应用概率统计　第三十卷
第四期　2014年8月

Chinese Journal of Applied Probability
and Statistics Vol.30 No.4 Aug. 2014

基于组块3×2交叉验证的预测误差估计的方差 *

杨杏丽　　王　钰　　王瑞波　　李济洪*

(山西大学数学科学学院, 太原, 030006)　　(山西大学计算中心, 太原, 030006)

摘　要

本文对文献中新提出的预测误差的组块3×2交叉验证估计的方差进行了研究, 给出了其方差的更为精细的表达式, 且从理论上证明了不存在其方差的通用(对所有分布都适用的)无偏估计.

关键词:　组块3×2交叉验证, 无偏估计, 预测误差估计的方差.

学科分类号:　O212.

§1. 引　　言

在统计机器学习中, 常常需要进行算法性能好坏的评价, 而从统计意义上来说就是检验不同算法的预测误差的估计是否有显著差异, 就需要构造检验统计量. 而统计量的构造, 需要估计预测误差估计量的方差. 交叉验证(Cross-Validation, 简记为CV)是一种简单、实用的数据重用方法, 被广泛应用于给定算法的预测误差(Prediction Error)估计. 实际应用中交叉验证也有多种形式, 正如文献[1]的综述文章中所提到的包括标准K折交叉验证, RLT (Repeated Learning-Testing)交叉验证, MCCV (Monte-Carlo CV)等. 在这些交叉验证中, 标准2折交叉验证相对计算量较小, 且在分类学习算法的模型选择问题中具有选择的一致性(见文献[2]). 因此, 在模型选择, 算法性能对照等问题中2折交叉验证有特别的价值.

为了提高估计的精度, 在计算量容许下, 人们考虑对同一个样本的多次随机划分, 重复进行多次2折交叉验证. 重复实验是统计中得到方差估计以及进行假设检验的前提. 比如, 较为有影响的是5×2交叉验证检验方法. [3][4]在研究两个算法的性能对照时, 重复了5次标准2折交叉验证, 而每次交叉验证都是通过随机把训练集切分为2份而进行的, 其文中称为5×2交叉验证方法. 以5次交叉验证的结果获得预测误差的估计及其方差的估计, 最后通过模拟实验证明了其方法比其它方法有更好的性能. 特别地, [5]将数据集均衡地分为四份, 每一份所包含的样本数大致相同, 并且每一类别在每份数据集中的个数也大致相同, 然后将四份中的任两份两两结合, 将其中互为补集的两个组合看为对数据集的一次切分, 这样可以得到3组数据集的切分, 在每一组数据集的切分上做一次标准的2折交叉验证, 然后将3组交叉验证的估计平均起来作为对预测误差的估计, [5]把这种交叉验证称为组块3×2交

*山西省科技基础条件平台建设项目(20130910030101)资助.

*通讯作者, E-mail: lijh@sxu.edu.cn.

本文2014年2月20日收到, 2014年5月15日收到修改稿.

doi: 10.3969/j.issn.1001-4268.2014.04.004

叉验证. 随后, [5]在块层面分析并给出了组块3×2交叉验证估计的方差, 并证明了基于组块3×2交叉验证检验的性能与文献[3][4]给出的5×2交叉验证检验可比, 但却有较少的计算量, 并且数据的切分方式可以较好地平衡分类数据的各个类别. 然而, 文献[5]中并没有给出组块3×2交叉验证更为精细的方差表达式. 并且我们注意到无论是5×2交叉验证检验中的方差的样本方差估计, 还是组块3×2交叉验证检验中的方差的保守估计, 都不是其方差的无偏估计, 但是方差的估计好坏又直接影响检验的性能. 那么, 预测误差的组块3×2交叉验证估计的方差是否和K折交叉验证估计的方差一样不存在无偏估计呢?

基于此, 本文给出了预测误差的组块3×2交叉验证估计的方差的更为精细的表达式, 并依此证明其方差不存在通用(对所有分布都适用的)无偏估计.

§2. 预测误差的组块3×2交叉验证估计

设从分布$\mathcal{F}$中得到容量为n的样本集为$D_n=\{\xi_1,\xi_2,\ldots,\xi_n\}$, $\xi_j=(x_j,y_j)$. x_j是输入向量, y_j是输出变量. 如果$f=\mathcal{A}(D_n)$表示在数据集D_n上训练由算法$\mathcal{A}$返回的预测函数, 损失函数$L(f(x),y)$ 表示预测与观测之间差异的度量(其中$L(f(x),y)$可以是连续的(如平方损失), 也可以是离散的(如$L(f(x),y)=I[f(x)\neq y]$, 即0-1损失)), 那么算法(模型)的预测误差定义为

$$\mu\triangleq \mathrm{PE}(\mathcal{A}(D_n))=\mathsf{E}_{\xi}[L(\mathcal{A}(D_n),\xi)],$$

这里, $\xi\backsim\mathcal{F}$.

所谓组块3×2交叉验证是指将数据D_n均匀地分为大致相等且不相交的四个子集, 记为L_j, $j=1,2,3,4$, 然后两两结合, 组成3组6个不同的组合: $\{(L_1,L_2),(L_3,L_4)\}$, $\{(L_1,L_3),(L_2,L_4)\}$, $\{(L_1,L_4),(L_2,L_3)\}$. 在每一组上做一个标准的2折交叉验证, 最后把3组2折交叉验证的结果进行平均, 即为组块3×2交叉验证, 具体地:

$$\widehat{\mu}_{3\times2}\triangleq\frac{1}{3}\sum_{i=1}^{3}\widehat{\mu}^{(i)}=\frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2}\widehat{\mu}_k^{(i)},$$

这里, $\widehat{\mu}_{3\times2}$表示的是预测误差的组块3×2交叉验证估计, $\widehat{\mu}_k^{(i)}$为第i组, 第k折交叉验证所得到的预测误差的估计, $\widehat{\mu}^{(i)}=(1/2)\sum\limits_{k=1}^{2}\widehat{\mu}_k^{(i)}$. 当把$\widehat{\mu}_{3\times2}$表示到样本层面上时, 有

$$\widehat{\mu}_{3\times2}\triangleq\frac{1}{3}\sum_{i=1}^{3}\frac{1}{n}\sum_{j=1}^{n}e_j^{(i)},\tag{2.1}$$

其中, $e_j^{(i)}=L(\mathcal{A}(D_n^{(i)}),\xi_j)$, $D_n^{(i)}$为第i组的训练集.

为了进一步深入分析组块3×2交叉验证的估计, 做以下的记号: 设$\mathcal{I}=\{1,2,\ldots,n\}$为数据集$D_n$的角标集, $\mathcal{S}_i=\{I_i^{(1)},I_i^{(2)}\}$, $i=1,2,3$是对D_n角标集的第i次切分, 即要求对任意的i有, $I_i^{(1)}\cap I_i^{(2)}=\Phi$, $I_i^{(1)}\cup I_i^{(2)}=\mathcal{I}$, 且要求每个$I_i^{(1)}$和$I_i^{(2)}$包含的角标个数相同(这里总假设$n$为4的倍数).

§3. $\widehat{\mu}_{3\times 2}$的方差

当$i=i'$, $i,i'=1,2,3$时, 记

$$\mathsf{Cov}\,(e_j^{(i)},e_{j'}^{(i)})=\begin{cases}\sigma^2 & j=j',\ j,j'\in\mathcal{I};\\ \omega & j\neq j',\ j,j'\in I_i^{(k)},\ k=1,2;\\ \gamma & j\in I_i^{(k)},\ j'\in I_i^{(3-k)},\ k=1,2.\end{cases}$$

此时, 由于$i=i'$, 考虑的是一个标准2折交叉验证的估计在样本层面上的协方差, 我们所作的假设与文献[6]一致. 当文献[6]中交叉验证的折数$K=2$时, 训练集的大小为$n/2$, σ^2是此时测试样例的损失的真正方差; ω是测试样例的块内协方差, 是由相同的训练集而导致的; γ是测试样例的块间协方差, 是由2折交叉验证的测试集和训练集互换引起的.

当$i\neq i'$, $i,i'=1,2,3$时, 记

$$\mathsf{Cov}\,(e_j^{(i)},e_{j'}^{(i')})=\begin{cases}\sigma^* & j=j',\ j,j'\in\mathcal{I};\\ \omega^* & j\neq j',\ j,j'\in\{I_i^{(k)}\cap I_{i'}^{(k')}\},\ k,k'=1,2;\\ \gamma^* & j\in\{I_i^{(k)}\cap I_{i'}^{(k')}\},\ j'\in\{I_i^{(3-k)}\cap I_{i'}^{(3-k')}\},\ k,k'=1,2;\\ \tau^* & j\in\{I_i^{(k)}\cap I_{i'}^{(k')}\},\ j'\in\{I_i^{(k)}\cap I_{i'}^{(3-k')}\}\cup\{I_i^{(3-k)}\cap I_{i'}^{(k')}\},\ k,k'=1,2.\end{cases}$$

此时, $i\neq i'$, $\mathsf{Cov}\,(e_j^{(i)},e_{j'}^{(i')})$与$\mathsf{Cov}\,(e_j^{(i)},e_{j'}^{(i)})$不同, $\mathsf{Cov}\,(e_j^{(i)},e_{j'}^{(i)})$是同组内不同样本之间的协方差, 而$\mathsf{Cov}\,(e_j^{(i)},e_{j'}^{(i')})$是不同组的两样本之间的协方差. 此时, 由于任意两组数据的切分具有组块的特性, 两组之间的训练集有$n/4$的样本相同, σ^*应与σ^2不同, 它表示训练集有$n/4$的样本相同, 而测试样例相同时的协方差; 同样的, ω^*表示在每一块上测试样例不同时的协方差; γ^*表示的是训练集有$n/4$的样本相同时, 第i组的测试样例j出现在第i'组的训练集中, 而第i'组的测试样例j'出现在第i组的训练集时两组之间的测试样例的协方差; τ^*则是第i组的测试样例j出现在第i'组的训练集中, 而另一组i'的测试样例j'却没有出现在第i组的训练集时的协方差.

命题 3.1 组块3×2交叉验证估计的方差为

$$\mathsf{Var}\,(\widehat{\mu}_{3\times 2})=\frac{1}{3n}\sigma^2+\frac{n-2}{6n}\omega+\frac{1}{6}\gamma+\frac{2}{3n}\sigma^*+\frac{n-4}{6n}\omega^*+\frac{1}{6}\gamma^*+\frac{1}{3}\tau^*. \tag{3.1}$$

证明: 组块3×2交叉验证估计的方差为

$$\mathsf{Var}\,(\widehat{\mu}_{3\times 2})=\frac{1}{9n^2}\sum_{i,i'=1}^{3}\sum_{j,j'=1}^{n}\mathsf{Cov}\,(e_j^{(i)},e_{j'}^{(i')}).$$

记组块3×2交叉验证的损失向量为$e'=(e_1^{(1)},\ldots,e_n^{(1)},e_1^{(2)},\ldots,e_n^{(2)},e_1^{(3)},\ldots,e_n^{(3)})$, 则$e$的协方差矩阵有下图的块结构:

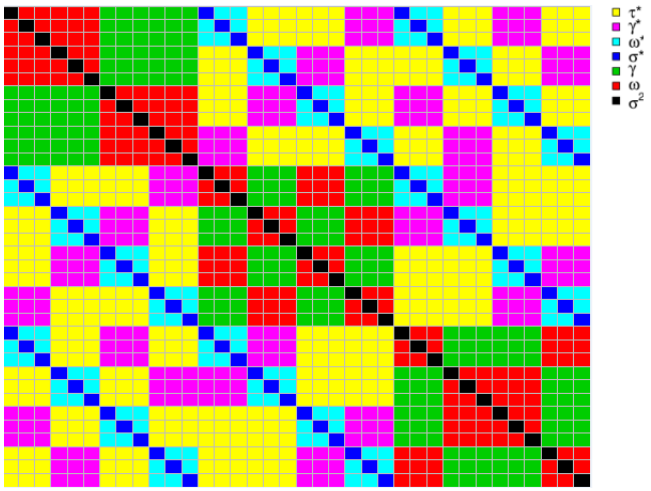

图1　组块3×2交叉验证损失向量的协方差矩阵

这时, 组块3×2交叉验证估计的方差可以写为

$$\begin{aligned}\mathrm{Var}\,(\widehat{\mu}_{3\times 2}) &= \frac{1}{9n^2}\Big\{\sum_{i=1}^{3}\Big[n\sigma^2+n\Big(\frac{n}{2}-1\Big)\omega+\frac{n^2}{2}\gamma\Big]\\ &\quad+\sum_{i=1}^{3}\sum_{i'\neq i}\Big[n\sigma^*+n\Big(\frac{n}{4}-1\Big)\omega^*+\frac{n^2}{4}\gamma^*+\frac{n^2}{2}\tau^*\Big]\Big\}\\ &= \frac{1}{3n}\sigma^2+\frac{n-2}{6n}\omega+\frac{1}{6}\gamma+\frac{2}{3n}\sigma^*+\frac{n-4}{6n}\omega^*+\frac{1}{6}\gamma^*+\frac{1}{3}\tau^*.\end{aligned}$$

证毕. □

模拟实验 1　组块3×2交叉验证估计的方差

设$\xi=(X,Y)$, $\mathrm{P}(Y=1)=\mathrm{P}(Y=0)=1/2$, $X|Y=0\sim N(0,I_{10})$, $X|Y=1\sim N(1,2I_{10})$. 使用分类树作为分类器, 考虑样本个数$n=64,128,256,512,1024$时, 组块3×2交叉验证方差中七部分的贡献.

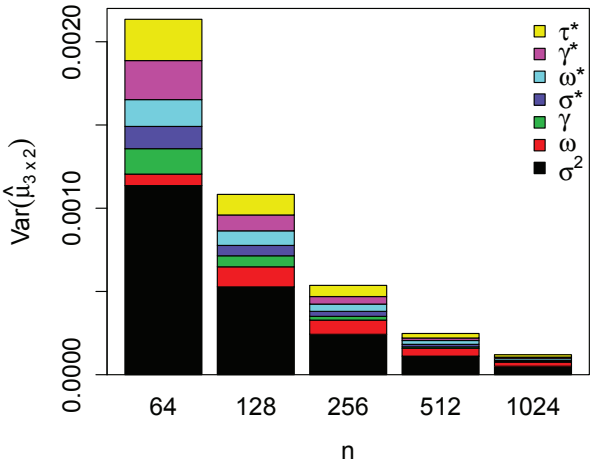

图2　组块3×2交叉验证估计的方差中$\sigma^2,\omega,\gamma,\sigma^*,\omega^*,\gamma^*,\tau^*$七部分占总方差的柱状图

从图2可以看出, 除了σ^2之外其它六部分的总和对组块3×2交叉验证估计的方差的贡献是不可忽略的, 则仅考虑σ^2而忽略其它六部分会对组块3×2交叉验证估计的方差造成很大的偏差. 另外, 随着样本个数n的增加, 可以发现σ^2逐渐减小, 并且对总方差的贡献从53.2%降低到42.5%, 而其它六部分虽然也随着样本个数的增加而减小, 但是其它六部分的总和对总方差的贡献却在增加.

§4. 基于组块3×2交叉验证的预测误差估计的方差的通用无偏估计不存在

在统计机器学习中, 一般不对总体的分布形式做任何假定, 所以, 是否能得到预测误差估计的方差的一个通用(对所有分布都适用的)无偏估计就十分有意义. 许多学者都试图找到这样的一个无偏估计, 如文献[7]就给出K折交叉验证作为预测误差估计时方差的两个估计, 但这两个估计或者是过估或者是欠估了这个方差, 都不是方差的无偏估计. 随后文献[6]证明了, 在标准的K折交叉验证情形下, 不存在其方差的通用无偏估计. 所谓的通用(universal)在文献[6]中指与总体分布无关, 换句话说就是适用于所有分布. 但当误差的分布和具体学习算法给定时, 对K折交叉验证来说是否也不存在方差的无偏估计? 文献[8]就给出了K折交叉验证作为预测误差估计时方差的一个几乎是无偏的估计, 但这个结论的得出依赖于具体的学习算法和误差的分布.

那么在组块的实验设置下, 3×2交叉验证估计的方差的无偏估计是否也不存在呢? 首先我们实验验证了文献[3][5]中给出的方差估计都不是无偏估计. 文献[3]考虑的是5×2交叉验证, 它把5×2交叉验证看成是标准2折交叉验证的简单重复, 记每次2折交叉验证的样本方差为$S_i^2=(\widehat{\mu}_1^{(i)}-\widehat{\mu}^{(i)})^2+(\widehat{\mu}_2^{(i)}-\widehat{\mu}^{(i)})^2$, $i=1,2,\ldots,5$, 则5×2交叉验证的方差估计为$(1/5)\sum\limits_{i=1}^{5}S_i^2$, 为了和组块$3\times2$交叉验证的方差作比较, 我们把文献[3]中的方差改为3×2, 记为$\widehat{\mathsf{Var}}_2(\widehat{\mu}_{3\times2})=(1/3)\sum\limits_{i=1}^{3}S_i^2$. 文献[5]研究的是组块$3\times2$交叉验证, 对这个方差的估计它同时考虑了组内方差和组间方差, 给出了方差的一个保守估计, 记为$\widehat{\mathsf{Var}}_1(\widehat{\mu}_{3\times2})=(1/6)\cdot\sum\limits_{i=1}^{3}\sum\limits_{k=1}^{2}(\widehat{\mu}_k^{(i)}-\widehat{\mu}_{3\times2})^2$. 以上的记号的意义与第2节相同.

模拟实验 2 组块3×2交叉验证估计的方差和它的两个估计

考虑和模拟实验1相同的设置, 分类器仍然是分类树, 但这时我们考虑的样本量$n=40,60,80,100,120,140,160,180,200,500,1000$时, 组块$3\times2$交叉验证估计的真实方差和它的两个估计值.

从图3可以看出, 显然这两个估计都不是组块3×2交叉验证的方差的无偏估计, 两者都大于组块3×2交叉验证的真实方差. 文献[5]中给出的保守估计相比样本方差估计来说要好点, 但差别不大. 以下, 我们来证明组块3×2交叉验证估计的方差不存在对所有分布都适用

的无偏估计.

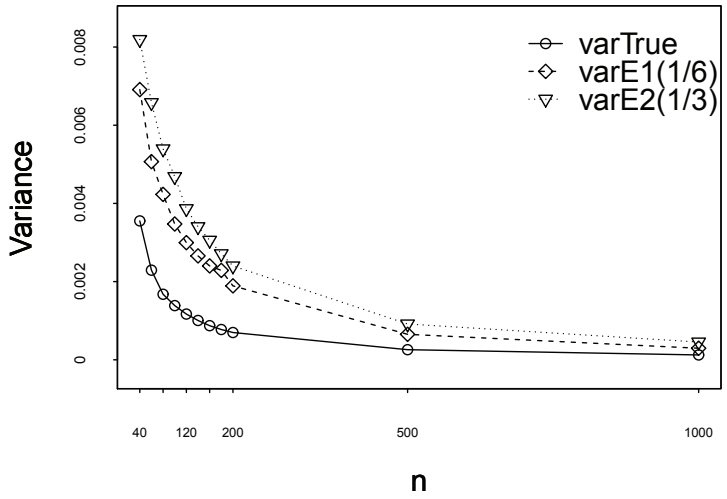

图3　图中带圈的线表示的是组块3×2交叉验证估计的真实方差, 带菱形的线表示的是文献[5]中给出方差的保守估计$\widehat{\mathsf{Var}}_1(\hat{\mu}_{3\times 2})$, 带三角形的线是方差的样本方差估计$\widehat{\mathsf{Var}}_2(\hat{\mu}_{3\times 2})$

假设组块3×2交叉验证估计的方差的估计$\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2})$可以用下列的Taylor展式表示:

$$\begin{aligned}\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2}) &= \alpha_0+\sum\alpha_1(i,j)e_j^{(i)}+\sum\alpha_2(i,j,i',j')e_j^{(i)}e_{j'}^{(i')}\\ &\quad+\sum\alpha_3(i,j,i',j',i'',j'')e_j^{(i)}e_{j'}^{(i')}e_{j''}^{(i'')}+\cdots.\end{aligned}\tag{4.1}$$

同文献[6]的分析, 要说明估计的无偏性, 即要说明$\mathsf{Var}(\hat{\mu}_{3\times 2})=\mathsf{E}(\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2}))$, 仔细观察$\mathsf{Var}(\hat{\mu}_{3\times 2})$, 它不包含与$i,j$无关的常数项, 则$\alpha_0=0$, 若设对任意的$i,j$有$\mathsf{E}(e_j^{(i)})=\mu$, 而$\mathsf{Var}(\hat{\mu}_{3\times 2})$中不包含$\mu$这一项, 则应有$\alpha_1(i,j)=0$. 同理, 可以推出在$\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2})$的表达式中除了$e_j^{(i)}e_{j'}^{(i')}$的系数不为0外, 其余全为0. 不妨设$\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2})$的表达式为: $\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2})\triangleq\sum\limits_{i,j,i',j'}W_{jj'}^{ii'}e_j^{(i)}e_{j'}^{(i')}$, 这和我们一般把方差的估计看成是预测误差的二次型的观点相一致.

命题 4.1　对$\mathsf{Var}(\hat{\mu}_{3\times 2})$的估计不存在通用(对所有分布都适用的)无偏估计.

证明:　一般来说, 与分布无关的方差估计采用类似于矩估计的形式, 可设组块3×2交叉验证估计的方差的估计为$\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2})\triangleq\sum\limits_{i,j,i',j'}W_{jj'}^{ii'}e_j^{(i)}e_{j'}^{(i')}$, 对任意的$i,j$, $i=1,2,3$, $j=1,2,\ldots,n$, 设$\mathsf{E}(e_j^{(i)})=\mu$.

$$\begin{aligned}\mathsf{E}(\widehat{\mathsf{Var}}(\hat{\mu}_{3\times 2})) &= \sum_{i,j,i',j'}W_{jj'}^{ii'}\mathsf{E}(e_j^{(i)}e_{j'}^{(i')})\\ &= \sum_{i=i',j,j'}W_{jj'}^{ii}\mathsf{E}(e_j^{(i)}e_{j'}^{(i)})+\sum_{i\neq i',j,j'}W_{jj'}^{ii'}\mathsf{E}(e_j^{(i)}e_{j'}^{(i')}).\end{aligned}$$

等式右边第一部分为

$$\begin{aligned}\sum_{i=i',j,j'} W_{jj'}^{ii}\mathsf{E}(e_j^{(i)}e_{j'}^{(i)}) &= \sum_{i=1}^{3}\sum_{j=j'\in\mathcal{I}} W_{jj}^{ii}\mathsf{E}(e_j^{(i)})^2+\sum_{i=1}^{3}\sum_{k=1}^{2}\sum_{j\in I_i^{(k)}}\sum_{j'\neq j,j'\in I_i^{(k)}} W_{jj'}^{ii}\mathsf{E}(e_j^{(i)}e_{j'}^{(i)})\\ &\quad+\sum_{i=1}^{3}\sum_{k=1}^{2}\sum_{j\in I_i^{(k)}}\sum_{j'\in I_i^{(3-k)}} W_{jj'}^{ii}\mathsf{E}(e_j^{(i)}e_{j'}^{(i)})\\ &= (\sigma^2+\mu^2)\sum_{i=1}^{3}\sum_{j=1}^{n} W_{jj}^{ii}+(\omega+\mu^2)\sum_{i=1}^{3}\sum_{k=1}^{2}\sum_{j\in I_i^{(k)}}\sum_{j'\neq j,j'\in I_i^{(k)}} W_{jj'}^{ii}\\ &\quad+(\gamma+\mu^2)\sum_{i=1}^{3}\sum_{k=1}^{2}\sum_{j\in I_i^{(k)}}\sum_{j'\in I_i^{(3-k)}} W_{jj'}^{ii},\end{aligned}$$

等式右边第二部分为

$$\begin{aligned}\sum_{i\neq i',j,j'} W_{jj'}^{ii'}\mathsf{E}(e_j^{(i)}e_{j'}^{(i')}) &= \sum_{i=1}^{3}\sum_{i'\neq i}\sum_{j=j'\in\mathcal{I}} W_{jj}^{ii'}\mathsf{E}(e_j^{(i)}e_j^{(i')})\\ &\quad+\sum_{i=1}^{3}\sum_{i'\neq i}\sum_{k=1}^{2}\sum_{k'=1}^{2}\sum_{j\in\{I_i^{(k)}\cap I_{i'}^{(k')}\}}\sum_{j'\neq j,j'\in\{I_i^{(k)}\cap I_{i'}^{(k')}\}} W_{jj'}^{ii'}\mathsf{E}(e_j^{(i)}e_{j'}^{(i')})\\ &\quad+\sum_{i=1}^{3}\sum_{i'\neq i}\sum_{k=1}^{2}\sum_{k'=1}^{2}\sum_{j\in\{I_i^{(k)}\cap I_{i'}^{(k')}\}}\sum_{j'\in\{I_i^{(3-k)}\cap I_{i'}^{(3-k')}\}} W_{jj'}^{ii'}\mathsf{E}(e_j^{(i)}e_{j'}^{(i')})\\ &\quad+\sum_{i=1}^{3}\sum_{i'\neq i}\sum_{j,j'\in\text{o.w.}} W_{jj'}^{ii'}\mathsf{E}(e_j^{(i)}e_{j'}^{(i')})\\ &= (\sigma^*+\mu^2)\sum_{i=1}^{3}\sum_{i'\neq i}\sum_{j=1}^{n} W_{jj}^{ii'}\\ &\quad+(\omega^*+\mu^2)\sum_{i=1}^{3}\sum_{i'\neq i}\sum_{k=1}^{2}\sum_{k'=1}^{2}\sum_{j\in\{I_i^{(k)}\cap I_{i'}^{(k')}\}}\sum_{j'\neq j,j'\in\{I_i^{(k)}\cap I_{i'}^{(k')}\}} W_{jj'}^{ii'}\\ &\quad+(\gamma^*+\mu^2)\sum_{i=1}^{3}\sum_{i'\neq i}\sum_{k=1}^{2}\sum_{k'=1}^{2}\sum_{j\in\{I_i^{(k)}\cap I_{i'}^{(k')}\}}\sum_{j'\in\{I_i^{(3-k)}\cap I_{i'}^{(3-k')}\}} W_{jj'}^{ii'}\\ &\quad+(\tau^*+\mu^2)\sum_{i=1}^{3}\sum_{i'\neq i}\sum_{j,j'\in\text{o.w.}} W_{jj'}^{ii'}.\end{aligned}$$

当把$(\sigma^2+\mu^2)$, $(\omega+\mu^2)$, $(\gamma+\mu^2)$, $(\sigma^*+\mu^2)$, $(\omega^*+\mu^2)$, $(\gamma^*+\mu^2)$, $(\tau^*+\mu^2)$前的系数相应地记为a, b, c, a', b', c', d'后有

$$\begin{aligned}\mathsf{E}(\widehat{\mathsf{Var}}(\widehat{\mu}_{3\times2})) &= a(\sigma^2+\mu^2)+b(\omega+\mu^2)+c(\gamma+\mu^2)+a'(\sigma^*+\mu^2)\\ &\quad+b'(\omega^*+\mu^2)+c'(\gamma^*+\mu^2)+d'(\tau^*+\mu^2)\\ &= (a+b+c+a'+b'+c'+d')\mu^2+a\sigma^2+b\omega+c\gamma\\ &\quad+a'\sigma^*+b'\omega^*+c'\gamma^*+d'\tau^*.\end{aligned}$$

要使得$\mathsf{E}(\widehat{\mathsf{Var}}\,(\widehat{\mu}_{3\times2}))=\mathsf{Var}\,(\widehat{\mu}_{3\times2})$, 则须使每一项前的系数相等, 这样得到下列的等式组:

$$\begin{cases} a+b+c+a'+b'+c'+d'=0,\\ a=1/(3n),\\ b=(n-2)/(6n),\\ c=1/6,\\ a'=2/(3n),\\ b'=(n-4)/(6n),\\ c'=1/6,\\ d'=1/3. \end{cases}$$

明显有上述等式组不成立, 则说明基于组块3×2交叉验证的预测误差估计的方差没有对所有分布都适用的无偏估计. 证毕. □

§5. 结论与展望

这篇文章主要分析了最近提出的组块3×2交叉验证的预测误差估计的方差, 并在组块的定义下, 根据一些可以接受的基本假设, 可以写出组块3×2交叉验证的方差更为精细的表达式. 最后, 我们基于此表达式证明了组块3×2交叉验证的方差不存在对所有分布都适用的无偏估计. 但当给定具体的学习算法和误差的分布时, 是否存在无偏估计就需要深入分析.

更进一步, 是否也可以构造出组块m×2交叉验证? 和组块3×2交叉验证相比, 组块m×2交叉验证有什么优点和缺点? 组块$m\times2$交叉验证的方差是否也不存在对所有分布都适用的无偏估计? 这些问题都是我们进一步研究的方向.

参 考 文 献

[1] Arlot, S. and Celisse, A., A survey of cross-validation procedures for model selection, *Statistics Surveys*, **4**(2010), 40–79.

[2] Yang, Y., Comparing learning methods for classification, *Statistica Sinica*, **16(2)**(2006), 635–657.

[3] Dietterich, T.G., Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Computation*, **10(7)**(1998), 1895–1923.

[4] Alpaydin, E., Combined 5×2 cv F test for comparing supervised classification learning algorithms, *Neural Computation*, **11(8)**(1999), 1885–1892.

[5] Wang, Y., Wang, R., Jia, H. and Li, J., Blocked 3×2 cross-validated t-test for comparing supervised classification learning algorithms, *Neural Computation*, **26(1)**(2014), 208–235.

[6] Bengio, Y. and Grandvalet, Y., No unbiased estimator of the variance of K-fold cross-validation, *Journal of Machine Learning Research*, **5**(2004), 1089–1105.

[7] Nadeau, C. and Bengio, Y., Inference for the generalization error, *Machine Learning*, **52(3)**(2003), 239–281.

[8] Markatou, M., Tian, H., Biswas, S. and Hripcsak, G., Analysis of variance of cross-validation estimators of the generalization error, *Journal of Machine Learning Research*, **6**(2005), 1127–1168.

Variance of Estimator of the Prediction Error Based on Blocked 3×2 Cross-Validation

Yang Xingli

(*School of Mathematical Sciences, Shanxi University, Taiyuan, 030006*)

Wang Yu Wang Ruibo Li Jihong

(*Computer Center of Shanxi University, Taiyuan, 030006*)

This paper studies the variance of blocked 3×2 cross-validation estimator of the prediction error recently proposed in the literature. A more accuracy representation of the variance is provided and the main theorem shows that there exists no universal (valid under all distributions) unbiased estimator of the variance.

Keywords: Blocked 3×2 cross-validation, unbiased estimator, variance of estimator of the prediction error.

AMS Subject Classification: 62F10, 62F40, 62F86.