新冠无症状比例问题——比较机器学习西瓜分类

鲁晨光

摘要：为澄清新冠无症状比例减少问题，本文拿机器学习中典型的西瓜分类问题做比较。通过图解医学检验和西瓜分类原理，本文得出结论：新冠检测的特异性不够高以及基础概率增大是假阳性和无症状比例由大变小的主要原因。而核酸检测特异性不够高和我们为提高敏感性（为了清零）而增大 CT 划分点有关。最后文中提出一些改进措施。

1. 引言

我曾在科学网发表博文解释说：阳性无症状比例下降是因为检测的特异性不够高，导致感染的基础概率由小变大时，假阳性比例快速下降。假阳性当然无症状。详见：科学网—为什么现在新冠无症状比例下降那么大？——从检测特异性看 - 鲁晨光的博文 (sciencenet.cn)

现在我发现：人工智能中的二元分类(比如西瓜熟和不熟，或好和不好分类)存在同样问题；阳性无症状比例还和 CT 值划分点有关。

本文目的是通过比较新冠检测和西瓜分类，澄清新冠检测存在的技术问题和数学原因，并提出改进检测或分类方法，

2. 新冠检测和西瓜分类原理

周志华的《机器学习》【1】是很多人学习机器学习的入门书。其中用西瓜分类为例说明机器学习方法，由浅入深，引人入胜。我们用西瓜分类做对比解释医学检验，说明医学检验中的困难也是机器学习中的困难。

我们根据西瓜外部特征判断西瓜成熟没有，特征比如瓜皮颜色、条纹清晰度、拍打声音，瓜蒂光滑和卷曲程度等。我们用 x_1 表示熟西瓜，x_0 表示生西瓜。假设特征是一维的，用 z(变量)表示，划分点是 z‘，两类的名字是 C_0 和 C_1，标签是 y_1 和 y_0。大写 X,Y,Z 表示相应的随机变量。医学检验或新冠检测类似(参看图 1)。

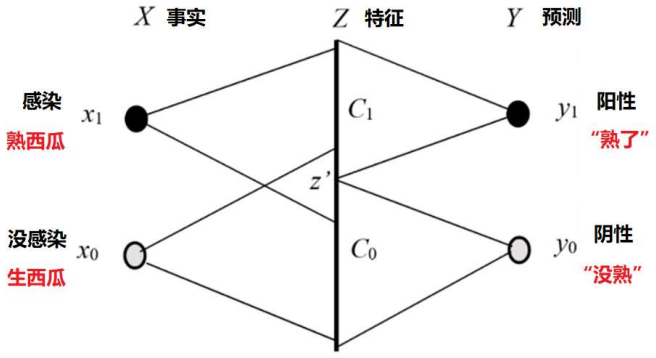

图 1. 医学检验和西瓜分类(根据特征 z)图解

假设西瓜成熟收获期是 6 月份，我们五月下旬按照同样标准(比如花纹清晰度或拍打声音)选瓜，选出来的生瓜比例一定很大。原因是，决定生瓜熟瓜的特征有

很多，它们并不是严格相关的。如果我们只用一两种， 就难免会按一定的比例误判。早期熟瓜比例少时，选出不熟的瓜比例就大。而到六月下旬，生瓜熟判的比例就大大下降，相反，熟瓜生判的比例会大大增加。

医学检测是类似的。对于新冠核酸检验，特征值就是 CT 值。核酸检测利用聚合酶链式反应，放大扩增特定的核酸片段，完整的反应步骤称为循环。CT 值是核酸扩增到检测阈值时所经历的循环数，CT 值越小表示病毒数量越多。(参看：http://www.sz.gov.cn/ztfw/ylws/wyw_183957/ywzsk_184570/content/mpost_10347545.html
)。

设真的感染者和未感染者的在 z(CT)上的概率分布是 $P(z|x_1)$和 $P(z|x_0)$。如图 2 所示。

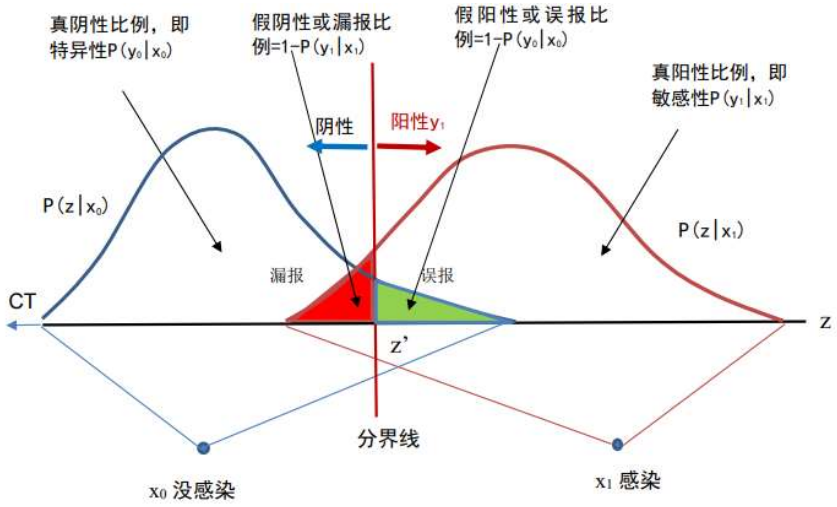

图 2. 医学检验分类图解(误报和漏报不可避免)。假设基础概率 P(x1)=0.5, 绿色面积和红线下面右侧部分面积之比就是假阳性和真阳性之比。

图 2 中其中两条曲线下面面积都是 1，红色曲线下面右边部分面积是敏感性，蓝色曲线下面左边部分面积是特异性。绿色三角形就反映误报的相对比例(即假阳性比例)，红色三角形面积反映漏报的相对比例。绝对比例是多少呢？那要看感染的基础概率 $P(x_1)$是多少。图 3 反映了基础概率很小时的情况。

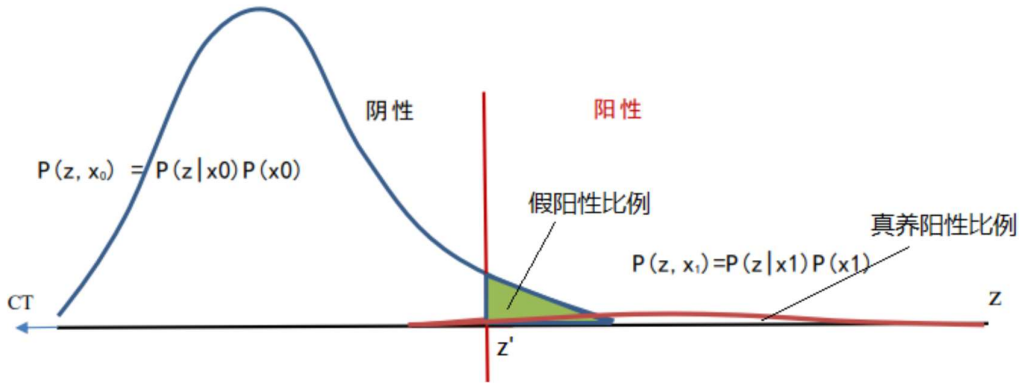

图 3. 基础概率 $P(x_1)$很小时假阳性比例很大。这时两条曲线是联合概率分布，下面面积反映感染和未感然人数相对比例。绿色面积和红线右边下面面积之比就是假阳性和真阳性比例。

图 3 反映了中国 2022 年 12 月之前的情况，那时候大部分地区感染的基础概率 $P(x_1)$很小，所以按照同样的 z'=35(循环次数)标准，假阳性比例很高。2022 年年底和 2023 年年初，感染的基础概率增大到 50%，假阳性的比例就如图 2 所示。这也就是为什么无症状比例下降了。其实主要原因是假阳性的比例下降了。假阳性当然无症状。

境外为什么没有报告如此高比例的无症状？因为境外基础概率大，并且没有用核酸普查。另一个原因是境外普遍用较低的 CT 分界点 z'=30.

3. CT 分界点(z')和无症状比例的关系

新冠检测有核酸检测和抗原检测。核酸检测敏感性不高(通常只有 0.6 左右)特异性高(据说接近 1)。抗原检测特异性和敏感性都比较高，但是也不太高(大概在 0.9-0.98 之间)。但是据说抗原在感染早期发现不了病毒——相当于早期敏感性很低。敏感性高就相当于不放过坏人，特异性高就相当于不冤枉好人。

为什么中国检测主要用核酸？因为当初感染比例小，用抗原冤枉好人太多，在清零政策下，导致放舱放不下。但是清零又要求不放过坏人，所以采用以下一种或几种措施：

1) 重复多次核酸检测(一次漏报 0.4，两次漏报就只有 0.16);

2) 提高 CT 值划分点(西方使用 z'=30，中国开始使用 z'=40，后来改为 z'=35，等于图 2 图 3 中分界红线左移);

3) 用抗原辅助(比如：有症状但是核酸阴性就采用抗原捡漏，或者先用抗原从有症状人群筛选感染者，再用核酸核实)。

从图 2 和图 3 可以看出，假阳性比例大主要因为检测的特异性不高——导致绿色区域较大。专家都知道抗原特异性不够高，但是都认为核酸特异性很高，接近 1. 按说核酸特异性高，假阳性比例应该很小。但是，上海曾报道核酸检测出现大量假阳性(见 http://www.gov.cn/fuwu/2022-05/23/content_5691963.htm)，官方解释说是样本遭到污染。但是我以为核酸的特异性并没有大到 1，即使是 0.999，上海几千万人口的 1/1000(被判为假阳性)也是数万人。**特别是，当我们把 CT 值划分点定 z'从 0.3 改为 0.35 或 0.4 时，增大敏感性的同时也降低了特异性**。左移图 2 中垂直红线，绿色区域增大，增大部分就是特异性降低部分。

由此可见，当初无症状比例大是因为假阳性比例大，而假阳性比例大是因为 CT 划分点 z'较大， 特异性达不到 1. 而使用较大的 z'值也是当时清零的需要。

4. 二分类划界方法和准则——从西瓜分类看

机器学习方法是先用样本得到学习函数，然后根据某种准则确定分类边界 z' 和分类函数 $y=f(z)$.

学习函数通常有：

1）似然函数

一个样本包含很多样例，一个样例是(x_i, z_k). 假设样本很大，我们就有不同样例(z, x)发生的相对频率，即样本分布 $P(x, z)$。从样本分布 $P(x, z)$, 我们得到条件

概率 $P(z|x_1)$和 $P(z|x_0)$. 但是它们还不是学习函数，因为它们不平滑，甚至不连续。为此，我们定义逼近它们的平滑的似然函数 $P(z|\theta_{x1})$和 $P(z|\theta_{x0})$, 其中 θ 表示模型参数。改变模型参数最大化对数似然度或负的交叉熵，就得到优化的模型参数和学习函数。比如：

$$P(z\,|\,\theta_{x1}) = \min_{P(z|\theta_{x1})} \sum_{z} P(z\,|\,x_1)\log P(z\,|\,\theta_{x1}).$$

假设似然函数是正态分布，则参数是期望和标准偏差。

2）分类函数(我叫它参数化的转移概率函数)

$P(x_i|z)$是转移概率函数(其中 x_i 是常量， z 是变量)。分类函数 $P(\theta_{x1}|z)$就是参数化的 $P(x_i|z)$。

这时候我们用样本分布 $P(x|z)$训练 $P(\theta_{x1}|z)$. 比如

$$P(\theta_{x1}\,|\,z) = \min_{P(\theta_{x1}|z)} \sum_{z}\sum_{i} P(z)P(x_i\,|\,z)\log P(\theta_{x1}\,|\,z).$$

常用的二分类函数是 Logistic 函数，

$$P(\theta_{x1}\,|\,z) = \frac{1}{1+\exp[-b(z-a)]},$$

$$P(\theta_{x0}\,|\,z) = 1 - P(\theta_{x1}\,|\,z) = \frac{\exp[-b(z-a)]}{1+\exp[-b(z-a)]}.$$

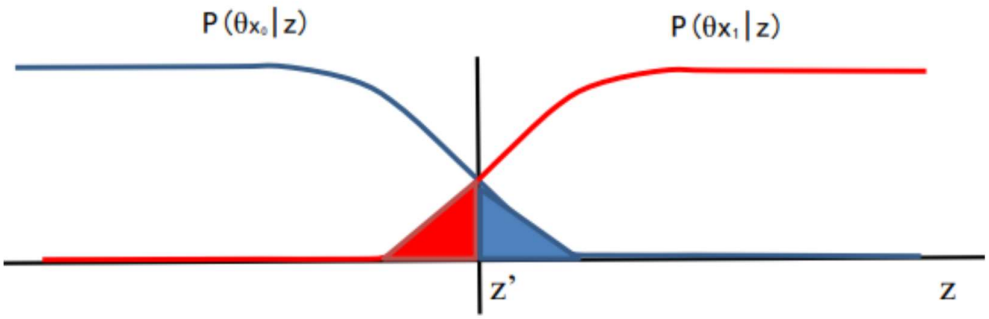

图 4. Logistic 函数用于二分类

3）真值函数

真值函数(即模糊真值函数或隶属函数)作为学习函数是笔者提出的【2】。真值函数 $T(\theta_{xi}|z)$和 $P(\theta_{xi}|z)$比，前者是对 $P(x_i, z)/[P(x_i)P(z)]$的逼近再归一化(使最大值为 1)，后者是对 $P(x_i|z)$的逼近。前者和 $P(x_i)$无关，而后者有关。真值函数更加适合多分类。

分类准则通常有：

1） **最大后验概率准则**(即最大正确率准则或最小误差准则)

按照这一准则，要比较两条分类函数曲线，选择图 4 中两条曲线的交叉点下面的 z 作为分界点 z'即可。这一准则的优点是比较直观——反映正确率。但是，如果 $P(x_1)$变了，以前得到的 Logistic 函数就不适应了，就不能反映正确率了。

2)**最大似然准则**

按照这一准则，选择图 2 两条曲线交点下方的 z 作为 z'就行了。最大似然准则就是最大信息准则(给定 z 时用最大信息准则，比较 $I(x_1; z)$和 $I(x_0; z)$，选信息量

大的 x【3】.

最大似然准则和最大正确率准则的异同是：当 $P(x_1)= P(x_0)$时，两者等价。当 $P(x_1)<< P(x_0)$时，最大似然准则能减少小概率事件的漏报。比如地震预报，而用最大后验概率或正确率准则，我们永远报“明天无地震”好了。但是这样的预报不含有信息。这也就是《机器学习》中讲到的类别不平衡问题。但是使用最大似然准则的缺点是正确率不高，因为它重视相对正确率。

3)**最大互信息准则**(或最大似然比准则)

最大互信息分类不用 x 和 z 之间的信息，而用 x 和分类标签 y 之间的信息优化 z’。

笔者贡献就是：找到一种迭代方法，它能求出最大互信息划分点 z’*【2】，而且还通过信息率-逼真函数 $R(G)$(信息率失真函数的推广)证明了迭代收敛；并且证明了最大互信息分类等价于最大似然比分类【3】。

这种迭代方法是：

- 先假设一个划分点 z’(比如有最大似然准则), 于是得到分类函数 $y=f(z)$。然后得到一点 z 对应的语义信息 $I(x_i; \theta_{yj})$和平均语义信息：

$$I(X_i;\theta_{y1}|z)=\sum_i P(x_i\mid z)I(x_i;\theta_{y1}),$$

$$I(X_i;\theta_{y0}|z)=\sum_i P(x_i\mid z)I(x_i;\theta_{y0}).$$

它们显示为 z 上的两条分布曲线。

- 用上面两条曲线交点对应的 z 做为新的划分点 z’.

重复上面两个步骤就得到 z‘*.

这种分类的优点是，提供的香农互信息最大; $P(x_1)$变化时，正确率下降较小(和最大后验概率准则比)。缺点是计算麻烦些。

4)**最小损失准则(贝叶斯分类器)**

我们定义预测成功和失败四种情况下有四种不同损失(负数为增益)。分界 z’使得平均损失最小。比如医学检验时，假设一个损失函数如表 1 所示：

表 1. 医学检验的损失函数

	y_0	y_1
x_0	0 (真阴性)	1 (假阳性)
x_1	3 (假阴性)	-1 (真阳性)

其中假设假阴性损失最大, 因为漏报损失最大。

一般情况下，小概率事件漏报损失大。最大互信息准则也减少小概率事件漏报，所以两者通常一致。当然也有例外的情况，比如大概事件漏报损失大时。

4） 正则化最小误差准则

这一准则和最大似然准则以及最大互信息准则类似。最大互信息或语义信息准则可以看做是一个特殊的正则化最小误差准则【2】。

5. 医学检验的改进方法——仿照西瓜分类

1）**根据任务特殊性或损失优化分类边界 z’**。

西瓜分类，卖家和买家需求不同，分类标准也不同。比如卖家希望减少熟瓜

生判，而买家希望减少生瓜熟判。

对于医学检验，目的不同，划分边界也应有改变。需要减少漏报(即需要不放过坏人)时，敏感性要高；希望减少误报(即希望不冤枉好人)或怕医疗资源挤兑时，特异性要高。

另外，有时候正确率重要，有时候信息重要，有时候减少损失重要。选择 z’要综合考虑。

2）**根据基础概率 $P(x)$变化调整 z’**.

西瓜成熟早期，熟瓜比例少，成熟标准要严格一些，不然生瓜熟判的比例大。而成熟晚期，成熟的标准要松一些，不然熟瓜生判的比例大。

医学检验类似。应根据基础概率调整 z’. 基础概率大可能因为被感染者比例大，也可能是因为我们选择了高危或有症状人群来检测。要想控制误报或漏报比例，划分边界也应调整(除了使用最大似然准则)。比如，基础概率小时，要想减少假阳性比例，就要降低 CT 的划分点 z’。基础概率大时，要减少假阴性比例，就要增大 z’.

3）**使用多种检测手段**。

同时根据西瓜外表、拍打声音和瓜蒂形状就能更可靠筛选出好瓜。因为每种特征都可能有不足之处，用每种特征淘汰一批，剩下的好瓜比例就比较大。

对于新冠检测，不妨用抗原减少漏报(因为它敏感性较高)，再用核酸减少误报(因为它特异性较高)，两种检测结果应该比两次重复用核酸好。当然这是假设抗原不是用在感染早期，敏感性确实高。

4）**如实报告检测可靠性，让大家有合理预期**。

当我们提供阳性和阴性判断时，最好同时注明基础概率和假阳性或假阴性比例，使大家对检测结果的误判有所预期。比如，当初核酸检测用 CT 的划分点 z’=35(或者用抗原)，如大海捞针发现阳性时，假阳性和无症状比例必然很高。但是这并不意味着病毒广泛传染后或基础概率增大后，假阳性和无症状比例还是这么高。如果大家对此有所预期，就不会错误地轻视病毒，以至于大面积发烧时不知所措。

6. 结束语

新冠传染早期无症状比例大，扩散期无症状比例小，这和西瓜分类类似——按照同样标准选瓜，早期不熟的比例大，中晚期不熟比例小。这是由于 1)分类技术有限，减少漏报(假阴性)必然增大误报(假阳性)；2)基础概率增大导致假阳性比例减小。

当初无症状或假阳性比例大，并不是有人造假。检测技术有限和清零需求是主要原因，我们为了减少漏报，提高了 CT 划分点 z’，就必然减少检测的特异性，从而带来更多假阳性和无症状。承认检测结果可靠性有限，让大家有合理预期，将有助于应对未来疫情。

参考文献

【1】 周志华，《机器学习》，清华大学出版社，2017.

【2】 鲁晨光，Semantic Information G Theory and Logical Bayesian Inference for Machine Learning，*Information* **2019**, *10*(8), 261; **https://doi.org/10.3390/info10080261**；中文：《语义信息 G 理论和逻辑贝叶斯推理用于机器学习》http://www.survivor99.com/lcg/cm/gtheory/index.html

【3】 鲁晨光，Semantic Channel and Shannon Channel Mutually Match and Iterate for Tests and Estimations with Maximum Mutual Information and Maximum Likelihood，2018 IEEE International Conference on Big Data and Smart Computing (BigComp)，15-17 Jan. 2018，https://ieeexplore.ieee.org/document/8367121

作者更多文章见：http://survivor99.com