



超平面支持向量机简化性能分析

程 辉¹, 方景龙¹, 王大全², 王兴起¹

(1. 杭州电子科技大学复杂系统建模与仿真教育部重点实验室 杭州 310018;

2. 杭州电子科技大学信息工程学院 杭州 310018)

摘 要: 与传统支持向量机相比, 针对复杂分类问题的超平面支持向量机和针对高噪声数据回归问题的回归型超平面支持向量机, 具有支持向量少、测试速度快、计算精度高的优点。然而对不同的样本集, 超平面支持向量机的简化效果有所不同, 仔细分析了这一现象的原因, 得出在支持向量中非约束支持向量所占比率越低则超平面支持向量机简化效果越明显的结论。

关键词: 模式识别; 超平面支持向量机; 简化支持向量机; 性能分析

doi: 10.11959/j.issn.1000-0801.2015161

Performance Analysis of Simplification of Hyperplane Support Vector Machine

Cheng Hui¹, Fang Jinglong¹, Wang Daquan², Wang Xingqi¹

(1. Education Ministry Key Laboratory of Complex Systems Modeling and Simulation,

Hangzhou Dianzi University, Hangzhou 310018, China;

2. Information Engineering School, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Comparing with traditional support vector machine, hyperplane support vector machine (HPSVM) and hyperplane support vector machine for regression (HPSVMR) not only reduce the number of support vectors and calculation time but also have comparable accuracy. However, they have different simplification effect on different problems. The reasons for this phenomenon were analyzed, the conclusion that the lower the percentage of non-bound support vectors is, the more obvious the effect of simplification was pointed out.

Key words: pattern recognition, hyperplane support vector machine, reduced support vector machine, performance analysis

收稿日期: 2015-01-05; 修回日期: 2015-06-15

基金项目: 国防基础科研计划基金资助项目 (No.JCKY2013415C001), 国防技术基础科研计划基金资助项目 (No.JSZL2014415B002), 武器装备预研基金资助项目 (No.9140A15040214DZ04221)

Foundation Items: Defense Basic Research Program (No.JCKY2013415C001), Defense Technical Basic Research Program (No.JSZL2014415B002), Weapon Equipment Advance Research Program (No.9140A15040214DZ04221)

论文引用格式: 程辉, 方景龙, 王大全等. 超平面支持向量机简化性能分析. 电信科学, 2015161

Cheng H, Fang J L, Wang D Q, *et al.* Performance analysis of simplification of hyperplane support vector machine. Telecommunications Science, 2015161

1 引言

支持向量机(support vector machine, SVM)的复杂度取决于支持向量的数目,减少支持向量的数目从而简化支持向量机是一项非常有意义的工作,到目前为止,国内外在简化支持向量机方面已经取得诸多研究成果。参考文献[1]通过在给定精度损失下生成支持向量缩减集合的方式得到一种简化支持向量机;参考文献[2]提出了一种简化支持向量机(reduced support vector machine, RSVM),在训练样本很多且支持向量所占比例极高的情况下,将支持向量机限制在一个训练样本子集中,能减少支持向量个数;参考文献[3]采用关系模糊 C-均值聚类(RFCM)算法来减少回归型支持向量的计算时间,在一定程度上降低了数据噪声和异常点的影响;参考文献[4]通过自适应遗传算法从受过训练的支持向量中选择最佳近似原始判别函数的向量解,从而优化 SVM 解向量;参考文献[5]提出一种平滑的可分示例近似求解(smoothed separable case approximation, SSCA)方法,用来减少支持向量数目;参考文献[6]允许用户通过迭代删除多余的支持向量,与当前逐一删除多余支持向量的做法相比,能够显著提高裁剪速度;参考文献[7]利用有向无环图(directed acyclic graph, DAG)简洁地表示代表支持向量机训练算法的树,通过避免共享结构上冗余的评估来减小总的计算量。参考文献[8,9]则提出了超平面支持向量机(hyperplane support vector machine, HPSVM)和回归型超平面支持向量机(hyperplane support vector machine for regression, HPSVMR),用错分间隔代替传统支持向量机中的错分样本与分类超平面距离之和,将其与机器泛化能力进行折中,从而简化 SVM 的支持向量数目。

但是在使用中发现, HPSVM 和 HPSVMR 对有些问题简化效果明显,对有些问题简化效果并不明显。为此仔细分析了支持向量的结构分布对 HPSVM 和 HPSVMR 简化效果的影响,发现非约束支持向量(non-bound support vector)占整个支持向量的比率和 HPSVM (HPSVMR)支持向量数与 SVM(SVMR)支持向量数的比值正相关,非约束支持向量在整个支持向量中的比率是影响 HPSVM 和 HPSVMR 简化效果的关键,支持向量中非约束支持向量比率越低, HPSVM 和 HPSVMR 简化效果就越明显。

2 超平面支持向量机

支持向量机考虑如下优化问题^[2],表示为:

$$\min_{w, b, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^L \xi_i$$

$$\text{s.t. } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i=1, 2, 3, \dots, L \quad (1)$$

其目标函数是最少错分样本和机器泛化能力(与 $w^T w$ 有关)的折中,其中, w 是超平面法向量, b 是偏置, $C > 0$ 是控制对错分样本惩罚程度的惩罚系数, $\phi(x_i)$ 是特征映射, ξ 是松弛因子, x_i 是训练样本, y_i 是 x_i 的标记,为 1 或 -1, L 是训练样本个数。

在实际求解中,不是求解式(1)中的问题,而是求解它的对偶问题^[2],表示为:

$$\max_{\alpha} \sum_{i=1}^L \alpha_i - \frac{1}{2} \sum_{i,j=1}^L \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{s.t. } \sum_{i=1}^L y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, i=1, 2, 3, \dots, L \quad (2)$$

其中, $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ 是核函数。这是一个二次凸规划问题,其局部最优解就是全局最优解。

支持向量机的分类判别函数是 $f(x) = \text{sgn}(w^T \phi(x) + b) = \text{sgn}(\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b)$, 其中 SV 是支持向量的下标集合(即对应的 $\alpha_i \neq 0$)。显然,支持向量个数越多,支持向量机越复杂,识别速度越慢。下面讨论影响支持向量个数的因素。

根据优化理论的 KKT 条件,对于式(1)和式(2)中的问题,有 $\alpha_i[y_i(w^T \phi(x) + b) - 1 + \xi_i] = 0, i=1, 2, 3, \dots, L$ 成立。所以,当 $\alpha_i \neq 0$ 时,必有 $y_i(w^T \phi(x) + b) - 1 + \xi_i = 0$ 成立。由于每个等式中有一个变量 ξ_i ,所以 $y_i(w^T \phi(x) + b) - 1 + \xi_i = 0$ 成立的可能性较大,从而传统支持向量机的支持向量较多。显然,如果减少变量 ξ_i 的个数,使得 $\alpha_i[y_i(w^T \phi(x) + b) - 1 + \xi_i] = 0$ 不能轻易成立,就有望降低支持向量个数。

为此, HPSVM 考虑如下优化问题^[8]:

$$\min_{w, b, \xi} \frac{1}{2} w^T w + C \xi$$

$$\text{s.t. } y_i(w^T \phi(x_i) + b) \geq 1, i \in V - MV; y_i(w^T \phi(x_i) + b) \geq 1 - \xi, i \in MV, \xi \geq 0 \quad (3)$$

其目标函数是最小错分间隔和机器泛化能力的折中,其中, MV 是一次训练形成的错分样本下标集合, V 是二次训练样本下标集合,可以将 V 取为整个训练样本的下标集



合,也可以为了减少二次训练时间,将 V 取为一次训练的支持向量的下标集合,但不管怎样,应有 $V \supseteq MV$ 。

根据优化理论的 KKT 条件,对于传统支持向量机,若 $\alpha_i=0$,则 $y_i(w^T\phi(x_i)+b) \geq 1$;若 $\alpha_i>0$,则 $y_i(w^T\phi(x_i)+b) \leq 1$ 。所以,传统支持向量机的支持向量是使 $y_i(w^Tx+b) \leq 1$ 成立的几乎所有样本,包括所有错分样本、超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 之间的所有样本以及超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 上面的部分样本。而 HPSVM 的支持向量是使 $y_i(w^Tx+b)=\min_j y_j(w^Tx_j+b)$ 成立的样本和使 $y_i(w^Tx+b)=1$ 成立的部分样本,仅仅包括错分样本中离分类超平面最远的样本、超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 上面的部分样本。

从式(1)和式(3)可明显看出,HPSVM 对训练样本的识别率必大于或等于传统支持向量机对训练样本的识别率。在测试样本识别率上,由于式(3)中的问题是用错分间隔代替式(1)中的错分样本离分类超平面距离之和,将其与机器泛化能力进行折中,所以一般情况下,HPSVM 对测试样本的识别率也会比传统支持向量机的识别率高。

同理,对于回归型支持向量机,可以构造一个 HPSVMR。根据优化理论的 KKT 条件,若 $\alpha_i=\alpha_i^*=0$,则 $|w^T\phi(x_i)+b-y_i| \leq \varepsilon$;若 $\alpha_i>0$ 或 $\alpha_i^*>0$,则 $|w^T\phi(x_i)+b-y_i| \geq \varepsilon$ 。所以,传统回归型支持向量机的支持向量是使 $|w^T\phi(x_i)+b-y_i| \geq \varepsilon$ 成立的几乎所有样本,其中包括所有超出允许误差 ε 的向量。而 HPSVMR 的支持向量是使 $|w^T\phi(x_i)+b-y_i|=\min_j |w^T\phi(x_j)+b-y_j|$ 成立的样本和使 $|w^T\phi(x_i)+b-y_i|=\varepsilon$ 成立的部分样本,仅仅包括超出允许误差 ε 最多的训练向量以及超平面 $w^T\phi(x_i)+b=y_i+\varepsilon$ 和 $w^T\phi(x_i)+b=y_i-\varepsilon$ 上面的部分样本。

3 超平面支持向量机简化性能分析

由于传统支持向量机的支持向量是使 $y_i(w^Tx+b) \leq 1$ 成立的几乎所有样本,包括所有错分样本、超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 之间所有样本以及超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 上面的部分样本。而 HPSVM 的支持向量是使 $y_i(w^Tx+b)=\min_j y_j(w^Tx_j+b)$ 成立的样本和使 $y_i(w^Tx+b)=1$ 成立的部分样本,仅仅包括错分样本中离分类超平面最远的样本以及超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 上面的部分样本。所以有理由相信,HPSVM 的支持向量要比传统支持向量机的支持向量少,而且,如果超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$

上的支持向量比较少,那么 HPSVM 的简化效果会更加明显。

由于 $0<\alpha<C$ 对应的支持向量必在超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 上, $\alpha=C$ 对应的支持向量满足 $y_i(w^T\phi(x_i)+b) \leq 1$,包括所有错分样本、超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 之间的所有样本以及超平面 $w^T\phi(x)+b=-1$ 和 $w^T\phi(x)+b=1$ 上面的部分样本。所以,非约束支持向量(对应 $0<\alpha<C$ 的支持向量)在整个支持向量中的比率是影响超平面支持向量机简化效果的关键,支持向量中非约束支持向量比率越低,HPSVM 的简化效果就越明显。

传统回归型支持向量机的支持向量是使 $|w^T\phi(x_i)+b-y_i| \geq \varepsilon$ 成立的几乎所有样本,其中包括所有超出允许误差 ε 的向量。而 HPSVMR 的支持向量是使 $|w^T\phi(x_i)+b-y_i|=\min_j |w^T\phi(x_j)+b-y_j|$ 成立的样本和使 $|w^T\phi(x_i)+b-y_i|=\varepsilon$ 成立的部分样本,仅仅包括超出允许误差 ε 最多的训练向量以及超平面 $w^T\phi(x_i)+b=y_i+\varepsilon$ 和 $w^T\phi(x_i)+b=y_i-\varepsilon$ 上面的部分样本。

所以,对于回归型问题,HPSVMR 的支持向量要比传统回归型支持向量机的支持向量少,而且,如果超平面 $w^T\phi(x_i)+b=y_i+\varepsilon$ 和 $w^T\phi(x_i)+b=y_i-\varepsilon$ 上的支持向量比较少,那么 HPSVMR 的简化效果会更加明显。由于 $0<\alpha<C$ 或 $0<\alpha^*<C$ 对应的支持向量必在超平面 $w^T\phi(x_i)+b=y_i+\varepsilon$ 和 $w^T\phi(x_i)+b=y_i-\varepsilon$ 上, $\alpha=C$ 或 $\alpha^*=C$ 对应的支持向量满足 $|w^T\phi(x_i)+b-y_i| \geq \varepsilon$,包括所有超出允许误差 ε 的向量。所以,非约束支持向量(对应 $0<\alpha<C$ 或 $0<\alpha^*<C$ 的支持向量)在整个支持向量中的比率也同样是影响 HPSVMR 简化效果的关键,支持向量中非约束支持向量比率越低,HPSVMR 的简化效果就越明显。

对于高噪声数据的回归问题,为了降低误差,允许误差 ε 不能取得太大,因而会有很多向量的误差超过允许误差 ε ,从而产生大量的支持向量,影响支持向量机的应用。采用 HPSVMR 可明显降低支持向量个数。

在公共测试数据库 UCI Adult、Breast Cancer、Heart、German 和 Splice 上对 SVM 和 HPSVM 方法进行了对比计算,在公用测试数据库 Abalone 上对 SVMR 和 HPSVMR 方法进行了对比计算,总共计算了 18 个例子。为方便起见,用 HPSVM1 表示二次训练采用整个训练样本的 HPSVM;用 HPSVM2 表示二次训练采用一次训练的支持向量的 HPSVM;用 HPSVMR1 表示二次训练采用整个训练样本的 HPSVMR;用 HPSVMR2 表示二次训练采用一次训练的支持向量的 HPSVMR。

计算结果见表 1~表 3,其中括号里的数据分别是非约束支持向量占整个支持向量的比率和 HPSVM1 (HPSVMR1)支持向量数与 SVM(SVMR)支持向量数的比值。由于 HPSVM1 和 HPSVM2 的计算结果差不多, HPSVMR1 和 HPSVMR2 的计算结果也差不多,所以只列出 HPSVM1 和 HPSVMR1 的计算结果。采用径向基核函数 $K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|)$ 进行计算,对于不同的测试样本,核函数参数和支持向量机惩罚参数的选取采用不同的数据。对于 UCI Adult 样本,采用参考文献[10]中的数据,即核参数取 $\gamma=0.005$,惩罚参数取 $C=1$;对于 Breast Cancer、Heart、German 和 Splice 样本,采用参考文献 [12] 中的数据,即

Breast Cancer 样本的核参数取 $\gamma=0.002$,Heart 样本的核参数取 $\gamma=0.008\ 3$,German 样本的核参数取 $\gamma=0.018\ 1$,Splice 样本的核参数取 $\gamma=0.014\ 2$,惩罚参数都取 $C=100$;对于 Abalone 样本,采用参考文献[12]中的数据,即核参数取 $\gamma=0.2$,惩罚参数取 $C=1\ 000$ 。

为了减小计算量,参考文献[10]中 A3a~A9a 的训练数据只取了前面一部分,所以表 1 的计算结果与参考文献[10]中的数据不完全一致。

由于 Breast Cancer、Heart 和 German 只有一组数据,所以把这组数据分为两部分,其中一部分作为训练样本,另一部分作为测试样本。

表 1 SVM 和 HPSVM 计算结果的比较 (UCI Adult)

数据集	支持向量个数				计算时间/ms		计算精度	
	$0<\alpha<C$	$\alpha=C$	SVM	HPSVM1	SVM	HPSVM1	SVM	HPSVM1
A1a	27	758	785	63	13 688	1 844	82.67%	82.65%
A2a	23	1 082	1 105	72	18 937	1 954	83.46%	83.405%
A3a	39	1 412	1 451	87	3 734	1 000	83.55%	83.52%
A4a	40	2 051	2 091	97	7 390	1 141	83.74%	83.86%
A5a	52	2 662	2 714	111	28 234	1 891	84.06%	84.10%
A6a	55	4 422	4 477	156	24 297	1 719	84.06%	84.10%
A7a	68	6 251	6 319	184	39 703	1 782	84.43%	84.48%
A8a	59	8 668	8 727	225	3 657	500	84.87%	84.87%
A9a	83	12 082	12 165	265	18 141	1 375	84.89%	84.80%

表 2 SVM 和 HPSVM 计算结果的比较 (Breast Cancer、Heart、German 和 Splice)

数据集	支持向量个数				计算精度	
	$0<\alpha<C$	$\alpha=C$	SVM	HPSVM1	SVM	HPSVM1
Breast Cancer	9	35	44	10	82.67%	82.65%
Heart	23	46	69	25	83.46%	83.40%
German	213	161	374	215	75.94%	75.94%
Splice	565	0	565	565	78.71%	79.81%

表 3 SVMR 和 HPSVMR 计算结果的比较 (Abalone)

数据集	ε	支持向量个数				均方误差	
		$0<\alpha<C$	$\alpha=C$	SVM	HPSVMR1	SVM	HPSVMR1
Abalone	4	63	242	305	61	5.52%	5.33%
	3	71	486	557	70	4.56%	4.68%
	2	82	952	103 4	93	4.18%	4.27%
	1.5	92	138 3	147 5	97	4.17%	4.26%
	1	104	201 1	211 5	101	4.21%	4.30%

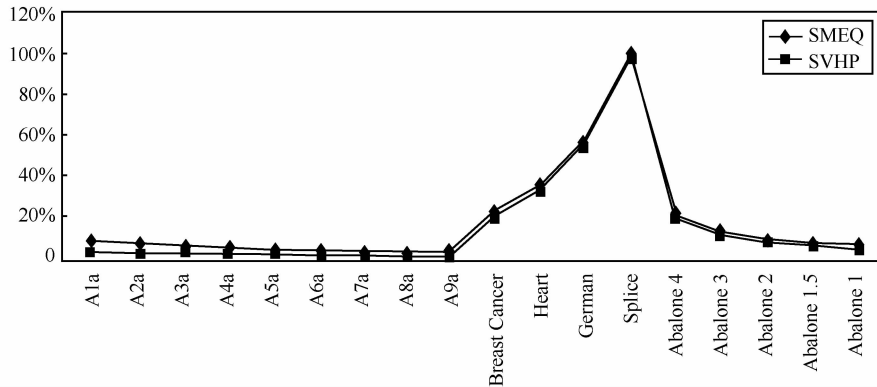


图1 非约束支持向量数占整个支持向量数的比率曲线 SMEQ 和 HPSVM(HPSVMR)
支持向量数占 SVM(SVMR)支持向量数比率曲线 SVHP

从表1~表3可以看出,对于这18个算例,HPSVM1和HPSVMR1的支持向量数明显比SVM和SVMR的支持向量数少得多,最低只有其2.2%,从而使得HPSVM1和HPSVMR1的计算时间(分类识别或数据拟合)比SVM和SVMR少得多,而计算精度则大体相当。

为了看清楚非约束支持向量占整个支持向量总的比率对超平面支持向量机简化效果的影响,将非约束支持向量占整个支持向量的比率曲线SMEQ和HPSVM(HPSVMR)支持向量数与SVM(SVMR)支持向量数的比值曲线SVHP画在了一张图上,如图1所示。从图1中可以看出,比率数据正相关非常明显,这充分说明了非约束支持向量在整个支持向量中的比率是影响超平面支持向量机和回归型超平面支持向量机简化效果的关键,支持向量中非约束支持向量比率越低,超平面支持向量机和回归型超平面支持向量机的简化效果就越明显。

4 结束语

HPSVM和HPSVMR对不同问题的简化效果有所不同。本文通过分析得出结论:非约束支持向量占整个支持向量的比率和HPSVM(HPSVMR)支持向量数与SVM(SVMR)支持向量数的比值正相关;非约束支持向量在整个支持向量中的比率是影响HPSVM和HPSVMR简化效果的关键;支持向量中非约束支持向量比率越低,HPSVM和HPSVMR简化效果就越明显。

本文的结论也充分说明HPSVM和HPSVMR对支持向量的减少程度与分类问题或拟合问题的复杂度有关,错分样本越多或超出允许误差的样本越多,HPSVM和HPSVMR的简化效果就越明显,如果错分样本很少或基本

上没有超出允许误差的样本,HPSVM和HPSVMR的简化效果就不够明显,所以说HPSVM和HPSVMR针对复杂问题的简化效果更好。

参考文献

- Burges C J C. Simplified support vector decision rule. Proceedings of the 13th International Conference on Machine Learning, San Mateo, CA, USA, 1996: 71~77
- Lin M K, Lin J C. A study on reduced support vector machines. IEEE Transactions on Neural Networks, 2003, 14(6): 1449~1459
- Shieh L H, Kuo C C. A reduced data set method for support vector regression. Expert Systems with Applications, 2010, 37(12): 7781~7787
- Lin H J, Yeh J P. Optimal reduction of solutions for support vector machines. Applied Mathematics and Computation, 2009, 214(2): 329~335
- Geebelen D, Suykens J A K, Vandewalle J. Reducing the number of support vectors of SVM classifiers using the smoothed separable case approximation. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(4): 682~688
- Liang X, Ma Y, He Y, et al. Fast pruning superfluous support vectors in SVMs. Pattern Recognition Letters, 2013, 34(10): 1203~1209
- Severyn A, Moschitti A. Fast support vector machines for convolution tree kernels. Data Mining and Knowledge Discovery, 2012, 25(2): 325~357
- 方景龙, 陈钰, 潘志庚等. 复杂分类问题支持向量机的简化. 电子学报, 2007, 35(5): 858~861
- Fang J L, Chen S, Pan Z G, et al. A simplification to support vector machine for complicated recognition problem. Acta Electronica Sinica, 2007, 35(5): 858~861
- Fang J L, Chen S, Liang X W, et al. An improved model of support

vector machine for high noise data regression problem. Proceedings of the IET China Ireland International Conference on Information and Communications Technologies, Dublin, Ireland, 2007

- 10 Schölkopf B, Burges C J C, Smola A J. Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1998
- 11 Lee M, Keerthi S, Ong C, *et al.* An efficient method for computing leave-one-out error in support vector machines with Gaussian kernels. IEEE Transactions on Neural Networks, 2004, 15 (3): 750~757
- 12 全勇, 杨杰, 姚莉秀等. 基于连续过松弛方法的支持向量回归算法. 软件学报, 2004, 15(2): 200~206
Quan Y, Yang J, Yao X L, *et al.* Successive over relaxation for support vector regression. Journal of Software, 2004, 15(2): 200~206

[作者简介]



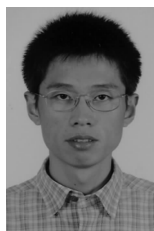
程辉,男,杭州电子科技大学硕士生,主要研究方向为机器学习。



方景龙[通信作者],男,博士,杭州电子科技大学研究员、博士生导师,主要研究方向为机器学习、软件工程。



王大全,男,杭州电子科技大学信息工程学院讲师,主要研究方向为嵌入式系统、智能识别。



王兴起,男,博士,杭州电子科技大学副教授,主要研究方向为数据挖掘、软件测试。

2015161-6

勘误说明

王忠伟,陈叶芳,肖四友等同志所撰写的稿件《一种高维大数据全 k 近邻查询算法》(编号:2015171)已在我刊 2015 年第 7 期上刊登,刊登的作者简介中“肖四友[通信作者],男,博士,宁波大学副教授”,应更改为“肖四友[通信作者],男,宁波大学博士生”,特此更正,并向读者、作者致歉!

《电信科学》编辑部
2015 年 8 月