{X}$ 表示 p 维特征空间的特征观测值,$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)$表示 p 维特征参数向量,$\varepsilon\sim N(0,\sigma^2)$为高斯噪音.

以 x^n 表示观测数据,$f_\theta(x^n)$表示从样本得到的似然函数,由此得到定义在特征空间 Θ 上的模型参数族 $M(x^n)=\{f(x^n\mid\theta):\theta\in\Theta\subset R^p\}$.依据香农的源信道编码理论可知采用参数模型 $M(x^n)$对数据 x^n 进行无损编码所能达到的平均最短编码长度为 $-\log f_\theta(x^n)$,即熵的下确界.

接下来考虑对参数模型自身的编码,由于对模型复杂度的直接编码已经被证明是不可计算的,因此在随机复杂度理论中,采用 *K-L* 距离(Kullback-Leibler Divergence)作为衡量观测模型与真实模型复杂度之间相对偏差的测度[13].设数据的真实生成模型为 $q(x^n)$,由 *K-L* 距离的定义:

$$K\text{-}L(q,f_\theta)=E_\theta\log\frac{q(x^n)}{f_\theta(x^n)}=E_\theta\{-\log f_\theta(x^n)-[-\log q(x^n)]\} \tag{2}$$

K-L 距离也被称为模型冗余度,表示用估计分布 $f_\theta(x^n)$替代真实分布 $q(x^n)$时所需的额外编码长度.由此得到参数模型的随机复杂度定义[13]:

$$SC(x^n)=-\log f_\theta(x^n)+K\text{-}L(q,f_\theta) \tag{3}$$

Rissanen 证明了模型冗余度具有如下 Minimax 意义上的最优下界[16]:

$$\min_{f_\theta}\sup_{\theta\in\Theta}K\text{-}L(q,f_\theta)=\frac{k}{2}\log\frac{n}{2\pi e}+\log\int_K\sqrt{\det\boldsymbol{I}(\theta)}\,\mathrm{d}\theta+O(1) \tag{4}$$

其中 K 为参数空间 Θ 的紧子集,k 表示进入模型的参数个数,$\boldsymbol{I}(\theta)$为参数向量 θ 的 Fisher 信息矩阵.根据式(1)和式(4),考察特征参数 θ_i 对应的 Fisher 信息量有:

$$\sqrt{\boldsymbol{I}(\theta_i)} = \sqrt{E[(\frac{\mathrm{d}}{\mathrm{d}\theta_i}\log f(x_i|\theta_i))^2]}$$
$$= \sqrt{\int_{-\infty}^{+\infty}(\frac{x_i-\theta_i}{\sigma^2})^2 \frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_i-\theta_i)^2}{2\sigma^2}\mathrm{d}\theta_i}$$
$$= \sqrt{\frac{1}{\sigma^2}} \quad (5)$$

上式表明 θ_i 对应的 Fisher 信息量在实数域上服从均匀分布(Jeffrey's Prior 的基本性质),这与 Laplace 的概率无差别原则是一致的.由于随机复杂度理论考虑的是对数据的理想编码长度,而不关心参数模型的具体形式,因此可根据概率无差别原则进一步假设 $\boldsymbol{I}(\theta)$ 正交,于是有:

$$\sqrt{\det\boldsymbol{I}(\theta)} = \det\sqrt{\boldsymbol{I}(\theta)} = (\frac{1}{\sigma^2})^{\frac{k}{2}} \quad (6)$$

将 σ^2 的极大似然估计 $\hat{\sigma}^2 = \| \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \|^2/n$ 带入上述关系式中,经整理得到:

$$\inf_{f_\theta}\sup_{\theta\in\Theta} SC(x^n) = -\log f_\theta(x^n) + k\log n - \frac{k}{2}\log\| \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \|^2 + C \quad (7)$$

其中 C 为与参数模型无关的常数.以下本文约定以 $M^-(x^n)$ 表示当前的备选模型,$M^+(x^n)$ 表示新增一维特征后得到的新备选模型.以 $f_\theta(x^n, M^-)$ 表示采用 $M^-(x^n)$ 得到的极大似然,$f_\theta(x^n, M^+)$ 表示新增一维特征后的极大似然.当在模型 $M^+(x^n)$ 和 $M^-(x^n)$ 之间进行比较时,由式(7)出发直接得到二者的随机复杂度之差的计算表达式如下,本文将其定义为 SCOC 判据:

$$\mathrm{SCOC} = SC(M^+) - SC(M^-) = \log\frac{f_\theta(x^n, M^-)}{f_\theta(x^n, M^+)}$$
$$+ \log n - \frac{1}{2}\log\| \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} \|^2 \quad (8)$$

3.2 采用 SCOC 判据进行特征选择

根据 SCOC 判据的性质将特征选择问题定义为一个零-范数约束的优化问题,目标是在特征空间中寻找满足如下优化条件的 $\hat{\beta}$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\{-\log f_\theta(x^n)\}, \quad (9)$$

其中:$\| \boldsymbol{\beta} \|_0 = \sum_{i=1}^{p} ind(\beta_i) = c(c > 0)$

由多元正态分布公式可知对于线性模型(1)有如下关系成立:

$$-2\log f(x^n) = n\log\| \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \|^2 + n\log(\frac{2\pi e}{n}) \quad (10)$$

将式(10)代入式(9),根据 Wolfe 对偶理论可知优化问题(9)的对偶问题可以表达为:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\{\frac{n}{2}\log\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 + \lambda \| \beta \|_0\} \quad (11)$$

其中参数 λ 表示参数向量 $\hat{\boldsymbol{\beta}}$ 中的单位元素向模型中引入的模型复杂度增量.注意到式(8)等号右侧除第一项外的其余两项之和即为新增特征所带来的模型复杂度增量,由于式(7)给出的随机复杂度达到 Minimax 意义上的最优,因此 λ 可以显式地表示为:

$$\lambda = \log n - \frac{1}{2}\log\| \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \|^2 \quad (12)$$

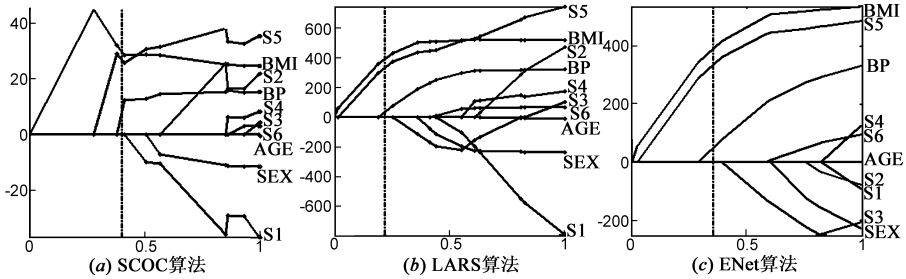
注意 λ 的取值在特征选择过程中随参数模型的复杂度增加呈非线性递增,这一性质确保了节俭原则的实现.通过上述推导将非凸优化问题(9)简化为一个凸优化问题,确保了解的存在性和唯一性.该问题在实际应用中可通过分步回归法自动求得经验解,从而实现自动特征选择.方法是:设初始模型为空集,每轮迭代向模型中增加一维使似然估计值最大的特征,然后通过计算相邻模型的 SCOC 值决定模型的取舍,如果 SCOC 值小于 0,则拒绝 $M^-(x^n)$ 模型,并接受 $M^+(x^n)$ 模型,继续执行下一轮迭代;反之若 SCOC 值大于 0,则拒绝 $M^+(x^n)$ 模型,特征选择过程结束,输出 $M^-(x^n)$.

4 实验结果分析

4.1 仿真实验结果分析

本文采用糖尿病公开数据集进行仿真测试,并就实验结果与 LARS 和 ENet 等两种经典 L1 方法进行比较*.该数据集包含 442 个糖尿病实例,每组实例包含 10 个基本特征**.

图 1 记录了 SCOC 与 LARS 和 ENet 等算法在糖尿病数据集上的实验结果,图 1(a)、(b)和(c)分别显示了三种算法在范数约束条件逐步放宽的情况下,参数特征先后进入模型的顺序.图中的横坐标为归一化后的模型复杂度约束条件,纵坐标为相应的特征回归系数,图中与纵坐标平行的虚线指出了根据残差平方和的 10 轮交叉验证结果选出的最优模型的位置.



(a) SCOC算法 (b) LARS算法 (c) ENet算法

图1 SCOC、LARS和ENet算法在糖尿病基准数据集上的特征选择顺序

图 1(a)清晰地展示了 SCOC

* LARS 和 ENet 算法程序来源:http://www2.imm.dtu.dk/pubdb/views/publication_details.php? id=3897

** 糖尿病数据下载地址:http://www-stat.stanford.edu/~hastie/Papers/LARS/

的自动特征选择能力,从图中可以看出特征按照 BMI、S5、BP、S1、SEX、S2、S4、S6、S3、AGE 的顺序逐一进入模型,与 LARS 和 ENet 的实验结果相比较,可以看出三者选出的最优模型是一样的,即特征集合{BMI,S5,BP}.需要说明是为了绘图演示的需要,试验中对 SCOC 算法的模型约束条件进行了人为调整以便使 10 个特征全部得以进入模型.由于 SCOC 判据是可证优化的,因此在实际应用中不需要像 LARS 和 ENet 一样借助交叉验证等手段来经验地选择最优模型,这也是 SCOC 算法的一个突出的优点.为验证这一点,采用分布回归法进行测试,结果表明 SCOC 恰好自动选出了经上述交叉验证得出的最佳模型,即特征集合{BMI,S5,BP},由此进一步验证了 SCOC 算法具有比 L1 方法更为理想的自动特征选择能力.

4.2　真实数据实验结果分析

为客观验证 SCOC 算法的实验性能,本文选择三组真实基因数据进行测试,并就有关结果与 FAIR 和 EB 算法进行比较[14,15].三组公开数据集分别是肺癌数据集(Lung cancer)、前列腺癌数据集(Prostate cancer) 和白血病数据集(Leukemia),其共同特点是特征维度远高于样本容量,特征相关程度高,是典型的稀疏建模问题.关于这三组数据集的基本信息总结于表 1(数据来源参见文献[14]).

表 1　三组基因芯片公开数据集的组成情况一览表

数据集	样本数	特征数	训练集			测试集		
			正例	反例	合计	正例	反例	合计
Leukemia	72	7129	27	11	38	20	14	34
Lung cancer	181	12533	16	16	32	15	134	149
Prostate	136	12600	52	50	102	25	9	34

表 2　SCOC、FAIR 与 EB 算法的基因选择实验结果一览表

	Leukemia			Lung cancer			Prostate cancer		
	TRE	*TEE*	# *G*	*TRE*	*TEE*	# *G*	*TRE*	*TEE*	# *G*
SCOC	0/38	1/34	8	0/32	13/149	20	1/102	4/34	14
EB	0/38	3/34	all	0/32	1/149	all	38/102	4/34	all
FAIR	1/38	1/34	11	0/32	7/149	31	10/102	9/34	2

表 2 记录了三种算法在全部数据集上的实验结果,内容包括训练误差(*TRE*),测试误差(*TEE*),以及各算法选出的基因表达式特征数目(# *G*).从中可以看出在白血病数据集上 SCOC 的训练和测试误差均优于 FAIR 和 EB.在前列腺癌数据集上,SCOC 的训练和测试误差均优于 FAIR,其测试误差与 EB 相同,但训练误差远低于 EB 算法.然而在肺癌数据集上,SCOC 的测试误差明显高于其他两种算法,经分析造成这种情况的原因可能有两种,一是由于肺癌数据集的训练集和测试集的样本分布不均衡(正反样本比例分别为 16:16 和 134:15)而造成训练时的参数估计有偏差,第二种可能是 SCOC 算法选出的特征集合不能够准确反映数据的真实情况.

为检验上述两种假设中哪一个是影响算法性能的主要因素,本文参照文献[14]的试验方法进行了交叉验证测试.方法是将肺癌数据集的训练集与测试集合并后,从中抽取其中的 $S\%$ 作为新的测试集,以剩余的$(1-S\%)$作为训练集进行特征选择实验.分别取 $S=\{0.3,0.4,0.5,0.6,0.7,0.8\}$,对每个 S 的取值均重复随机采样测试 100 次.另外分别从正反样本中各随机抽取 16 个实例组成新的训练集,以剩余的 149 个样本作为测试集同样进行 100 轮测试,所得结果作为基准数据(baseline).

图 2 以箱线图的形式记录了 SCOC 算法在上述实验条件下的测试误差情况,其中横坐标表示六种不同的采样比例以及基准测试(虚线右侧),纵坐标给出了采用 SCOC 算法建立的特征模型在测试集上的分类误差占测试样本的百分比.从图中可以看出 SCOC 算法在肺癌数据集的各种划分情况下的测试分类误差水平变化不大,且优于 FAIR 算法给出的结果(7/149≈4.7%),说明采用 SCOC 算法建模能够准确筛选出包含数据真实信息的特征集合.通过进一步查看 SCOC 在基准划分方式下的交叉验证结果,可以看到其分类错误率指标接近于 FAIR 的测试结果,由此证明了前述分析中的第一种原因成立,即训练样本的不均衡是导致分类误差水平较高的根本原因.

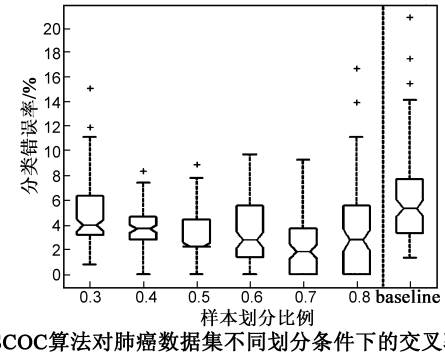


图2　SCOC算法对肺癌数据集不同划分条件下的交叉验证结果

综上所述,SCOC 算法在真实基因数据上表现出的实验性能从整体上优于 FAIR 和 EB 算法,采用该算法对数据进行建模能够获得良好的分类准确性,算法稳定性和泛化能力较好.

5　结论

本文从随机复杂度理论的基本结论出发推导得出了一个可证优化的特征选择判据(SCOC),并据此提出了一个基于零 - 范数约束的特征选择算法模型.实验结果表明 SCOC 算法模型具备接近于主流 L1 方法的自动特征选择能力,在处理特征高度稀疏的基因选择任务时,算法性能总体优于现有的代表性稀疏建模方法.

从 SCOC 模型的推导过程可知模型优化方程的凸性严格,确保了问题解的唯一性,且算法不受样本容量影响,从而克服了 L1 方法的理论局限性.本文理论研究价值在于证明了 L0 方法同样适用于稀疏建模任务,为稀疏建模研究突破 L1 方法框架的束缚提供了实验证据.由于稀疏建模方法在生物信息学、图像分析和文本数据挖掘等众多相关领域有着现实而广泛的应用需求,因此本文的成果不仅具有理论意义,也具有一定的实践推广价值.

参考文献:

[1] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. The Journal of Machine Learning Research, 2003, 3(3): 1157 – 1182.

[2] 王雪松,张依阳,程玉虎.基于高斯过程分类器的连续空间强化学习[J].电子学报,2009,37(6):1153 – 1158.
Wang Xue-song, Zhang Yi-yang, Cheng Yu-hu. Reinforcement learning for continuous spaces based on gaussian process classifier[J]. Acta Electronical Sinica, 2009, 37(6): 1153 – 1158. (in Chinese)

[3] 蒋盛益,郑琪,张倩生.基于聚类的特征选择方法[J].电子学报,2008,36(12A):157 – 160.
Jiang Sheng-yi, Zheng Qi, Zhang Qian-sheng. Clustering-based feature selection[J]. Acta Electronical Sinica, 2008, 36(12A): 157 – 160. (in Chinese)

[4] Hesterberg T, Choi N H, Meier L, Fraley C. Least angle and L1 penalized regression: A review[J]. Statistics Surveys, 2008, 2(1): 61 – 93.

[5] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society Series B, 2005, 67(2): 301 – 320.

[6] Lin D, Pitler E, Foster D P, Ungar L H. In Defense of l0[R]. In: ICML/UAI/COLT Workhsop on Sparse Optimization and Variable Selection. Helsinki, 2008.

[7] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [M]. New York: Springer-Verlag, 2001. 193 – 210.

[8] Grunwald P D, Rissanen J. The Minimum Description Length Principle[M]. The MIT Press, 2007. 3 – 40.

[9] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression[J]. The Annals of statistics, 2004, 32(2): 407 – 451.

[10] Zou H. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418 – 1429.

[11] Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n[J]. Annals of Statistics, 2007, 35(6): 2313 – 2351.

[12] Efron B, Hastie T, Tibshirani R. Discussion: The dantzig selector: Statistical estimation when p is much larger than n[J]. The Annals of Statistics, 2007, 35(6): 2358 – 2364.

[13] Hansen M H, Yu B. Model selection and the principle of minimum description length[J]. Journal of the American Statistical Association, 2001, 96(454): 746 – 774.

[14] Fan J, Fan Y. High dimensional classification using features annealed independence rules[J]. Annals of statistics, 2008, 36(6): 2605 – 2637.

[15] Greenshtein E, Park J. Application of Non parametric empirical Bayes estimation to high dimensional classification[J]. Journal of Machine Learning Research, 2009, 10(7): 1687 – 1704.

[16] Rissanen J. Strong optimality of the normalized ML models as universal codes and information in data[J]. IEEE Transactions on Information Theory, 2001, 47(5): 1712 – 1717.

作者简介:

刘 峤 男,1974 年生于陕西西安,1996 年毕业于电子科技大学获学士学位,现为该校信息安全专业博士研究生,主要研究方向为统计机器学习算法及其在信息安全和生物信息学中的应用.

E-mail: cnliuqiao@gmail.com

王 娟 女,1980 年生于四川成都,讲师,博士生,2003 年于电子科技大学获计算机科学与技术学士学位,2006 年于电子科技大学获得计算机系统结构硕士学位.2007 – 2008 年作为访问学者在美国北卡罗来纳大学夏洛特分校进行网络异常相关研究.研究兴趣包括:数据挖掘,网络异常检测等.