小议RPD

江 苏

（Email: jiangsukust@163.com）

时至今日，笔者对十几年前被用户问及模型的RPD是多少却未能马上解答一事仍记忆犹新。因为，在平常的工作中大家主要关注的是内部交叉验证的$RMSECV$和R^2及外部验证的$RMSEP$、R^2、MAE（平均绝对误差）和MRE（平均相对误差）等统计指标，未曾计算过RPD，尽管长期的实践证明这样做并无不妥，但当年的尴尬却一直如鲠在喉。

RPD的概念最早是由 Phil Williams（P.C. Williams）于 1986 年提出的，目的是为了引入一种快速评估近红外模型的简单方法[1]。1993 年，Williams 与 Sobering（二者当时均任职于加拿大谷物委员会下属的谷物研究实验室）合著论文[2]发表在 Journal of Near Infrared Spectroscopy 上，这是提出者第一次将RPD带到同行评审刊，将其用于谷物和种子物化参数的近红外模型评估。

注：值得一提的是，Williams 与已故近红外之父 Karl Norris 有过诸多合作，Williams 也凭借其在近红外光谱分析领域的卓越贡献获得了多个奖项，其中就包括 2015 年 ICNIRS Karl Norris Award（2014 年首届奖章颁给了 Karl Norris 本人）。

原文中对RPD的定义为：“the RPD, which is the ratio of the SEP to the standard deviation (SD) of the original data.”；同时给出了具体的示例：“For example, if the standard deviation of the original data is 1.83 and the SEP is 0.27, the RPD is given by 1.83/0.27 = 6.78.”；因此，其公式如下：

$$RPD = \frac{SD}{SEP} \tag{1}$$

式中：

SD — n个样本参考值的标准偏差，其值为$\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$

y_i — 第i个样本的参考值（$i = 1, 2 \ \ldots \ n$）

$\bar{y}$ — n个样本参考值的平均值

SEP — Standard Error of Prediction，预测标准误差

看到这里，不知道读者们有没有发现一个问题，RPD的文字定义与公式之间是矛盾的。我们也可以在 IM Publications 下的近红外论坛上看到同样的疑问，“We probably need to ask Phil why it is RPD and not RDP!”，见 http://www.impublications.com/discus/messages/5/2136.html?1190896505=。

当然，我们现在无法揣测 Williams 当初如此定义的用意。有意思的是，Williams 在 2017 年

的一篇论文中描述RPD时使用的是“Ratio of SD to SE in validation or CV”[3]。不管如何，自 1993 年之后，RPD开始在近红外分析领域流行起来了，数以千计的期刊论文、教科书、会议论文、海报中都能看到它的身影。

事实上，在多年的使用过程中，RPD这个名词有过多种释义，例如：ratio of standard deviation of reference results and standard error of performance of NIR data[4]、ratio of the standard error of performance to the standard deviation[5]、ratio of performance to deviation[6]、residual predictive deviation[7; 8]……。而且，RPD从其首次提出时使用测试集表征预测性能（分母为SEP），也发展到了用于校正阶段（分母为SEC）和内部交叉验证阶段（分母为$SECV$）。

下面，我们来聊聊与RPD相关的两个问题。

1. RPD与R^2的关系

首先，我们有必要讨论一下SEP（Standard Error of Performance/Prediction，性能/预测标准误差），不同文献对SEP的定义不尽相同，分别记为SEP_1[9]、SEP_2[10]和SEP_3[11]。假定一组测试集（验证集）包含n个样本，则公式分别为：

$$SEP_1 = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}} = \sqrt{\frac{\sum(d_i)^2}{n}} \tag{1}$$

（1）式也称为$RMSEP$（预测均方根误差），其中：

$\hat{y}_i$ — 第i个样本的预测值（$i = 1, 2 \ \ldots \ n$）

y_i — 第i个样本的参考值（$i = 1, 2 \ \ldots \ n$）

d_i — 预测误差，其值为$\hat{y}_i - y_i$

$$SEP_2 = \sqrt{\frac{\sum(d_i - bias)^2}{n - 1}} \tag{2}$$

（2）式中：

$bias$ — 预测误差的平均值，又可称为系统误差，其值为$\frac{\sum d_i}{n}$。

$$SEP_3 = \sqrt{\frac{\sum(d_i)^2}{n - 1}} \tag{3}$$

（3）式与（1）式的区别仅在于根号下的分母不同；当$bias$趋近于 0 时，（3）式趋近于（2）式。

注 1：$\sum d_i^2$即SSE（Sum of Squares Error）或$PRESS$（Predictive Residual Error Sum of Squares）

注 2：$nSEP_1{}^2 = (n-1)SEP_2{}^2 + nBias^2$（推导过程略）

注 3：SEP_2（扣除了$bias$）的值并非肯定小于SEP_1（$RMSEP$），当$bias$较小且n不大时，SEP_2甚至会大于SEP_1。

注 4：笔者个人倾向于将SEP和$RMSEP$的定义区分开来，即将SEP定义为预测误差的标准偏差，如此便应采用SEP_2的表达，根号下分母为$n-1$（样本标准偏差）；而$RMSEP$的定义似乎从未产生过争议（分母为n），亦即$\sqrt{MSE}$（均方误差开平方）。MSE在机器学习和深度学习领域应用更广泛，是一种常见的损失函数，通常将其称作 L2-Loss，例如：mean_squared_error（scikit-learn、Keras）、MSELoss（PyTorch、PaddlePaddle）、L2Loss（mxnet）……

接下来，我们再回顾一下决定系数R^2的定义，此处采用R_3^2对应的公式，即

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$$

结合SD的公式，可得：

$$R^2 = 1 - \frac{\sum d_i^2}{(n-1)(SD)^2}$$

那么有：

$$1 - R^2 = \frac{\sum d_i^2}{(n-1)(SD)^2} = \frac{n(SEP_1)^2}{(n-1)(SD)^2} = \frac{(n-1)(SEP_3)^2}{(n-1)(SD)^2}$$

因为有$RPD = \frac{SD}{SEP}$，不难看出：

1) 如果采用SEP_3，那么$RPD^2 = \frac{1}{1-R^2}$

2) 如果采用SEP_2且$bias$趋近于 0，有$RPD^2 \approx \frac{1}{1-R^2}$

3) 如果采用SEP_1且n较大，即可忽略n和$n-1$之间的差异，也有$RPD^2 \approx \frac{1}{1-R^2}$

因此，不管采用何种SEP定义，也无论$bias$和n的大小，RPD与R^2之间都存在正相关关系，即RPD随着R^2的增大而增大，随着R^2的减小而减小。

注：因为$\boldsymbol{R_3^2} \in (-\infty, \mathbf{1}]$，当模型性能较差时，$R_3^2$可能出现负值，此时$RPD < 1$（$SEP > SD$）。

也有学者认为RPD与R^2本质上是相同的评价指标，不应该同时引用[12]。

2. 采用RPD给模型分级？

在 1993 年的论文中，Williams 指出RPD的值应越高越好（这一点，我想大家应该都会同意），

同时给出了一些指导意见：

- $RPD > 10$，excellent for use in process control
- $5 < RPD < 10$，adequate for quality control
- $2.5 < RPD < 5$，satisfactory for screening
- $RPD < 1$，not capable of predicting the parameter accurately

不少文献将上述意见视为圭臬，当作模型评价的黄金标准。当然，也有其他领域的学者提出了各自的推荐值[6; 13-17]。2014 年，Williams 在 NIR news 上发表了《The RPD statistic: a tutorial note》一文[18]，可以看作是前文[2]的补充和总结性发言。

然而，争议总是相伴而行。有学者就此提出了不同意见，他们指出：尽管RPD值大幅低于某个阈值，但模型仍是可用的，应该由用户来评估多个统计量，从而决定是否应用该模型[19]；没有统计或实用基础来支撑这样的阈值[12]。同在 2014 年的 NIR news，三位北欧知名学者 Kim H. Esbensen、Paul Geladi 和 Anders Larsen 联合发表了《The RPD myth...》[20]，罕见地针对《The RPD statistic: a tutorial note》进行了逐段评述，并以“It is a myth that the RPD statistics furthers an objective, across-model, comparative, unambiguous prediction validation figure-of-merit. **MYTH BUSTED!**”结尾。

让我们用具体的案例来讨论这个问题。首先，我们引入拙文《小议决定系数R^2》中 “R^2的欺骗性”章节的数据（详见 https://mp.weixin.qq.com/s/s2CeV62OjmQl3-dImxvuXQ），同时计算RPD值（采用SEP_2），结果见下表（蓝色字体为新增计算数据）：

表 1 不同含量范围总芳烃预测结果统计表

	Range-1	**Range-2**	**Range-3**	**Range-Full**	**Range-1-Modified**
$\boldsymbol{SSE}$	15.04196	25.03622	29.53437	55.98760	16.96802
$\boldsymbol{SST}$	29.44667	192.73745	484.93387	4317.27695	488.87712
$\boldsymbol{k}$ or $\boldsymbol{n}$	30	55	75	118	32
$\boldsymbol{R^2}$	0.48918	0.87010	0.93910	0.98703	0.96529
$\boldsymbol{RMSEP(SEP_1)}$	0.70810	0.67469	0.62753	0.68882	0.72818
$\boldsymbol{MAE}$	0.56543	0.53096	0.49706	0.54555	0.59045
$\boldsymbol{MRE}$	1.74498%	1.63659%	1.54526%	1.88746%	1.95375%
$\boldsymbol{SD}$	1.00767	1.88924	2.55991	6.07452	3.97118
$\boldsymbol{SEP_2}$	0.71714	0.68006	0.62846	0.69081	0.72947

$\boldsymbol{RPD}$	1.40512	2.77803	4.07331	8.79337	5.44393

通过RPD的定义及前文推导出来的RPD与R^2之间存在正相关关系，可以想见，RPD也具备R^2统计量的一些特点——**RPD的大小严重依赖于样本集参考值的分布情况**。从上表可知，仅仅只是在 Range-1 测试集中增加了两个样本，RPD瞬时就从 1.40512 变成了 5.44393，其根本原因在于当SEP的值非常接近的情况下（分别是 0.71714 和 0.72947），SD从 1.00767 飙升到了 3.97118。

如果采用SEP_3和SD来计算RPD，其结果见表 2：

表 2 不同方式计算 RPD 的结果

	Range-1	Range-2	Range-3	Range-Full	Range-1-Modified
$\boldsymbol{SEP_3}$	0.72020	0.68091	0.63175	0.69176	0.73983
$\boldsymbol{RPD}$	1.39916	2.77459	4.05207	8.78130	5.36765
$\boldsymbol{RPD^2}$	1.95764	7.69834	16.41931	77.11131	28.81169
$\boldsymbol{1/(1-R^2)}$	1.95764	7.69834	16.41931	77.11131	28.81169

显然，表 2 的结果证实了前文“如果采用SEP_3，那么$RPD^2 = 1/(1-R^2)$”这一结论。

综上，笔者认为，任何一个近红外模型都有其适用范围，如果抛开具体的应用场景，仅使用一个数值来判断模型是否可用，未免过于“武断”。这个案例再次提醒我们，进行模型评价时应该综合考虑多个统计量的结果，此外，预测值 vs.参考值的散点图尤为重要——一图胜千言。

…… 分割线 ……

本节所涉及的数据及 Python 代码已上传至笔者在 GitHub 社区维护的光谱多元校正算法库 https://github.com/freesiemens/SpectralMultivariateCalibration，SpecMC 软件的核心算法也在库中，欢迎交流。

…… 分割线 ……

由于本人水平和经验有限，文中难免有错漏之处，其中观点也仅代表个人意见，敬请读者批评指正。

参考文献

[1] Williams P C: Variables affecting near-infrared reflectance spectroscopic analysis, Williams P C, Norris K H,

editor, Near-infrared technology in the agricultural and food industries, 1987: 143-167.
[2] Williams P C, Sobering D C. Comparison of Commercial near Infrared Transmittance and Reflectance Instruments for Analysis of Whole Grains and Seeds[J]. Journal of Near Infrared Spectroscopy, 1993, 1(1): 25-32.
[3] Williams P, Dardenne P, Flinn P. Tutorial: Items to be included in a report on a near infrared spectroscopy project[J], 2017.
[4] Pandord J A, Williams P C, Deman J M. Analysis of oilseeds for protein, oil, fiber and moisture by near-infrared reflectance spectroscopy[J]. Journal of the American Oil Chemists' Society, 1988, 65(10): 1627-1634.
[5] Batten G D. Plant analysis using near infrared reflectance spectroscopy: the potential and the limitations[J]. Australian Journal of Experimental Agriculture, 1998, 38(7): 697-706.
[6] Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, et al. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy[J]. TrAC Trends in Analytical Chemistry, 2010, 29(9): 1073-1081.
[7] Soriano-Disla J M, Gómez I, Guerrero C, et al. The potential of NIR spectroscopy to predict stability parameters in sewage sludge and derived compost[J]. Geoderma, 2010, 158(1): 93-100.
[8] Agussabti, Rahmaddiansyah, Satriyo P, et al. Data analysis on near infrared spectroscopy as a part of technology adoption for cocoa farmer in Aceh Province, Indonesia[J]. Data in Brief, 2020, 29: 105251.
[9] Ozaki Y, Mcclure W F, Christy A. Near-Infrared Spectroscopy in Food Science and Technology[M]. II. John Wiley & Sons, Inc., 2006: 1-408.
[10] Naes T, Isaksson T, Fearn T, et al. A user friendly guide to multivariate calibration and classification[M]. NIR publications, 2002.
[11] Windham W R, Mertens D R, Barton F E: Protocol for NIRS calibration: sample selection and equation development and validation, Marten G C, Shenk J S, Barton F E, editor, Near infrared reflectance spectroscopy (NIRS): Analysis of forage quality: Agriculture Research Service, 1989: 96-103.
[12] Mcbratney A, Minasny B. Why you don't need to use RPD[J]. Pedometron, 2013, 33.
[13] Chang C-W, Laird D A, Mausbach M J, et al. Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties[J]. Soil Science Society of America Journal, 2001, 65(2): 480-490.
[14] Kovalenko I V, Rippke G R, Hurburgh C R. Determination of Amino Acid Composition of Soybeans (Glycine max) by Near-Infrared Spectroscopy[J]. Journal of Agricultural and Food Chemistry, 2006, 54(10): 3485-3491.
[15] Kapper C, Klont R E, Verdonk J M a J, et al. Prediction of pork quality with near infrared spectroscopy (NIRS): 1. Feasibility and robustness of NIRS measurements at laboratory scale[J]. Meat Science, 2012, 91(3): 294-299.
[16] Ge Y, Atefi A, Zhang H, et al. High-throughput analysis of leaf physiological and chemical traits with VIS–NIR–SWIR spectroscopy: a case study with a maize diversity panel[J]. Plant Methods, 2019, 15(1): 66.
[17] Liu J, Han J, Xie J, et al. Assessing heavy metal concentrations in earth-cumulic-orthic-anthrosols soils using Vis-NIR spectroscopy transform coupled with chemometrics[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 226: 117639.
[18] Williams P. The RPD Statistic: A Tutorial Note[J]. NIR news, 2014, 25(1): 22-26.
[19] Reeves J B, Smith D B. The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America[J]. Applied Geochemistry, 2009, 24(8): 1472-1481.
[20] Esbensen K, Geladi P, Larsen A. The RPD myth…[J]. NIR news, 2014, 25: 24.