DOI:10.16016/j.1000-5404.1982.03.007

引入变量对减少错分率的效力

——一个定性分类变量筛选程序

数学教研室 苑锡光

摘 要

在定性分类变量的判别分析中常须对解释变量进行筛选。近年来发表过一些这种方法，但都是用于二分反应变量的，不能直接用于多类别（类别数＞2）反应变量的情况上。本文从减少错分率出发，分析了引入变量对减少错分率的效力，导出一个适用于多类别反应变量的筛选程序。程序中把引入变量的实际效力及似然比判据X_L^2两个统计量配合起来，同时做为选入变量的标准，且对后者的自由度的计算进行了相应的修正。这样就在变量筛选过程中，随时都能把待选入变量的效力正确地反映出来了。在选入过程的终止标准问题上，没有采用通常的以一个概率值为终止标准的做法，而是采用了比容——样本容量与样本空间非零状态数之比——为标准，看来这样既能较好地保证予报的可靠性，又可在此基础上尽可能地减少错分率。最后在例算中与Lachin[3]的筛选程序进行了某些比较。

一、引入变量的效力

设一母体中的个体可按照有限个定性分类变量来分类，其中一个（记为X）是反应变量，其他n个（记为V_1、V_2、……、V_n）是解释变量，它们都有确定的边缘分布，且都是多项分布。这些变量都有有限的类别数，X有I类，其他分别有J、K、……类。I、J、K、……等都不小于2。

用p_i（i＝1，2，……，I）表示变量X第i类出现的概率，p_i中最大者设为p_u。要预报某一个体属于那一类，如只以X的分布为依据，则应将它们都预报为第u类，于是预报的错分率R_e为

$$R_e = \sum_{i} p_i \qquad (i \neq u)$$

现引入另一变量V_1，其第j类出现的概率为q_j（j＝1，2，……，J）。将个体按X及V_1交叉分类，得一I×J二维列联表。记个体落入第i行第j列的概率为p_{ij}。此时必有$p_{i+} = p_i$（i＝1，2，……，I），式中下标中＋号表示对相应的下标求和，即

$$p_{i+} = \sum_{j} p_{ij}$$

同样$p_{+j} = q_j$即V_1的边缘分布。对一定的j，诸p_{ij}/q_j（i＝1，2，……，I）可视为一分布函数，把它叫做第j列的子分布。

如已知某个体属于V_1中第j类，即可按第j列子分布中概率最大者，设为P_{wj}，预报它属于变量X的第W类。考虑这样做的错分率。

在X与V_1独立情况下，$p_{ij} = p_i q_j$，各子分布与X的边缘分布等同，错分率与未引入V_1时相同。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

当X与V_1有连带（association）时，设

$p_{ij}=\Delta_{ij}p_iq_j$（$i=1,2,\cdots\cdots,I$；$j=1,2,\cdots\cdots,J$）其中$\Delta_{ij}\geqslant 0$且至少有部份不为1，Δ_{ij}受式

$$\sum_i \Delta_{ij}p_i=1，及\quad \sum_j \Delta_{ij}q_j=1$$

的约束。如由于Δ_{ij}对1的偏离使得在第j列中有p_{wj}（$W\neq u$）成为p_{ij}中之最大者，即存在有

$$p_{wj}=\max\left\{p_{ij}\right\}\ (W\neq u;\ i=1,2,\cdots\cdots,I),\quad\cdots\cdots\cdots\cdots①$$

则将使落入V_1中第j类而属于X中第W类的个体不再像未引入变量V_1时那样被预报为第u类，而是被正确地预报为第W类。于是错分率必减少（$p_{wj}-p_{uj}$）。和

$$\sum_j\ (p_{wj}-p_{uj})\quad\cdots\cdots\cdots\cdots②$$

即为引入变量V_1后错分率的减少量，把它除以V_1的类别数得

$$E_1=\sum_j\ (p_{wj}-p_{uj})\ /\ J。\quad\cdots\cdots\cdots\cdots③$$

我们把E_1叫做引入变量的实际效力。如用R_e去除②式，即得Goodman与Kruskal从最优预报提出的对连带的测度[2]，亦即预测能力指标。

如①式对所有i、j都不成立，显然不可能减少错分率。这种情况下，由于X与V_1有连带必将有部份子分布与X的边缘分布有差别。这种分布上的差别，若在给定V_1的类别的条件下再引入与X有连带的其他变量，可能继续加大。因此这种差别产生了减少错分率的可能性，或是引入变量对减少错分率的潜在效力。

很自然地想到用量

$$S=2\sum_{i,j}p_{ij}\ln\frac{p_{ij}}{p_iq_j}=2\sum_{i,j}p_{ij}\ln\Delta_{ij}\quad\cdots\cdots\cdots\cdots④$$

（ln表示自然对数）来量度引入变量的这两种效力。因为它（ ）可做为各子分布与X的边缘分布间“相似性”的量度。我们把S叫做引入变量对减少错分率的效力。

也可以从另一途径来考虑。Δ_{ij}对1的偏离可用

$$\sum_{i,j}\ (\Delta_{ij}-1)^2$$

量度。而错分率的减少除依赖于Δ_{ij}对1的偏离外，还依赖于个体落入相应格的概率，所以上式中各项以相应的概率加权是合理的。因此也可用量

$$S'=\sum_{i,j}p_{ij}\ (\Delta_{ij}-1)\quad\cdots\cdots\cdots\cdots⑤$$

表示引入变量的效力。

考虑再引入一个变量，设为V_2。用r_k（$k=1,2,\cdots\cdots,K$）表示变量V_2的边缘分布。于是我们有一个$I\times J\times K$三维列联表。用p_{ijk}表示个体落入第i行第j层第k列格中的概率。假设给定V_1类别的条件下，X与V_2独立。这就是说，我们考察在上次$I\times J$表的子分布上发生的引入V_2的效力。在这种条件独立假设下，个体落入第j层第i行第k列的概率应为

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$p_{ij+} \cdot p_{+jk}/p_{+j+}$$

此时错分率与不引入V_2时相同。而在假设给定V_1类别条件下X与V_2有连带时，则可设

$$p_{ijk} = \Delta_{ijk} \cdot p_{ij+} \cdot p_{+jk}/p_{+j+}。$$

于是引入变量V_2的效力即可表为

$$\quad = 2\sum_{i,j,k} p_{ijk} \ln \Delta_{ijk}。 \quad \cdots\cdots⑥$$

引入变量V_2时错分率的减少是在上一引入变量效力基础上发生的，所以V_2的实际效力E_2应为

$$E_2 = (\sum_{j,k}(p_{wjk} - p_{ujk}) - JE_1)/k,$$

式中$p_{wjk} = \max\{p_{ijk}\}$。（u如前为X边缘分布中具有最大密度的类别）。

不难把以上叙述引申到更多次地引入变量的情况去。

二、比较用统计量

如抽取了容量为N的样本。考察X与第一次待选入的各个变量（有n个）之间的连带，列出n个二向表。以V_1为例，列出$I \times J$表。用f_{ij}及$\hat{f}_{ij}$分别表示第i行第j列格中个体频数的观测值及在X与V_1独立假设下的期望值。将p_i、q_j及f_{ij}的极大似然估计量

$$\hat{p}_i = \frac{f_{i+}}{N}, \qquad \hat{q}_j = \frac{f_{+j}}{N}, \quad \hat{f}_{ij} = \frac{f_{i+} \cdot f_{+j}}{N}$$

等代入①式，再乘以N，经简单代数变换即得

$$NS = 2\Sigma f_{ij} \ln (f_{ij}/\hat{f}_{ij}),$$

可见NS即似然比χ_L^2统计量，它在待选入变量与反应变量独立的假设下近似地服从自由度为$(I-1)(J-1)$的χ^2分布，因此似然比判据的χ_L^2统计量结合其自由度及相应概率值可用来比较各待选入变量对减少错分率的效力。

不难看出NS′就是Pearson 的x^2统计量。这里我们得到了Lachin导出的同样的结论[3]，不过是在反应变量不限于二分的情况下。以上两种统计量我们倾向于前者，因它与分析列联表的对数线性模型有更紧密的联系。

第$a+1$（$a \geq 1$）次选入变量时，如已选入的a个变量为V_{s1}，V_{s2}，……，V_{sa}，它们的类别分别用j、k、……、l表示，类别数为J、K、……、L。待选入变量有V_{t1}、……、$V_{t(1-a)}$，其类别用m、……、r表示，其类别数为M、……、R。设从V_{t1}开始计算。此时独立性假设为：给定V_{s1}、……、V_{sa}类别条件下，反应变量X与解释变量V_{t1}独立，由是期望值应按

$$\hat{f}_{ij\cdots\cdots lm} = f_{ij\cdots\cdots l+} \cdot f_{+j\cdots\cdots lm}/f_{+j\cdots\cdots l+}$$

计算。似然比χ^2统计量为

$$\chi_L^2 = NS = 2\Sigma f_{ij\cdots\cdots lm} \ln (f_{ij\cdots\cdots lm}/\hat{f}_{ij\cdots\cdots lm}), \qquad ⑦$$

其自由度为

$$df = JK\cdots\cdots L(I-1)(M-1)。 \qquad ⑧$$

引入变量的实际效力可用统计量

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$NEa_{+1}=\sum_{j\cdots lm}(f_{wj\cdots lm}-f_{uj\cdots lm})/N-\sum_{\alpha=1}^{a}L\alpha E\alpha]/M \quad \cdots\cdots⑨$$

去比较，式中

$$f_{wj\cdots lm}=\max_{i}\left\{f_{ij\cdots lm}\right\},$$

Lα表示第α个引入变量的类别数。

在实际使用χ_L^2时往往遇到有的观测值为零，甚至边缘总计中出现零的情况。观测值为零时我们采用将⑦式中相应的项设为零的方式处理。边缘总计中出现零时，宜将该行或列从表中简约掉。于是计算自由度的公式⑧即应加以修正，其中df 有如JK……L层二向表自由度的和。如有一层中出现m′列（或i′行）的边缘总计为零则从该层的M（或I）中减去m′（或i′）。这样，可以较正确地反映正待选入的解释变量的效力。

三、变量筛选程序

1．设有a个已选入变量（延用上段中的符号），考虑从其它n－a个变量中再选入一变量$V_{S(a+1)}$。先按照⑦、⑧式计算各未选入变量的χ_L^2值、df值及各相应概率P_{t1}，P_{t2}，…，$P_{t(n-a)}$。记这些概率中最小者为$\delta\times10^{-E}$（$1\leqslant\delta<10$，E为正整数或零）。

指定最大可接受概率水平α。如$\delta\times10^{-E}>\alpha$则筛选过程停止。

2．以$\delta\times10^{-E}$为准定出可接受概率区间。为此指定因数 >1。如P_{t1}…，$P_{t(n-a)}$中有P_m符合条件

$$\min\left\{\alpha,\ \delta\times10^{-E}\right\}>P_m>\delta\times10^{-E},$$

则将相应变量列入集合Y。

3．计算各未选入变量的实际效力NE_{t1}，NE_{t2}，…，$NE_{t(n-a)}$。找出集合Y中实际效力最大的变量，把它选入，即$V_{S(a+1)}$。

如Y中具有最大实际效力值的变量不止一个，则选入其中概率最小者。

然后再次进入选入过程。

4．随时计算样本容量与各样本空间非零状态数（状态数指已选入的各解释变量类别数之积）之比，以下将此比简称比容θ。（n－a个未选入变量分别与已选变量组成n－a个样本空间）。为比容θ指定可接受下限θ′及上限θ″（θ″＞θ′）。

当$V_{S(a+1)}$的比容小于θ″时选入过程停止，并按下列标准重新考虑最后一个被选入变量（$V_{S(a+1)}$作废）。

5．列出所有比容大于θ′的变量，将其中实际效力最大者选入作为$V_{S(a+1)}$。此次选入时，如没有一个比容大于θ′，则返回上次选入过程，依此重新考虑V_{Sa}。此时如有变量其实际效力与实际效力最大值相等（或相差甚小），则取其中比容最大的变量。

四、实例及结语

全国胃癌病理协作组为研究中、晚期胃癌病理改变与予后的关系，于1980年按统一的选择规定及分类标准收集了8523例病例。经该组同意，利用其部份病例进行了几次例算，以考查以上程序。计算是在BCM—Ⅱ型计算机上进行的。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

病例中被予报的反应变量是病人手术后存活年限（记为X）。解释变量除记载上过分不完全者外有22个，它们是：性别（V_1），年龄（V_2），部位（V_3），肿瘤大小（V_4），大体类型（V_5），组织学类型（V_6），两种癌周间质反应（V_7及V_8），五种肿瘤侵犯情况（V_9～V_{13}），四种癌旁上皮反应（V_{14}～V_{17}），三种淋巴结反应（V_{18}～V_{20}），淋巴结转移情况（V_{21}）及手术类型（V_{22}）。其中有20种属于定性分类者，有的有序，有的无序。解释变量的类别数有2，3，4，5，7，8六种。

（1）例一

将反应变量划为三类，第一类存活年限在一年内（$X \leqslant 1$），第二类为$1 < X \leqslant 3$，第三类为$X > 3$。所用参数值如表1所示，其中各先验概率是根据6400例有随访病例算出的存活率估计的。对这里的例算，适用病例有2759例。随机取1963例为试验样本，其他796例为验证样本。

表1　　存活年限分三类时所用参数

先验概率：	可按受概率水平$\alpha = 0.01$
$p_1\{X \leqslant 1\} = 0.35$	概率区间常数$\beta = 10$
$p_2\{1 < X \leqslant 3\} = 0.36$	比容下限$\theta' = 15$
$p_3\{X > 3\} = 0.29$	比容上限$\theta'' = 20$

表2所列为在试验样本上进行的筛选结果。各行的次序按选入先后排列，第一次选入的是V_{21}（淋巴结转移情况），以下是V_9（肿瘤侵犯深度）、V_7（淋巴样细胞反应）及V_{18}。选入V_{18}后比容已降至θ'与θ''之间，故选入过程终止。

由以上结果，估计后验概率并用简单Bayes决策法（设错分损失都为1）得出予报法则。将此法则在试验样本上回报及用之于验证样本所得的错分率，如表3所示。

表2　　筛选结果

选入变量	χ^2_L	df	$p(\chi^2_L)$	NE	比容0
V_{21}	208.234	4	～0	0.0329916	490.75
V_9	100.88	18	1.80E-13	2.27243E-03	163.6
V_7	124.346	16	1.0E-07	5.8103E-03	56.08
V_{18}	225.556	163	5.0E-04	7.5885E-03	16.5

表3　　错分率

样本	存活年限			总错分率
	$X \leqslant 1$	$1 < X \leqslant 3$	$X > 3$	（Bayes估计量）
试验样本	0.540	0.467	0.450	0.488
验证样本	0.609	0.563	0.502	0.562

（2）例二

将反应变量划为两类，第一类存活年限在三年内（$X \leqslant 3$），第二类为$X > 3$年。所用参数值见表4。试验样本容量$N = 2274$例，验证样本容量=1054例。

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表4　　存活年限分为两类所用参数

先验概率	概率区间常数β = 10
$p_1\{X \leqslant 3\} = 0.71$	比容下限0′ = 10
$p_2\{X > 3\} = 0.29$	比容上限0″ = 15
可接受概率水平 = 0.01	

筛选结果：第一次选入V_{21}，第二次选入V_7，以下是V_8及V_{18}最后是V_1（性别）。前四个选入时都是该次概率值最小而实际效力也最大。第五个则只是落入集合Y中而实际效力最大者，但它的比容已小于0′。由于其它17个待选入变量的比容也都小于0′，故退至第四次选入过程。这时发现V_{10}（淋巴管有无瘤栓）的实际效力与V_{18}者相等，但其比容远大于V_{18}者，故将V_{10}选入。过程终止。

利用这四个解释变量在试验样本上估计出后验概率，得出予报法则。将此法则用于验证样本得到的错分率及其在试验样本上的回报列入表5中。

表5　　错　分　率

样本	存活年限 ≤3年	存活年限 >3年	总错分率（Bayes估计量）
试验样本（2274例）	0.0570	0.694	0.242
验证样本（1054例）	0.1053	0.709	0.280

我们又用Lachin的筛选程序处理了同一试验样本。用的最大可接受概率mpa为0.15，最小剔除概率也是0.15。使人感兴趣的是，所选出的变量同样是上面那四个，选出的次序也相同（未出现被剔除的情况）。只是第四个变量V_{10}是直接选入的。到第五次选入时所有变量的概率值都大于0.4，过程自然终止。

例算说明我们所用筛选程序可以选出对减少错分率起较大作用的变量集合。如表2所列在试验样本上总错分率为48.8%，由此可知予测能力指标达到23.75%，即四个变量的引入使错分率减少了23.75%。而且在试验及验证样本上的错分率，由表2及表3可见，都是基本一致的。说明得出的予报法则的可靠性是较好的。

由例二看来，在解释变量为二分时，似乎我们所用程序与Lachin的程序的效果相同。实际上并不总是这样。由于我们引用了引入变量的实际效力，并且对自由度的处理采用了不同的方法，在变量筛选过程中随时可以把待引入变量的实际及潜在效力清楚地表现出来。从而较多地提供了减少错分率的机会。再者最后一次选入变量不再考虑潜在效力，离开了概率值，以比容及实际效力为终止及变量选入的标准；这既保证了预报的可靠性、一致性，又可在必要的比容条件下得到尽量低的错分率。下面把以上实例引申一下以说明这里的看法。

把例一中两个样本加到一起做为试验样本，再用两种程序进行变量筛选。结果Lachin程序仍终止于原来的四个变量，而本文所用程序则选入了原被筛选掉的V_{18}及V_1，而未再选入V_{10}。两个结果对试验样本回报的错分率见表6。从各种错分率比较，总的来说本文程序的结果是较好的。特别是在样本中占比例较小的存活年限大于3年的那一类的错分率有明显下

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表6　　错　　分　　率

程　序	存活年限 $X \leqslant 3$	存活年限 $X > 3$	总错分率（Bayes估计量）	予测能力指标
Lachin	0.0718	0.699	0.2537	13.52%
本　文	0.0824	0.586	0.2285	21.20%

降，这正是在类别比例悬殊情况下所希望的。同样的方法用在例一的样本中，也可再选入另一变量V_4，从而使回报总错分率降至42.66%。

参 考 文 献

(1) Goldstein M and Dillon WR : "Discrete Discriminant Analysis", New york: John Wiley & sons, 1978

(2) Goodman LA and Kruskal WH: "Measures of association for cross-classificafions", J Amer Statist Ass, pp.732--764, 1954

(3) Lachin JM: "On a stepwise procedure for two populations Bayes decision rule using discrete variables". Biometrics, 29: 551—564, 1973

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net