PLOS ONE

研究论文

在评估不平衡数据集上的二分类器时，精确率 - 召回率曲线比 ROC 曲线更具信息量。

Takaya Saito*, Marc Rehmsmeier*

挪威卑尔根大学信息学系计算生物学研究室，邮编 N-5020，卑尔根，挪威，邮局信箱 7803 号

* takaya.saito@ii.uib.no（TS）； marc.rehmsmeier@ii.uib.no（MR）

摘要

二分类器通常通过诸如灵敏度和特异性之类的性能指标进行评估，其性能也经常通过受试者工作特征（ROC）曲线来展示。诸如阳性预测值（PPV）之类的替代指标以及相关的精确率/召回率（PRC）曲线则使用得较少。许多生物信息学研究开发并评估的分类器将应用于正负样本极不均衡的数据集，其中负样本的数量远远超过正样本的数量。虽然 ROC 曲线在视觉上具有吸引力，并能提供分类器在广泛特异性范围内的性能概览，但人们可能会质疑，在不均衡分类场景中使用 ROC 曲线是否具有误导性。本文表明，在不均衡数据集的背景下，由于对特异性的直观但错误的解读，ROC 曲线的视觉可解释性可能会对关于分类性能可靠性的结论产生误导。另一方面，PRC 图能够为观察者提供对未来分类性能的准确预测，原因在于其评估的是正预测中真正例的比例。我们的研究结果对大量在不平衡数据集上使用 ROC 图的研究的解读具有潜在影响。

参考文献：斋藤 T，雷姆斯迈尔 M（2015）在评估不平衡数据集上的二分类器时，精确率-召回率图比 ROC 图更具信息量。《公共科学图书馆·综合》10（3）：e0118432。doi: 10.1371/journal.pone.0118432

学术编辑：盖伊·布罗克，美国路易斯维尔大学

收件日期：2014 年 6 月 23 日

接受日期：2015 年 1 月 16 日

发布日期：2015 年 3 月 4 日

数据获取声明：数据可从 http://dx.doi.org/10.6084/m9.figshare.1245061 获取。

资金来源：作者无资金来源或支持可报告。

利益冲突：作者已声明不存在利益冲突。

介绍

二分类器是将数据集分为正类和负类的统计和计算模型。近年来，它们已成功应用于广泛的生物学和医学问题[1-3]。为了能够判断分类器的有用性，尤其是在与其他方法进行比较时，对其预测性能的评估至关重要。在模型构建阶段，常用的分类器性能度量指标有准确率、错误率和受试者工作特征曲线（ROC）下的面积（AUC）[4]。在评估最终模型时，还有许多其他有用的度量指标，而诸如 ROC 和精确率-召回率（PRC）曲线等几种图表则提供了直观的表示[5]。

类别不平衡——即正例和负例数量的差异，通常负例多于正例——在众多科学领域中都存在，包括生命科学，其中不均衡的类别分布是自然产生的[6-9]。不平衡数据集的分类是机器学习领域中一个相对较新的挑战[5, 10]。尽管针对不平衡数据的二分类问题已提出了许多解决方案[5, 11]，但这些方案大多与数据重采样[7, 12-14]或模型训练[15-19]有关。尽管在利用不平衡数据构建分类器方面已开发出最先进的解决方案[5, 11, 20]，但选择合适的性能评估方法往往被低估。

重要的是要认识到，训练阶段的评估与最终模型的评估是不同的。第一阶段是在训练过程中选择最有效和最稳健的模型。通常会将训练数据集进一步划分为训练子集和验证子集，例如用于交叉验证[21]。第二阶段是在训练完成后对最终模型进行评估。理想情况下，此阶段的测试数据应反映原始总体的类别分布，尽管这些分布通常是未知的。本文专门分析最终模型的性能评估。

高通量生物实验的迅速发展产生了大量规模庞大的数据集，而其中大多数数据集预计会是不平衡的[8, 22, 23]。在此，我们回顾了常用的评估指标的理论背景，特别是 ROC[24, 25]、PRC[26]、集中 ROC（CROC）[27]和成本曲线（CC）[28]。ROC 是二分类器最常用的评估方法，但在处理不平衡数据集时，对 ROC 曲线的解读需要特别谨慎[29]。ROC 的替代指标 PRC、CROC 和 CC 虽不如 ROC 流行，但它们在不平衡数据集下表现稳健[26-28]。在本研究中，我们旨在从多个不同角度阐明这些指标之间的差异，面向计算生物学/生命科学领域的受众。为实现这一目标，我们首先介绍诸如特异性和灵敏度等基本的单阈值指标，然后介绍 ROC 及其替代指标的图表。接下来，我们讨论在不平衡数据下具有信息量的精度指标，以及基于精度的 PRC。在一项模拟研究中，我们分析了在不平衡数据集的背景下应用 ROC、PRC、CROC 和 CC 的行为和效用。模拟使用了具有不同性能水平的随机生成样本。随后，我们展示了文献分析的结果，该分析调查了在关于不平衡数据集的真实世界研究中使用了哪些评估指标。文献分析基于两组 PubMed 搜索结果。此外，我们重新分析了先前发表的一项关于一种流行的微小 RNA 基因发现算法 MiRFinder [30] 的研究中的分类器性能。我们还简要回顾了可用的评估工具。

理论背景

在“理论背景”部分，我们回顾了包括混淆矩阵中的基本度量以及无阈值度量（如 ROC 和 PRC）在内的性能度量。必要时我们还给出了一些简单示例，并对相关工具进行了简要介绍。我们使用三个不同的标签，即 ROC、PRC 和 Tools，来组织这一部分。第一个标签 ROC 代表基本度量、ROC 以及除 PRC 之外的 ROC 替代度量的理论背景。第二个标签 PRC 代表精度、PRC 以及 ROC 和 PRC 之间比较的理论背景。最后，第三个标签 Tools 代表对 ROC、ROC 替代度量以及 PRC 相关工具的简要介绍。我们在各子部分标题的开头使用这些标签，以使整个部分易于理解。

ROC：混淆矩阵中四种结果的组合形成了各种评估指标。

在二分类中，数据被分为两类不同的类别，即正类（P）和负类（N）（见图 1A，左椭圆）。二分类器随后将所有数据实例分类为正类或负类（见图 1A，右椭圆）。这种分类会产生四种结果——两种正确的（或真实的）分类，即真阳性（TP）和真阴性（TN），以及两种错误的（或虚假的）分类，即假阳性（FP）和假阴性（FN）（见图 1B）。由这四种结果构成的 2×2 表格被称为混淆矩阵。所有这些

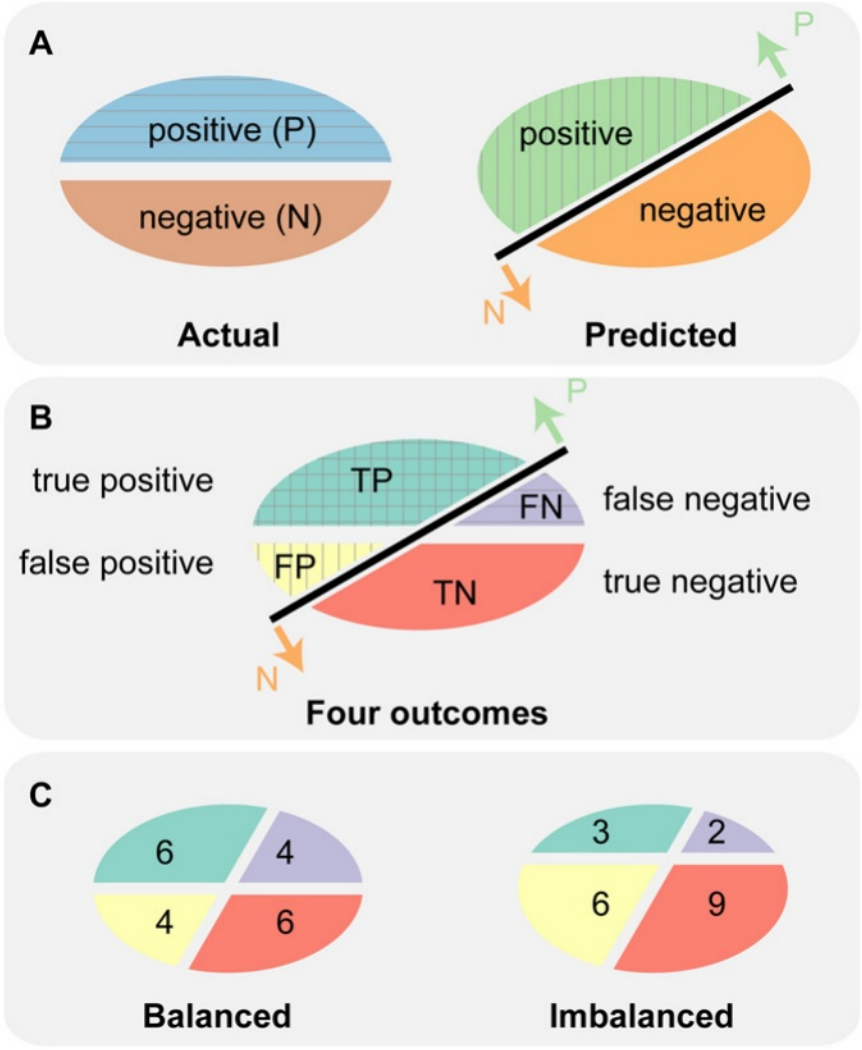

图 1. 实际标签和预测标签生成混淆矩阵的四种结果。（A）左侧椭圆显示了两个实际标签：正例（P；蓝色；上半部分）和负例（N；红色；下半部分）。右侧椭圆显示了两个预测标签：“预测为正例”（浅绿色；左上半部分）和“预测为负例”（橙色；右下半部分）。一条黑色线代表一个分类器，将数据分为“预测为正例”（由向上箭头“P”表示）和“预测为负例”（由向下箭头“N”表示）。（B）将两个实际标签和两个预测标签相结合，产生四种结果：真阳性（TP；绿色）、假阴性（FN；紫色）、假阳性（FP；黄色）和真阴性（TN；红色）。（C）两个椭圆分别展示了平衡数据（左）和不平衡数据（右）中 TP、FP、TN 和 FN 的示例。两个示例均使用 20 个数据实例，平衡数据包含 10 个正例和 10 个负例，不平衡数据包含 5 个正例和 15 个负例。

图 1 doi:10.1371/journal.pone.0118432.g001

表 1.来自混淆矩阵的基本评估指标。

测量	公式
ACC	（真阳性数 + 真阴性数）/（真阳性数 + 真阴性数 + 假阴性数 + 假阳性数）
错误	（假阳性数 + 假阴性数）/（真阳性数 + 真阴性数 + 假阴性数 + 假阳性数）
序列号，真阳性率，召回率	真阳性数 /（真阳性数 + 假阴性数）
SP	真阴性数 /（真阴性数 + 假阳性数）
假阳性率	真阳性数 /（真阴性数 + 假阳性数）
阳性预测值，阴性预测值	真阳性数 /（真阳性数 + 假阳性数）
移动通信运营商（Mobile Communication Carrier）	$(TP * TN - FP * FN) / ((TP + FP) (TP + FN) (TN + FP) (TN + FN))^{1/2}$
F0.5	1.5 * prec * rec /（0.25 * prec + rec）
F_1	2 倍的精度乘以召回率除以（精度加召回率）
F_2	5 乘以 精确率 乘以 召回率 除以 （4 乘以 精确率 加 召回率）

ACC：准确率；ERR：错误率；SN：灵敏度；TPR：真正例率；REC：召回率；SP：特异度；FPR：假正例率；PREC：精确率；PPV：阳性预测值；MCC：马修斯相关系数；F：F 值；TP：真阳性；TN：真阴性；FP：假阳性；FN：假阴性

文献引用：doi:10.1371/journal.pone.0118432.t001

二分类的基本评估指标源自混淆矩阵（见表 1）。

分类器性能最常用的两个基本度量指标是准确率（ACC）和错误率（ERR）[5]。灵敏度（SN）和特异度（SP）也很受欢迎[31]。灵敏度等同于真正例率（TPR）和召回率（REC），特异度等同于 1 - 假正例率（FPR）。另一个度量指标是精确率（PREC），PRC 就是基于它的。精确率也等同于阳性预测值（PPV）。

马修斯相关系数（MCC）[32] 和 Fβ 分数 [33] 也很有用，但使用频率较低。MCC 是根据混淆矩阵中的所有四个值计算得出的相关系数。Fβ 分数是召回率和准确率的调和平均值，其中 β 通常为 0.5、1 或 2。

所有这些度量标准都有各自的优缺点。由于它们在平衡数据集和不平衡数据集中的表现不同，因此在考虑当前数据或未来应用中待分析数据的类别分布时，选择合适的度量标准以进行有意义的性能评估就显得十分重要。

ROC 曲线图对二元分类器进行整体评估。

表 1 列出了用于评估分类器性能的基本度量。所有这些度量都是单阈值度量，也就是说，它们是针对分类器的单个分数阈值（截止值）定义的，无法给出不同阈值下性能范围的概览。虽然在特定应用中，将数据集划分为正预测类和负预测类的任何此类阈值都可能是合理的，但如何选择合适的阈值并不明显。一个强大的解决方案是使用无阈值度量，例如 ROC 曲线和 PRC 曲线。这些无阈值度量要求分类器生成某种分数，从而可以将数据集划分为正预测类和负预测类，而不仅仅是提供静态划分。大多数近期的机器学习库都能够生成判别值或后验概率，这些值可用作分数 [27, 34, 35]，但并非所有分类器都能提供此类值。

ROC 曲线展示了特异性和敏感性之间的权衡关系[24]。它是模型整体的，因为它展示了在所有可能阈值下计算出的特异性和敏感性值的成对情况。

在 ROC 图中，随机性能的分类器会呈现出一条从（0，0）到（1，1）的直线[24]，这条线可以被定义为 ROC 的基准线。ROC 曲线提供了一个单一的性能度量，称为 ROC 曲线下面积（AUC）得分。对于随机分类器，AUC 得分为 0.5，而对于完美分类器，AUC 得分为 1.0 [4]。AUC 得分便于比较多个分类器的性能。

ROC：集中 ROC（CROC）图用于评估分类器的早期检索性能。

ROC 曲线的早期检索（ER）区域（见图 2A 中的灰色矩形区域）对于评估具有高排名实例的数据部分很有用[36, 37]。例如，如果分类器将大量数据预测为正类，那么检查所有被预测为正类的实例可能会很耗时且成本高昂，尤其是在数据集很大时。因此，检查早期检索的性能是切实可行的，因为这只需检查有限数量的高分实例。

集中式 ROC（CROC）图有助于评估早期检索性能[27]。CROC 图是通过一个放大器函数构建的，该函数对 x 轴上的假正率（FPR）进行转换。例如，当使用指数函数且α = 7（见方法部分）时，该函数将 FPR 值[0.0, 0.5, 1.0]转换为[0.0, 0.971, 1.0]。0 到 0.5 的区域被放大，而 0.5 到 1.0 的区域则被缩小。与 ROC 图类似，CROC 曲线下的面积（AUC）同样可用于比较分类器[27]。虽然像 ROC_{50} 这样简单的单一阈值度量（在假正数达到 50 时累加真正数[38]）在评估早期检索性能时可能有用，但 CROC 图提供了整个性能范围的可视化表示，因而具有更高的实用性。

ROC：成本曲线（CC）将误分类成本考虑在内。

成本曲线（CC）是受试者工作特征曲线（ROC）图的另一种替代方案[12, 28]。成本曲线通过改变操作点来分析分类性能[5]。操作点基于类别概率和误分类成本。归一化预期成本或 NE[C] 在 y 轴上表示分类性能[28]。它类似于错误率，因此较低的 NE[C] 值表示更好的分类器。概率成本函数（+）或 PCF（+）在 x 轴上表示操作点[28]。PCF（+）基于正确分类正例的概率，通过类别概率和误分类成本计算得出[5]。PCF（+） 和 NE[C] 的实际计算比 ROC 和 PRC 图的计算复杂得多（有关 PCF（+） 和 NE[C] 的计算，请参阅 S1 文件中的补充方法）。

在评估不平衡数据集上的二分类器时，精确率是一种直观的度量标准。

为了探究分类器性能的基本度量在平衡数据集和不平衡数据集上的表现情况，我们构建了一个简单的示例（见图 1C）。两个数据集的样本量相同。真阳性、假阳性、真阴性和假阴性预测的数量（TP、FP、TN 和 FN）定义如图 1C 所示。表 2 列出了从这两个数据集中得出的基本度量结果。只有精确率、MMC 和三个 Fβ 分数在两个数据集之间存在差异，而大多数度量值保持不变（见表 2 中的“平衡”和“不平衡”两列）。更重要的是，这些未发生变化的度量值无法反映出分类器在不平衡样本中的不良表现。例如，准确率（ACC）表明

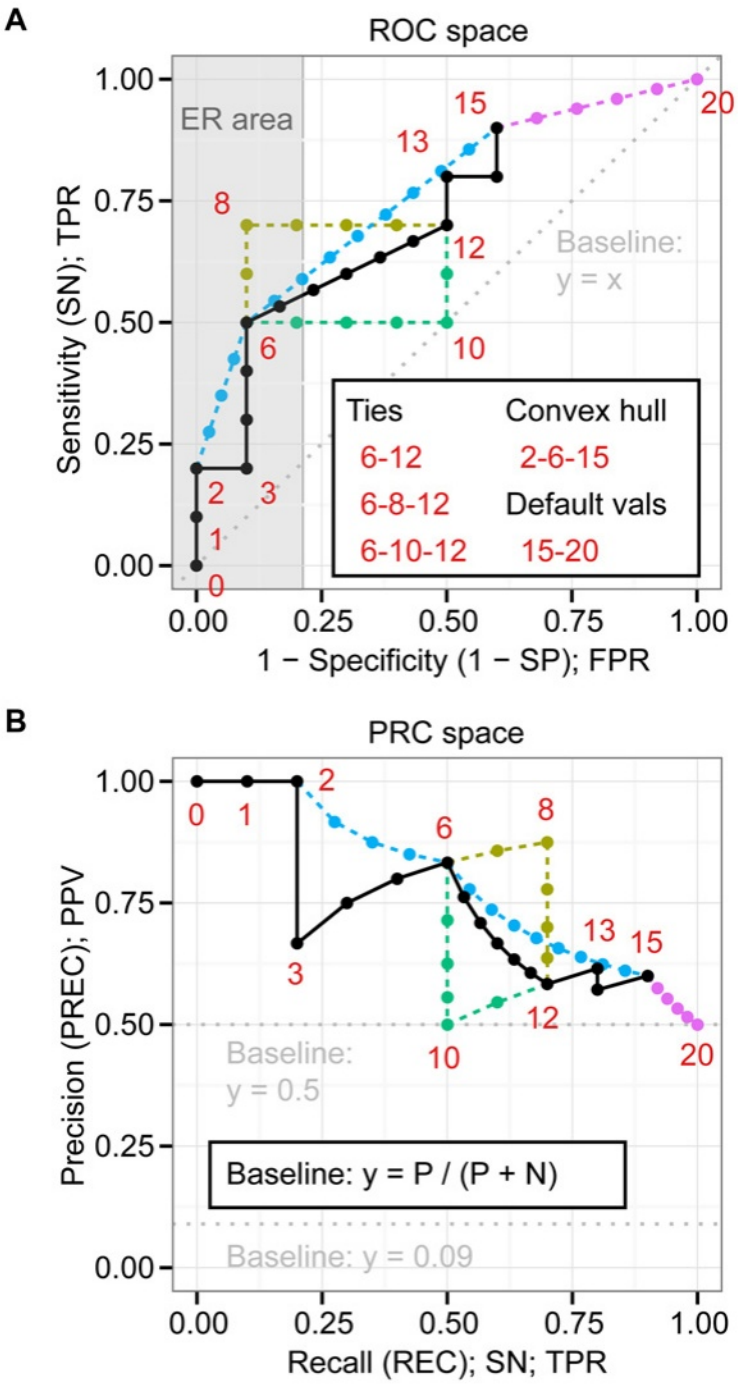

图 2. PRC 曲线与 ROC 曲线具有一一对应关系。（A）ROC 空间包含一条基本的 ROC 曲线和一些点（黑色），以及四条替代曲线和点：下限相等（绿色）、上限相等（深黄色）、凸包（浅蓝色）和缺失预测数据的默认值（品红色）。ROC 点旁边的数字表示从 10 个正样本和 10 个负样本计算 FPR 和 TPR 时的分数排名（实际分数见 S1 文件中的表 A）。（B）PRC 空间包含与 ROC 空间中点相对应的 PR 点。

图 2 doi:10.1371/journal.pone.0118432.g002

表 2. 平衡数据集和不平衡数据集的基本评估指标示例。

测量	平衡的	不平衡的
ACC	0.6	0.6
错误	0.4	0.4
SN（真阳性率，召回率）	0.6	0.6
SP	0.6	0.6
假阳性率	0.4	0.4
预测误差率（付费观看率）	0.6	0.33
移动通信运营商（Mobile Communication Carrier）	0.2	0.17
F0.5	0.6	0.37
F_1	0.6	0.43
F_2	0.6	0.52

有关两个数据集中的真阳性、假阳性、真阴性和假阴性数量，请参见图 1C。

文献引用：doi:10.1371/journal.pone.0118432.t002

该分类器对两个样本的性能表现良好（0.6）。然而，精确率（PREC/PPV）表明，该分类器在平衡数据集上的性能良好（0.6），但在不平衡数据集上的性能相对较差（0.33）。因此，精确率揭示了使用准确率时未被注意到的性能差异。

虽然 MMC 和三个 Fβ 分数在两个数据集之间也有所不同，但精度更容易理解。例如，精度为 0.33 可以立即理解为在所有正向预测中，有 33% 的预测是正确的。这种理解直接适用于将分类器应用于大型数据集的情况，在这种情况下，对正向分类实例（即“预测”）中正确分类的数量进行估计非常重要。精度是衡量这一性能方面的直接且直观的指标。

PRC：PRC 曲线展示了精度与灵敏度之间的关系，其基线会随着类别分布的变化而变化。

精确率 - 召回率（PRC）曲线展示了对应于不同召回率（灵敏度）值的精确率值。与 ROC 曲线类似，PRC 曲线也提供了模型整体的评估。PRC 曲线下的面积得分（AUC （PRC））在多分类器比较中同样有效[26]。

虽然 ROC 的基线是固定的，但 PRC 的基线由正样本（P）和负样本（N）的比例决定，即 y = P / (P + N)。例如，在类别分布平衡的情况下，y = 0.5；而在正负样本比例为 1:10 的类别分布不平衡的情况下，y = 0.09（见图 2B）。由于基线会移动，AUC（PRC）也会随 P:N 比例的变化而变化。例如，随机分类器的 AUC（PRC）仅在类别分布平衡时为 0.5，而在一般情况下，包括平衡和不平衡的分布，其值为 P / (P + N)。实际上，AUC（PRC）与 PRC 基线的 y 位置相同。

PRC：在对 PRC 曲线和 ROC 曲线进行点间插值时，需要采用不同的处理方法。

PRC 曲线与对应的 ROC 曲线之间存在一一对应关系[26]，即两条曲线中的任意一点都能唯一确定另一条曲线上的对应点。然而，在对两点之间进行插值时必须谨慎，因为 PRC 曲线和 ROC 曲线的插值方法不同——ROC 分析采用线性插值，而 PRC 分析采用非线性插值。在 PRC 空间中对两点 A 和 B 进行插值

可以表示为函数 y = （$(TP_A + x)$）/（$\{TP_A + x + FP_A + ((FP_B - FP_A)\ x)$）/（$(TP_B - TP_A)\}$），其中 x 可以是 TP_A 和 TP_B 之间的任意值 [26]。

与插值相关的三个实际的 ROC 特性示例为 ROC 凸包 [39]、处理平分情况以及缺失分数的默认值。为了探究这些特性，我们研究了一个包含 20 个实例的示例，其中正负样本数量相等（见图 2A；分数和标签见 S1 文件中的表 A）。

ROC 凸包给出了分类器可能达到的最佳性能的估计值[39]。它是通过仅连接部分点（图 2A 中的 0 - 2 - 6 - 13 - 15 - 20）的直线组合而成，而原始的 ROC 曲线则是通过将所有点（图 2A 中的 0 到 20 所有点）用直线连接起来。很明显，由于 ROC 凸包跳过了某些点，所以其 AUC 值优于原始 ROC 曲线的 AUC 值。

分类器有时会对预测的部分结果产生平局（即分数相同）的情况（图 2A 中的 6 - 12）。从这些平局中绘制 ROC 曲线的三种明显方法是：先计算正类再使用上限（图 2A 中的 6 - 8 - 12）、先计算负类再使用下限（图 2A 中的 6 - 10 - 12）以及取平均值（图 2A 中的 6 - 12）。ROC 绘图工具通常使用平均值和下限这两种方法[27, 40]。

分类器有时无法对预测的部分内容给出分数。一个这样的例子是在分类前使用过滤。被过滤排除的实例可能没有分配分数。在我们的示例中，ROC 曲线图显示了一个情况，即分类器仅对 15 个实例（图 2A 中的 0 - 15）给出了分数，而其余的 5 个实例（图 2A 中的 16 - 20）没有给出分数。如果为这 5 个实例分配相同的默认值作为补偿缺失分数的措施，ROC 曲线可以线性延伸至点（1，1）（图 2A 中的 15 - 20）。

尽管在 PRC 分析中进行插值所需的计算量比在 ROC 分析中多，但如果要避免产生误导性的图表，遵循正确的步骤仍然至关重要，尤其是在要插值的 PRC 点之间的距离非常大时。在我们的简单示例中，单个点之间的一一对应关系可以在图 2A 和 B 中的 0 - 20 点处看到。

工具：有许多用于绘制 ROC 曲线和 PRC 曲线的工具可免费获取。

有许多工具可以免费绘制 ROC 和 PRC 曲线图，但与 ROC 功能相比，PRC 功能通常较为欠缺。ROCR [40] 是一个流行的 R [41] 包，可用于绘制多种评估图，包括 ROC、PRC 和 CC。但它缺少非线性 PRC 插值计算的功能。AUCCalculator [26] 是一个 Java 应用程序，能够提供准确的 PRC 和 ROC 插值，但不具备绘图功能。CROC [27] 是一个用于 CROC 和 ROC 计算的 Python 包。一些集成的机器学习和生物信息学平台，如 WEKA [34] 和 Bioconductor [42, 43] 也具备绘制 ROC 和 PRC 曲线图的基本功能或库。总体而言，建议将 AUCCalculator 与任何绘图程序结合使用，以生成准确的 PRC 曲线图。ROCR 也可以推荐使用，但前提是不需要在 PRC 点之间进行插值。

材料与方法

基本评估措施

我们从混淆矩阵中计算了基本的评估指标。二分类器的混淆矩阵有四种结果，分别为真阳性（TP）、真阴性（TN）、假阳性（FP）和假阴性（FN）。本研究中讨论的指标有准确率（ACC）和错误率

（错误率）（ERR）、灵敏度（SN）、特异度（SP）、真正例率（TPR）、召回率（REC）、假正例率（FPR）、准确率（PREC）、阳性预测值（PPV）、马修斯相关系数（MCC）[32] 以及 Fβ 分数 [33]（其中 β 为 0.5、1 或 2）。表 1 总结了这些度量的公式。

模型整体评估指标

在我们的研究中分析的模型整体评估指标有 ROC、PRC、CROC 和 CC。我们使用内部的 Python 和 R 脚本计算生成这些指标所需的值。这些脚本还具备绘图功能。ROC 图的 x 轴为 FPR 或 1 - 特异性，y 轴为 TPR 或灵敏度。PRC 图的 x 轴为灵敏度/召回率，y 轴为精确度/阳性预测值。CROC 图的 x 轴为转换后的 FPR，y 轴为 TPR。我们使用指数函数 $f(x) = (1 - \exp(-\alpha x))/(1 - \exp(-\alpha))$，其中 $\alpha = 8$ 来转换 FPR。CC 图的 x 轴为概率成本函数（+）或 PCF（+），y 轴为归一化预期成本或 NE[C] [28]。PCF（+）基于正确分类正例的概率，而 NE[C] 表示分类性能（有关 PCF（+）和 NE[C] 的计算，请参阅 S1 文件中的补充方法）。我们使用 AUCCalculator [26] 和 CROC Python 库 [27] 来计算曲线下面积。

随机抽样模拟

为了分析和比较模型范围内的评估指标，我们通过分别从正样本和负样本的得分分布中随机抽取得分，生成了五个不同分类器性能水平的样本（表 3）。得分较高的实例表明其更有可能被标记为正样本。我们从正态（N）分布或贝塔分布中抽取正样本和负样本，以生成四个不同的水平：随机（Rand）、早期检索差（ER-）、早期检索好（ER+）和优秀（Excel）。对于“完美”（Perf）水平，我们未从分布中抽取，而是使用了固定的值 1（正样本）和 0（负样本）来生成得分。ER- 和 ER+ 的得分分布相似，但 ER+ 在较高（较好）的排名中正样本更多，而 ER- 在较低（较差）的排名中正样本更多。我们将生成的得分存储在数组中以便排序。随后，我们按照得分从低到高的顺序对它们进行排序。对于得分相同的样本，我们按照其在原始数组中的出现顺序分配排名。图 3 展示了这五个水平的得分分布情况。在我们的模拟中，对于平衡数据集，我们使用了 1000 个正样本和 1000 个负样本，而对于不平衡数据集，则使用了 1000 个正样本和 10000 个负样本。一轮模拟使用这些样本计算 ROC、PRC 以及其他图表所需的所有必要指标。然后我们从数据采样重新开始进行下一轮模拟。我们反复执行整个过程。

表 3. 性能模拟中正样本和负样本的得分分布。

等级	优点	负片
随机（兰德）	标准正态分布（均值为 0，方差为 1）	标准正态分布（均值为 0，方差为 1）
早期提取不良（ER-）	测试版（4.1）	贝塔（1，4）
良好的早期恢复（ER+）	贝塔（1，1）	贝塔(1, 4)
优秀的（Excel）	N (3,1)	标准正态分布（均值为 0，方差为 1）
完美（Perf）	1	0

N：均值和方差为参数的正态分布；Beta：形状参数为参数的贝塔分布。对于“完美”性能等级，使用了固定值 1 和 0 。

表 3 doi:10.1371/journal.pone.0118432.t003

图 3. 正负得分分布的组合为模拟分析生成了五个不同的层级。我们随机抽取了 250 个负样本和 250 个正样本，分别对应 Rand、ER-、ER+、Excel 和 Perf，然后将得分转换为从 1 到 500 的排名。红色圆圈代表 250 个负样本，而绿色三角形代表 250 个正样本。

图 3 doi:10.1371/journal.pone.0118432.g003

1000 次。为了绘制曲线，我们在 x 轴上设置了 1000 个区间，并计算了 y 轴上对应值的中位数。

PubMed 搜索

为了探究生命科学研究中二分类器所采用的评估指标，我们进行了两次 PubMed 检索。在第一次 PubMed 检索中，我们旨在了解 ROC 在总体上的流行程度，使用了“ROC 或（受试者工作特征曲线）”这一检索词。从检索结果中，我们收集了 2002 年至 2012 年每年的文章数量。在第二次 PubMed 检索中，我们旨在找出使用支持向量机分类器的全基因组研究，使用了“（支持向量机 AND 全基因组）NOT 关联”这一检索词。我们使用“支持向量机”来查找使用二分类器的研究，使用“全基因组”来查找不平衡数据集的研究。我们还添加了“NOT 关联”以排除全基因组关联研究（GWAS）[44]。截至 2013 年 5 月，检索结果共列出 63 篇文章（见 S1 文件中的表 B）。其中 3 篇综述文章和 2 篇无法获取全文的文章被排除在进一步分析之外。

关于第二次 PubMed 搜索的文献分析

我们对第二次检索到的 58 篇文章进行了人工分析，并根据三个主要类别和 13 个子类别对其进行了分类（见 S1 文件中的表 C 和表 D）。这三个主要类别

类别包括支持向量机（SVM）类型、数据类型和评估方法。我们利用 SVM 类型这一类别来确定 SVM 分类器是否为二分类器，它包含两个子类别，即 BS（二分类 SVM）和 OS（其他 SVM）（见 S1 文件中的表 C）。我们利用数据类型这一类别来确定用于性能评估的数据集是否不平衡，它包含五个子类别，即 IB1（严重不平衡）、IB2（不平衡）、SS（小样本量）、BD（平衡数据）和 OD（其他类型数据）（见 S1 文件中的表 C）。我们利用评估方法这一类别来确定用于评估分类模型的方法，它包含五个子类别，即 ROC、STM1（仅单阈值度量，第 1 组）、PRC、pROC（部分 ROC）、STM2（仅单阈值度量，第 2 组）和 OE（其他评估方法）（见 S1 文件中的表 C）。我们选择了 BS、IB1、IB2、SS、ROC 和 PRC 这些子类别，并计算了每个子类别在总文章数中的比例。此外，我们使用“BS 且（IB1 或 IB2）且非 SS”这一筛选条件对文章进行了筛选。最终的 33 篇文章代表了针对大规模不平衡数据集的二元支持向量机分类研究。

采用 ROC 和 PRC 对 MiRFinder 研究进行再分析

我们为重新分析 MiRFinder 研究生成了两个测试数据集，并将其分别标记为 T1 和 T2（图 4）。数据集 T1 使用来自多个生物体的实际 miRNA 作为阳性样本，而阴性样本则是通过打乱真实 miRNA 的核苷酸生成的伪 miRNA。数据集 T2 使用由 RNAz [45] 生成的所有功能性 RNA 候选序列。为了提取这些候选序列，我们使用了来自加利福尼亚大学圣克鲁兹分校（USCS）基因组生物信息学网站（http://genome.ucsc.edu）的整个秀丽隐杆线虫（C. elegans）多序列比对数据（2008 年 5 月，ce6/WS190）。阳性候选序列是那些与 miRBase [46] 条目重叠的序列，

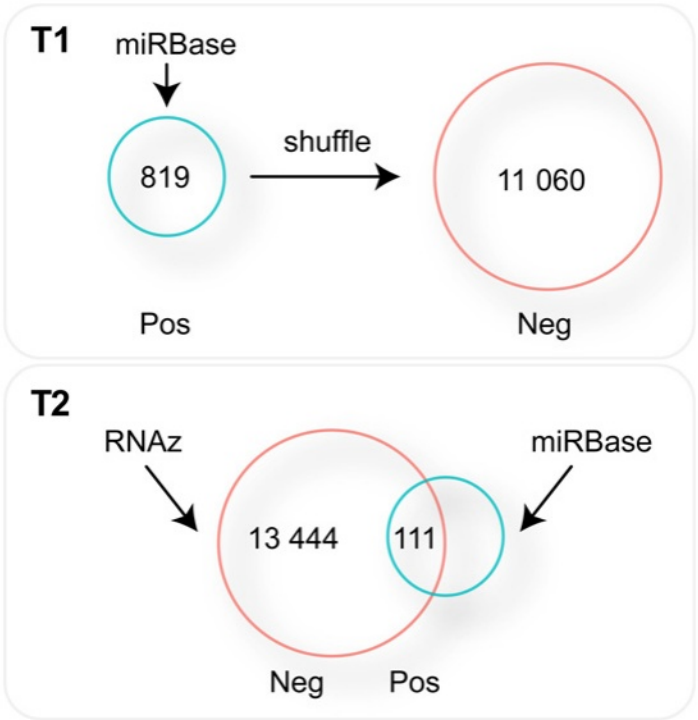

图 4. 数据集 T1 和 T2 生成的简单示意图。T1 包含来自 miRBase 的 miRNA 基因作为阳性样本。阴性样本是通过随机打乱阳性样本的核苷酸生成的。对于 T2，使用 RNAz 工具生成 miRNA 基因候选。阳性样本是与 miRBase 中实际 miRNA 基因重叠的候选基因。

图 4 的 DOI 为：10.1371/journal.pone.0118432.g004

阴性结果是其余的候选功能性 RNA。T1 包含 819 个阳性结果和 11060 个阴性结果，T2 包含 111 个阳性结果和 13444 个阴性结果。为了计算 miRNA 发现工具的得分，我们下载了 MiRFinder [30]、miPred [47]、RNAmicro [48]、ProMir [49] 和 RNAfold [50] 的源代码，并在本地安装。然后，我们在 T1 和 T2 上计算了这些工具的得分（有关测试数据和得分计算的更多详细信息，请参阅 S1 文件中的补充方法）。

结果与讨论

从评估指标的不同视角来看，在数据不平衡的情况下，PR 曲线比 ROC 曲线更具信息量。

在结果部分，我们旨在从多个不同角度展示评估指标在不平衡数据集中的表现情况。我们使用三个不同的标签——模拟、文献分析和再分析来组织结果部分。第一个标签“模拟”代表对随机生成样本的 ROC、CROC、CC 和 PRC 进行的模拟分析。第二个标签“文献分析”代表对两组 PubMed 搜索结果的分析，以探究评估指标在生命科学文献中的实际使用情况。最后，第三个标签“再分析”代表对 MiRFinder 研究的再分析，以揭示 ROC 和 PRC 在实际应用中的差异。我们在各子部分标题的开头使用这些标签，以使整个结果部分易于理解。

模拟：在评估不平衡数据集上的二元分类器时，PRC 图比 ROC、CROC 和 CC 图提供的信息更多。

为了探究 ROC、CROC、CC 和 PRC 曲线之间的差异，我们在平衡和不平衡的情况下进行了随机抽样模拟。为了涵盖广泛的实际相关分类器行为，我们研究了五种不同的性能水平——完美、优秀、早期检索良好（ER+）、早期检索差（ER-）和随机，并分别从不同的正负样本得分分布中随机抽取生成得分（见表 3）。平衡样本包含 1000 个正样本和 1000 个负样本，不平衡样本包含 1000 个正样本和 10000 个负样本。关于这四种不同类型的曲线，我们的观察结果如下。

ROC 曲线。平衡数据集和不平衡数据集的 ROC 曲线没有变化（图 5A），相应的所有 AUC（ROC）得分也未发生变化（S1 文件中的表 E）。图 5A 中 ER-（红色圆点，黑色圆圈）的两个点很好地说明了平衡数据集和不平衡数据集对曲线解释的不同。平衡数据集中的该点代表 160 个假阳性（FP）和 500 个真阳性（TP），如果用该点进行性能评估，ER- 可能被认为是一个好的分类器。相反，不平衡数据集中的同一点代表 1600 个假阳性和 500 个真阳性，在这种情况下，该分类器的性能可能被认为较差。ROC 曲线未能明确显示这种性能差异。此外，这也是解释早期检索区域的 ROC 曲线与 AUC（ROC）之间潜在不匹配的一个好例子。在早期检索区域，ER+ 明显优于 ER-，但 ER- 和 ER+ 的 AUC（ROC）得分相同或均为 0.8（S1 文件中的表 E）。因此，在这种情况下，AUC（ROC）不足以评估早期检索性能。另一个潜在的问题是，当两条 ROC 曲线相互交叉时，AUC（ROC）可能无法进行准确的比较。模拟结果表明，在数据不平衡的情况下，对 ROC 图的解读需要特别谨慎，并且需要检查早期检索区域。

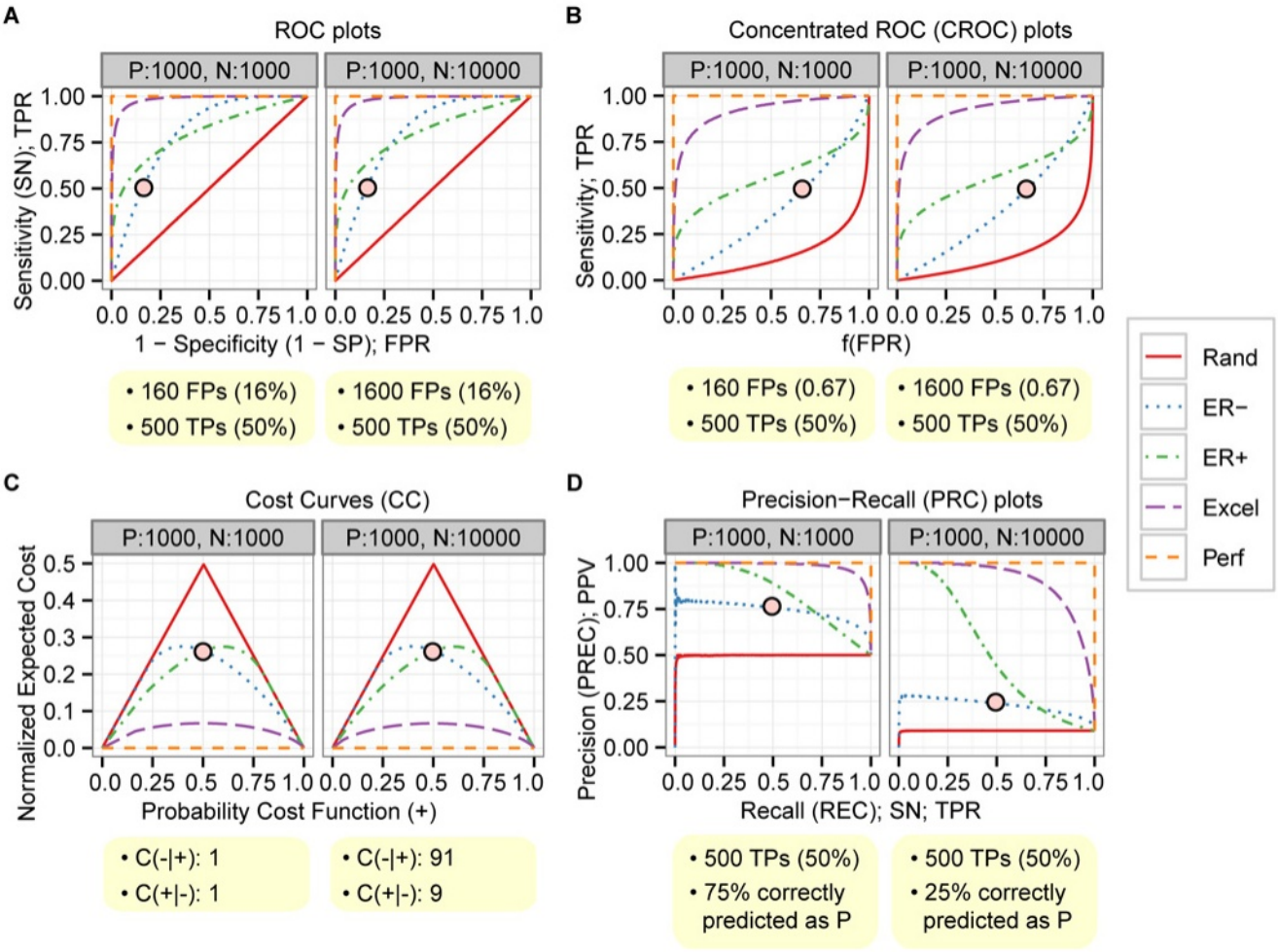

图 5. 在平衡数据和不平衡数据之间，PRC 曲线发生了变化，但其他曲线保持不变。每个面板包含两个图，分别为平衡数据（左）和不平衡数据（右），分别为（A）ROC 曲线，（B）具有指数函数 f(x) = (1 - exp(-α x))/(1 - exp(-α))（其中α = 7）的 CROC 曲线，（C）CC 曲线，以及（D）PRC 曲线。五条曲线代表了五种不同的性能水平：随机（Rand；红色）、早期检索差（ER-；蓝色）、早期检索良好（ER+；绿色）、优秀（Excel；紫色）和完美（Perf；橙色）。

图 5 doi:10.1371/journal.pone.0118432.g005

集中 ROC（CROC）图。与 ROC 图一样，CROC 图（图 5B）在平衡数据集和不平衡数据集之间没有变化。因此，所有 AUC（CROC）得分也保持不变（S1 文件中的表 E）。图 5B 中 ER-（红色圆点带黑色圆圈）的两个点分别代表 TPR 为 0.5 和 f（FPR） 为 0.67。由于当 f（FPR） 约为 0.67 时 FPR 为 0.16，所以这两个点在两种情况下都代表 500 个真阳性，但在平衡数据集中有 160 个假阳性，在不平衡数据集中有 1600 个假阳性。与 ROC 类似，CROC 曲线无法明确显示这种性能差异。不过，早期检索区域的性能差异是明显的，因为该区域大幅扩展，这是 CROC 相对于 ROC 的主要优势。因此，在比较分类器在早期检索区域的性能时，CROC 可能是有用的。然而，CROC 在曲线解释方面与 ROC 存在相同的问题，尤其是在数据集不平衡的情况下。此外，放大器函数的优化参数（如α）通常是未知且难以确定的，尤其是在多个 CROC 曲线相互交叉的情况下。

成本曲线（CC）。平衡数据集和不平衡数据集之间的成本曲线图也未发生变化（图 5C）。就成本曲线而言，它与 ROC 的其他变体有很大的不同。

对图的解读。它展示了基于误分类成本和类别概率的不同 PCF （+） 值的分类性能。C（-|+） 表示将正类误分类为负类的成本，C（+|-） 表示将负类误分类为正类的成本。p（+） 和 p（-） 分别表示正类和负类的概率。误分类成本通常是未知的，但可以根据类别分布进行估计。例如，对于平衡数据集，误分类成本可以是 C(-|+) = 1 和 C(+|-) = 1，而对于不平衡数据集，误分类成本可以是 C(-|+) = 91 和 C(+|-) = 9。这意味着将正类误分类为负类的成本要远高于将负类误分类为正类的成本。要得到 PCF（+） 值为 0.5（图 5C 中带有黑色圆圈的红色点），对于平衡数据集，对应的类别概率为 p(+) = 0.5 和 p(-) = 0.5，而对于不平衡数据集，对应的类别概率为 p(+) = 0.09 和 p(-) = 0.91。一旦确定了感兴趣的 PCF （+） 值，就很容易比较多个分类器的性能。当需要测试各种误分类成本和类别概率时，成本曲线很有用，但必须充分理解 PCF（+） 和 NE[C]。

精确率 - 召回率（PRC）曲线。与 ROC、CROC 和 CC 曲线不同，PRC 曲线在平衡数据集和不平衡数据集之间会发生变化（图 5D）。相应地，AUC（PRC）得分也会发生变化（S1 文件中的表 E）。图 5D 中 ER- 的两个点（红色圆圈中的黑点）分别表示在平衡数据集和不平衡数据集中，正确正预测的比例为 75% 和 25%，且这些正确正预测占所有正样本的 50%。因此，PRC 正确地表明 ER- 在平衡数据集中的表现良好，但在不平衡数据集中的表现不佳。AUC（PRC）得分也支持这一点（S1 文件中的表 E）。此外，PRC 显示 ER+ 在平衡和不平衡数据集中的表现均优于 ER-。同样，AUC（PRC）得分也支持这一点（S1 文件中的表 E）。总之，PRC 能够显示平衡数据集和不平衡数据集之间的性能差异，并且在揭示早期检索性能方面可能很有用。

模拟总结。模拟的总体结果表明，对于不平衡的情况，PRC 是最具信息量和最强大的图表，能够明确揭示早期检索性能的差异。

文献分析：在针对二分类器与不平衡数据集结合的研究中，大多数研究都将 ROC 曲线作为其主要的性能评估方法。

为了评估我们的研究结果在实际应用中的相关程度，我们分析了两组 PubMed 搜索结果（见方法）。第一次分析的目的是定量地确定 ROC 分析在总体上的流行程度。搜索结果表明，ROC 确实是一种流行的方法，且其受欢迎程度在过去十年中稳步上升（图 6；上图）。

第二次分析的目标是从不平衡数据集的二分类研究中进行筛选，以便进一步分析。我们使用 PubMed 检索词“（（支持向量机） AND 全基因组） NOT 关联”来查找使用支持向量机（SVM）进行分类且数据集不平衡的研究[51]。此次检索共得到 63 篇文章，其中 58 篇为可获取全文的研究论文（图 6；下图；完整文章列表及参考文献见 S1 文件中的表 B）。

我们根据三种类别对这 58 篇文章进行了分类：支持向量机（SVM）的类型、数据类型以及评估方法。汇总结果（表 4）显示，大多数研究使用 SVM 构建二分类器（表 4；BS；96.5%），且超过半数的研究使用不平衡数据集（表 4；B1 和 IB2；63.8%）。不出所料，ROC 是最常用的性能评估方法（表 4；ROC，所有；60.3%），其比例甚至更高。

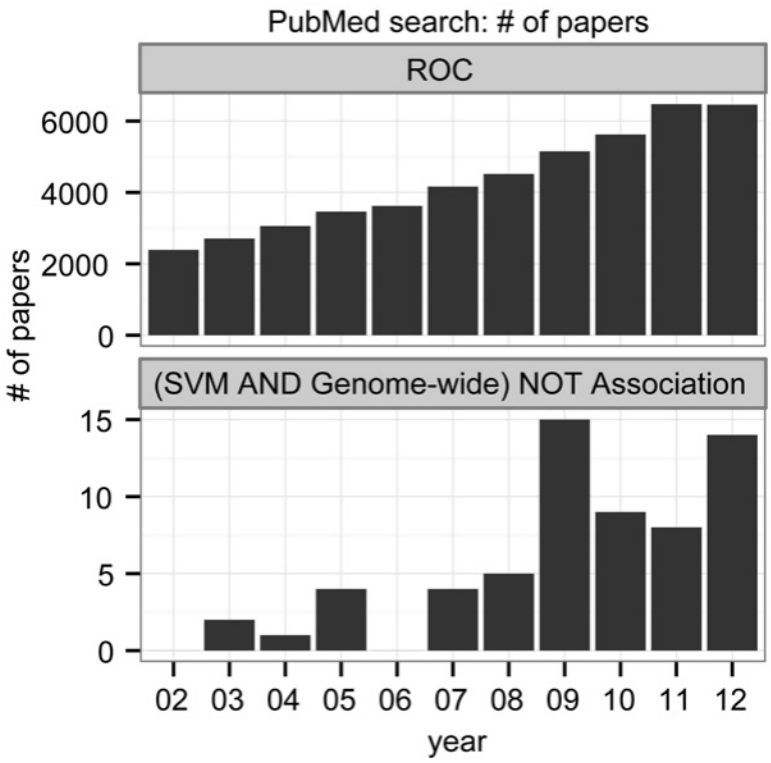

图 6. 两个 PubMed 搜索结果展示了 2002 年至 2012 年间每年找到的论文数量。上方的柱状图显示了通过“ROC”这一术语找到的论文数量，而下方的图表则展示了通过“（（支持向量机） 且 全基因组） 非 关联”这一术语找到的论文数量。

图 6 doi:10.1371/journal.pone.0118432.g006

经过对采用二分类器且数据集不平衡的研究进行筛选后，这一比例略有上升（表 4；ROC、BS 和 IB；66.7%）。这种筛选还排除了样本量较小的研究（表 4；SS；24.1%），因为解决小样本数据不平衡问题的方法可能与解决中等和大规模数据不平衡问题的方法不同[5, 52]。仅有 4 篇论文使用 PRC（表 4；PRC；6.0%）作为评估方法，而 22 篇论文使用 ROC（表 4；ROC；66.7%）。其中，有 3 篇论文同时使用了 ROC 和 PRC。其余 10 篇论文均使用单一阈值度量。

表 4. 按三个主要类别和六个子类别总结的文献分析。

Main[a]	子	描述	All[b]	BS 和 IB[c]
支持向量机	BS	支持向量机二分类器	56 (95.6%)	33 (100%)
数据	IB1	不平衡（负样本 10 折交叉验证）	28 (48.3%)	24 (72.7%)
	IB2	不平衡（负例是正例的 2 到 9 倍）	9 (15.5%)	9 (27.3%)
	SS	样本量小（少于 200）	14 (24.1%)	-
评估	“ROC”	ROC 或 AUC（ROC 曲线）	35 (60.3%)	22 (66.7%)
	中华人民共和国	中华人民共和国或中华民国（PRC）	4 (6.9%)	4 (12.1%)

[a]SVM: 支持向量机类型，数据：数据类型，评估：评估方法。

[b]The 文章总数为 58 篇。

[c]Filtered 通过支持向量机二分类（BS）和不平衡（IB1 或 IB2）且非小样本量（SS）的方法。这类文章的总数为 33 篇。

表 4 doi:10.1371/journal.pone.0118432.t004

文献分析的结果清楚地表明，ROC 是处理不平衡数据时最广泛使用的评估方法，这表明将主要评估方法从 ROC 改为 PRC 可能会对许多研究产生影响。

重新分析：对先前已发表的一项研究进行的重新评估证实了 PRC 图相较于 ROC 图的优势。

为了评估将二元分类器应用于不平衡数据集的研究在主要评估方法从 ROC 曲线变更为 PRC 曲线时可能会受到多大程度的影响，我们从 58 篇可获取全文的研究文章中选取了一篇。

尽管这 58 项研究涵盖了广泛的研究领域，但其中有 5 项研究来自微小 RNA（miRNA）基因发现领域（见 S1 文件中的表 F）。miRNA 是一类在植物和动物中具有重要调控作用的小 RNA [53]，而寻找 miRNA 基因的基因组位置是生物信息学中一个热门但具有挑战性的领域 [54]。我们选择 MiRFinder 研究 [30] 用 PRC 进行重新分析，原因有三：它结合了不平衡数据使用了 ROC，测试数据可用，且分类器能够生成分数，这是能够绘制 ROC 和 PRC 图所必需的。

最初的 MiRFinder 研究评估了另外七种工具（见 S1 文件中的表 G）。仅给出了 MiRFinder 分类器自身的 ROC 曲线，而其他七种工具则提供了 ROC 点（ROC 空间中的单个点）。在 MiRFinder 研究评估的七种额外工具中，我们选择了能够生成分数且源代码可用的三种工具，即 miPred [47]、RNAmicro [48] 和 ProMir [49]，并添加了 RNAfold [50] 作为第四种工具。RNAfold 通过最小化热力学自由能来预测 RNA 二级结构。它并非专门针对 miRNA 的工具，但包括我们重新分析所选的四种工具在内的大多数 miRNA 基因发现工具都高度依赖于最小自由能（MFE）计算。因此，确定更复杂的工具与 RNAfold 的 MFE 计算基准相比能提供多少额外性能是很有趣的。

由于在多种条件下测试性能很有意思，所以我们添加了一个额外的测试集。该测试集是通过使用 RNAmicro 研究 [48] 中描述的方法从秀丽隐杆线虫基因组生成的。

总的来说，我们在两个独立的测试集上评估了五种不同的工具。我们将来自 MiRFinder 研究的测试集标记为 T1，将我们从秀丽隐杆线虫基因组生成的测试集标记为 T2。评估结果如图 7 和表 5 所示，并在以下子章节中进行了描述和讨论。

重新分析：在 T1 上进行测试时，中华人民共和国（PRC）而非中华民国（ROC）表明某些工具表现不佳。

图 7A 中的 T1 上的 ROC 曲线表明，所有分类器的预测性能都非常好到优秀。表现最佳的两个分类器 MiRFinder 和 miPred 的 ROC 曲线相似，但在早期检索区域，miPred 的表现似乎优于 MiRFinder。ROC 曲线并不能直接说明这五种工具的预测结果有多可靠，需要对所显示的假阳性率的实际意义进行一些思考。AUC（ROC）得分（表 5）表明，在整个 FPR 范围内，MiRFinder 略优于 miPred，但这种差异太小，没有实际意义。AUC（ROC）得分与 ROC 曲线的直观印象基本一致，但在实际意义的解释方面同样存在不足。

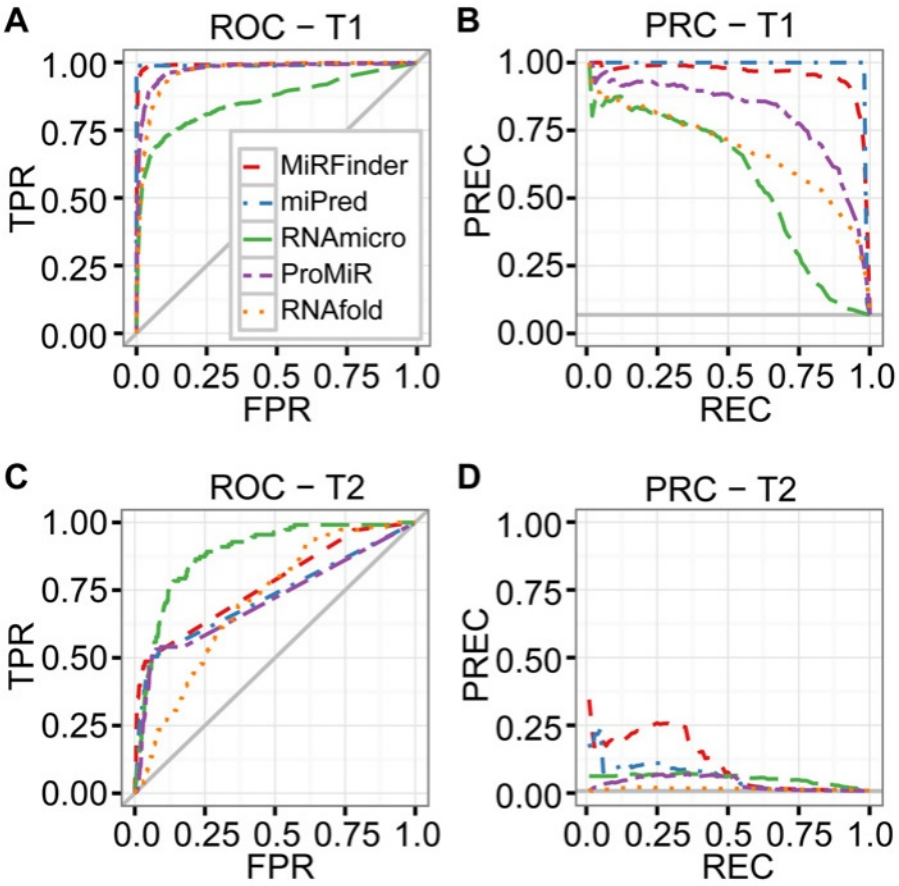

图 7. 对 MiRFinder 研究的重新分析表明，在不平衡数据上，PRC 比 ROC 更强。ROC 和 PRC 曲线展示了六种不同工具的性能，分别是 MiRFinder（红色）、miPred（蓝色）、RNAmicro（绿色）、ProMiR（紫色）和 RNAfold（橙色）。灰色实线代表基准线。重新分析使用了两个独立的测试集 T1 和 T2。四个图分别为（A）T1 上的 ROC 曲线，（B）T1 上的 PRC 曲线，（C）T2 上的 ROC 曲线，（D）T2 上的 PRC 曲线。

图 7 doi:10.1371/journal.pone.0118432.g007

图 7B 中的 PRC 曲线与图 7A 中的 ROC 曲线类似，表明所有分类器的预测性能都非常好到优秀。然而，这里我们可以看到，一些分类器的高召回率伴随着精度的下降，尤其是 RNAmicro，RNAfold 和 ProMiR 也有一定程度的下降，但程度较小。我们还可以看到，分类器的性能得到了更好的区分，使得差异更容易被发现。总体而言，PRC 曲线

表 5. T1 和 T2 的 ROC 曲线和 PRC 曲线的 AUC 分数。

	T1 受试者工作特征曲线（ROC）	中华人民共和国	T2 受试者工作特征曲线（ROC）	中华人民共和国
MiRFinder	0.992 乘以	0.945	0.972	0.106 乘以
miPred	0.991	0.976 乘以	0.707	0.024
RNA微小	0.858	0.559	0.886 乘以	0.054
ProMiR	0.974	0.801	0.711	0.035
RNA折叠	0.964	0.670	0.706	0.015

在数据集 T1 和 T2 上的 ROC 曲线和 PRC 曲线下的面积（AUC）得分。每列中最佳的 AUC 得分用星号（*）标注。

表 5 doi:10.1371/journal.pone.0118432.t005

该图能够快速直观地判断分类器的性能，因为它展示了最相关的两个指标——准确率和召回率之间的权衡。图 7B 中的 PRC 图还表明，所有分类器都明显优于随机分类器，随机分类器的表现由灰色水平基线表示。AUC（PRC）得分（表 5）与 PRC 图中确定的性能顺序一致，但由于它是整个曲线的总结，无法表达在召回率值范围内性能的变化情况。

重新分析：中国在 T2 测试中所有工具的表现都非常糟糕。

图 7C 中的 ROC 曲线与图 7A 大不相同，这应归因于测试数据的差异。RNAmicro 在较宽的假阳性率（FPR）范围内明显领先，尽管 MiRFinder 在早期检索区域表现更强。在 FPR 的中段区域，所有方法都表现出良好的性能，其中 RNAmicro 的表现非常出色，但在 FPR 较小时，真阳性率（TPR）较低。除了 ROC 曲线在判断实际性能方面的一般困难外，图 7C 还需要决定哪些 FPR 区域是相关且可接受的。观察者可能会对中段 FPR 区域的性能感到满意，却未意识到由于数据的严重不平衡，这些 FPR 可能会转化为大量的假阳性预测。AUC（ROC）得分（表 5）表明 RNAmicro 在此次性能竞赛中明显胜出，但自然无法体现性能在 FPR 值范围内的变化，尤其是在这种情况下早期检索区域的变化。

图 7D 中的 PRC 曲线清楚地表明，在此测试集下分类器的性能严重下降。在整个恢复率范围内，除了 MiRFinder 之外，所有方法的精度值都非常低，这对其实际应用价值提出了质疑。MiRFinder 的表现相对较好，在恢复率不是特别低的情况下，精度值也不算特别低，例如在 0.25/0.25 时。虽然图 7C 中的 ROC 曲线给人的印象不错，但图 7D 中的 PRC 曲线揭示了残酷的事实。从实际相关的精度指标来看，除了 MiRFinder 之外，所有方法的性能都接近随机分类器的性能，这由灰色水平线表示。此外，图 7D 中随机分类器的基线低于图 7B 中的基线，这表明测试数据的不平衡性更强，构建良好分类器的难度更大。AUC（PRC）得分（表 5）在候选方法的排名上与 PRC 曲线一致，但同样自然地无法捕捉到 MiRFinder 在不同恢复率范围内的性能变化。

重新分析：在 T1 和 T2 测试中，PRC 比其他指标更直观。

我们还使用 CROC 和 CC 对这五种工具在 T1 和 T2 上进行了评估（见 S1 文件中的图 A：A-D）。与 ROC 曲线相比，CROC 曲线（见 S1 文件中的图 A：A-B）在早期检索区域的分辨率更高，但同样不适合快速判断实际相关性。由于必须考虑转换函数 f，因此解释起来更加困难，不过可以通过在 x 轴上标注原始的 FPR 值来解决这一问题。与 ROC 曲线一样，CROC 曲线也没有显示出在 T2 测试集下性能下降的全部程度。对于成本曲线（见 S1 文件中的图 A：C-D）来说，情况也是如此，而且如果没有对 NE[C] 和 PCF （+） 有很好的理解，这些曲线也难以理解。

重新评估总结。我们重新分析的结果清楚地表明了 PRC 相对于 ROC 的优势。PRC 图展示了具有实际意义的指标，即准确率和召回率，其中准确率尤为重要，因为它衡量的是在所有正向预测中正确预测的比例。PRC 图反映了分类器对不同阈值的敏感度。

具有明显视觉线索的不平衡数据集。PRC 曲线还有助于估计创建良好分类器的难度，因为随机基线的位置取决于正负实例数量的比例。

结论

ROC 是评估二分类器性能的一种流行且强大的指标。然而，在处理不平衡数据集时使用它需要特别谨慎。有人提出 CROC、CC 和 PRC 作为 ROC 的替代方案，但它们的使用频率较低。在我们的综合研究中，我们从多个角度展示了这些不同指标之间的差异。只有 PRC 会随着正负样本的比例变化而变化。

随着高通量测序技术的迅速发展，采用机器学习方法的研究数量可能会增加。我们的文献分析表明，大多数此类研究使用不平衡数据集，并将 ROC 曲线作为主要的性能评估方法。我们在此表明，与 ROC 曲线不同，PRC 曲线能以清晰的视觉线索表达分类器对不平衡数据集的敏感性，并能对实际分类器性能进行准确直观的解释。我们的研究结果强烈建议将 PRC 曲线作为最具信息量的可视化分析工具。

补充信息

S1 文件。包含补充方法、一张补充图表、七张补充表格和补充参考文献。补充方法。成本曲线计算；两个独立测试集 T1 和 T2 的准备；安装四个 miRNA 发现工具和 RNAfold；五个工具在 T1 和 T2 上的预测得分。S1 文件中的图 A。测试数据集 T1 和 T2 上的 CROC 和 CC 图。S1 文件中的表 A。用于制作 ROC 和 PRC 曲线以计算插值的观察标签和预测得分示例。S1 文件中的表 B。通过 PubMed 搜索“支持向量机 AND 全基因组 AND NOT 关联”得到的 63 篇论文列表。S1 文件中的表 C。三个主要类别和 13 个子类别的描述。S1 文件中的表 D。三个主要类别和 13 个子类别对 PubMed 搜索得到的 58 篇研究论文的分类。S1 文件中的表 E。随机抽样模拟中 ROC、PRC 和 CROC 的 AUC 得分。S1 文件中的表 F。文献分析中选出的五项前体 miRNA 研究。S1 文件中的表 G。MiRFinder 研究中用于比较的七种工具。补充参考文献。

（DOCX）

致谢

作者们想感谢卑尔根大学计算生物学单元（CBU）的成员们对本文手稿早期版本提出的宝贵意见。

作者贡献

实验的构思与设计：TS、MR。实验操作：TS。数据分析：TS。论文撰写：TS、MR。

参考文献

1. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S. Machine learning and its applications to biology. PLoS Comput Biol. 2007; 3: e116. PMID: 17604446
2. Krogh A. What are artificial neural networks? Nat Biotechnol. 2008; 26: 195–197. doi: 10.1038/nbt1386 PMID: 18259176

3. Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Ratsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008; 4: e1000173. doi: 10.1371/journal.pcbi.1000173 PMID: 18974822

4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143: 29–36. PMID: 7063747

5. He H, Garcia E. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009; 21: 1263–1284.

6. Chawla N, Japkowicz N. Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explor. 2004;6.

7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling tech-nique. J Artif Intell Res. 2002; 16: 321–357.

8. Rao RB, Krishnan S, Niculescu RS. Data mining for improved cardiac care. SIGKDD Explor. 2006; 8: 3–10.

9. Kubat M, Holte RC, Matwin S. Machine Learning for the Detection of Oil Spills in Satellite Radar Im-ages. Mach Learn. 1998; 30: 195–215.

10. Provost F. Machine learning from imbalanced data sets 101. Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets. 2000.

11. Hulse JV, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th international conference on Machine learning. 2007: 935–942.

12. Guo H, Viktor HL. Learning from imbalanced data sets with boosting and data generation: the Data-Boost-IM approach. SIGKDD Explor. 2004; 6: 30–39.

13. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. In Proceed-ings of the Fourteenth International Conference on Machine Learning. 1997: 179–186.

14. Ling C, Li C. Data Mining for Direct Marketing: Problems and Solutions. In Proceedings of the Fourth In-ternational Conference on Knowledge Discovery and Data Mining. 1998: 73–79.

15. Elkan C. The foundations of cost-sensitive learning. Proceedings of the 17th international joint confer-ence on Artificial intelligence— Volume 2. 2001: 973–978.

16. Sun Y, Kamel MS, Wong AKC, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit. 2007; 40: 3358–3378.

17. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal. 2002; 6: 429–449.

18. Hong X, Chen S, Harris CJ. A kernel-based two-class classifier for imbalanced data sets. IEEE Trans Neural Netw. 2007; 18: 28–41. PMID: 17278459

19. Wu G, Chang E. Class-Boundary Alignment for Imbalanced Dataset Learning. Workshop on Learning from Imbalanced Datasets in ICML. 2003.

20. Estabrooks A, Jo T, Japkowicz N. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Comput Intell. 2004; 20: 18–36.

21. Ben-Hur A, Weston J. A user's guide to support vector machines. Methods Mol Biol. 2010; 609: 223– 239. doi: 10.1007/978-1-60327-241-4_13 PMID: 20221922

22. Mac Namee B, Cunningham P, Byrne S, Corrigan OI. The problem of bias in training data in regression problems in medical decision support. Artif Intell Med. 2002; 24: 51–70. PMID: 11779685

23. Soreide K. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive bio-marker research. J Clin Pathol. 2009; 62: 1–5. doi: 10.1136/jcp.2008.061010 PMID: 18818262 24. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006; 27: 861–874.

25. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240: 1285–1293. PMID: 3287615

26. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. 2006: 233–240.

27. Swamidass SJ, Azencott CA, Daily K, Baldi P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. Bioinformatics. 2010; 26: 1348–1356. doi: 10.1093/bioinformatics/btq140 PMID: 20378557

28. Drummond C, Holte R. Explicitly Representing Expected Cost: An Alternative to ROC Representation. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2000: 198–207.

29. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform. 2012; 13: 83–97. doi: 10.1093/bib/bbr008 PMID: 21422066

30. Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, Zhao SH. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics. 2007; 8: 341. PMID: 17868480

31. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994; 308: 1552. PMID: 8019315

32. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algo-rithms for classification: an overview. Bioinformatics. 2000; 16: 412–424. PMID: 10871264

33. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. Advances in Information Retrieval. 2005: 345–359.

34. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor. 2009; 11: 10–18.

35. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011; 2: 1–27.

36. Hilden J. The area under the ROC curve and its competitors. Med Decis Making. 1991; 11: 95–101. PMID: 1865785

37. Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007; 47: 488–508. PMID: 17288412

38. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate se-quence matching. Comput Chem. 1996; 20: 25–33. PMID: 16718863

39. Macskassy S, Provost F. Confidence bands for ROC curves: Methods and an empirical study. Pro-ceedings of the First Workshop on ROC Analysis in AI. 2004.

40. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21: 3940–3941. PMID: 16096348

41. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comput Graph Stat. 1996; 5: 299–314.

42. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open soft-ware development for computational biology and bioinformatics. Genome Biol. 2004; 5: R80. PMID: 15461798

43. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics. 2008; 9: 461. doi: 10.1186/1471-2105-9-461 PMID: 18959772

44. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005; 6: 95–108. PMID: 15716906

45. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding RNA detection. Pac Symp Biocomput. 2010: 69–79. PMID: 19908359

46. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011; 39: D152–157. doi: 10.1093/nar/gkq1027 PMID: 21037258

47. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 2007; 35: W339–344. PMID: 17553836

48. Hertel J, Stadler PF. Hairpins in a Haystack: recognizing microRNA precursors in comparative geno-mics data. Bioinformatics. 2006; 22: e197–202. PMID: 16873472

49. Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res. 2005; 33: 3570–3581. PMID: 15987789

50. Hofacker I, Fontana W, Stadler P, Bonhoeffer S, Tacker M, Schuster P. Fast Folding and Comparison of RNA Secondary Structures. Monatsh Chem. 1994; 125: 167–188.

51. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory. 1992: 144–152.

52. Raudys SJ, Jain AK. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Trans Pattern Anal Mach Intell. 1991; 13: 252–264.

53. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116: 281–297. PMID: 14744438

54. Gomes CP, Cho JH, Hood L, Franco OL, Pereira RW, Wang K. A Review of Computational Tools in microRNA Discovery. Front Genet. 2013; 4: 81. doi: 10.3389/fgene.2013.00081 PMID: 23720668