计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900216

分类学习算法的性能度量指标综述

杨杏丽
山西大学数学科学学院 太原 030006
山西大学计算机与信息技术学院 太原 030006

摘 要 在机器学习的分类问题研究中,对分类学习算法的正确评价是非常重要的。现实中,许多性能度量指标被从不同的角度提出,文中主要介绍了基于错误率的、基于混淆矩阵的和基于统计显著性检验的三大类性能度量指标,详细地讨论了分类学习算法各性能度量指标的提出背景、意义以及适用范围,分析了各种性能度量之间的差异,提出和分析了各方法中有待进一步研究的问题和方向。进一步,通过实验数据横向(每类度量中各方法之间的类内差异)和纵向(3 类度量之间的类间差异)对照了各性能度量指标之间的差异,分析了各性能度量指标在分类算法选择上的一致性。

关键词:性能度量;错误率;混淆矩阵;统计检验
中图法分类号 TP181

Survey for Performance Measure Index of Classification Learning Algorithm

YANG Xing-li
School of Mathematical Sciences, Shanxi University, Taiyuan 030006, China
School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

Abstract In the research of classification task of machine learning, it is important for correctly evaluating the performance of the learning algorithm. In practical application, many performance measure indexes are proposed based on different perspectives. Three kinds of performance measure indexes based on error rate, confusion matrix and statistical test are introduced in this paper. The background, significance and scope of each measure index are discussed. The differences of different methods are analyzed. The future research problems and directions are also put forward and analyzed. Furthermore, the differences of these performance measure indexes are also compared by experimental data in portrait and landscape. The consistency of these performance measure indexes is also analyzed in classification algorithm selection.

Keywords Performance measure, Error rate, Confusion matrix, Statistical test

1 引言

在人类的日常生活和活动中,人们对外界事物的认识几乎都是通过对事物的分类来进行的,如看到一个东西很自然地想知道它是一个房子,还是一棵树,或者是一个人,这就是最简单的分类。在人类复杂的活动中,如何进行所认识对象的分类是人类发展过程中必须解决的问题,这也是机器学习和模式识别等相关研究领域最基本的研究内容之一[1-3]。典型地,分类问题的结果度量是已知的类别度量,如心脏病的发作与不发作,所谓分类即是根据获得的一组训练样本数据(包括特征和响应)建立预测模型(学习算法),基于此模型去预测新的未知对象的类别结果[4-5]。但只进行简单的分类并不是我们的目的,得到好的分类预测结果才是我们所需要的,因此评价学习算法的分类性能至关重要[6-7]。

分类的目标是把每一个样本划分到两个或多个类别中的某一个类,并把它尽可能正确地分类(注意,这里不考虑多标签分类问题,即一个样本属于多个类别的问题)。由此可见,对学习算法有一个很直接的度量,即使得这种划分相比它的真实类别有尽可能小的错误率。错误率是指在所有可能的样本上类别决策错误的概率,然而,由于数据的分布未知,因此往往得不到错误率的精确表达式。现实中,通常采用经验的基于数据的方法来估计分类学习算法的错误率,如最简单的训练错误率估计、基于独立测试集的测试错误率估计、基于样本重用方法(交叉验证和 Bootstrap 方法)的错误率估计

到稿日期:2020-09-30 返修日期:2020-12-29 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(62076156,61806115);山西省应用基础研究项目(201901D111034,201801D211002);统计与数据科学前沿理论及应用教育部重点实验室开放研究课题(KLATASDS2007)
This work was supported by the National Natural Science Foundation of China(62076156,61806115),Shanxi Applied Basic Research Program (201901D111034,201801D211002) and Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE,ECNU(KLATASDS2007).
通信作者:杨杏丽(yangxingli@sxu.edu.cn)

等[1,4,8-13]。

然而，就像 Yildiz 等[14]于 2011 年指出的那样，错误率度量在两类分类问题中无法区分两类样本分错的程度，即可能第一类分错的样本占总错误率的 90%，而第二类只占 10%，这时把它们同等对待显然不合适。为此，多个基于混淆矩阵的分类算法性能度量指标被提出，包括准确率(precision)、召回率(recall)、F 度量、受试者工作特征曲线(Receiver Operating Characteristic，ROC)、ROC 曲线下面积(Area Under ROC Curve，AUC)度量等[14-21]。

上述两类方法都是直接基于某个度量指标的大小来进行学习算法性能的度量，因此，一些学者指出两个分类学习算法性能度量指标上的微弱差异极有可能是由数据的随机误差(方差)引起的，此时得到的算法度量结果就有可能是错误的。为此，基于统计显著性检验的许多性能度量方法被提出，如最常用的两类样本 t 检验方法、基于交叉验证的 t 检验方法等[22-27]。

在机器学习中，对于一个分类问题，可以提出很多的学习算法，要想度量不同算法之间的性能差异，必须给出相应的算法性能度量指标。算法性能度量用于指导整个学习算法的选择与应用过程，它为最终学习算法或模型的选定提供了质量保证。因此，基于不同的角度，上述提到的多种算法性能度量指标被提出。然而，对于这些度量指标的适用范围、相互关系与差异，现有文献的研究还较少，且大部分研究都是基于少数性能度量指标进行对照的，没有提供一个清楚广泛的综合分析。例如，Zadrozny 等[28]于 2001 年较早使用了不同的概率度量指标均方误差和错误率去评价多种分类学习算法。Cortes 等[29]和 Rosset 等[30]对 AUC 度量和错误率的关系进行了详细的统计分析。Flach 等[31]和 Fuernkranz 等[32]在 ROC 空间从理论上分析了 AUC、准确率和 F 度量的关系。Buja 等[33]于 2005 年通过研究对数损失和平方损失来判断它们是否为合适的得分度量准则。其他相关内容可参考文献[34-41]。

多种性能度量指标对照的相关文献还较少，具体如下。Caruana 等[42]和 Ferri 等[43]于 2004 年基于常用的支持向量机、人工神经网络、最近邻、决策树和 Boosting 5 种分类算法，简单从多维尺度分析和相关性分析的角度实验对照了精确率、F 值、AUC 度量、准确率、召回率等度量指标的性能。Ferri 等[6]于 2009 年通过实验对照了常用的精确率、F 度量、AUC 度量、平方误差和熵损失等性能度量指标的差异，他们主要关注的是这些性能度量指标对于不同的门限值、排序或先验分布等因素的敏感性分析。Sokolova 等[7]于 2009 年分别在两类、多类、多标签、分层分类 4 个方面讨论了常用的多种性能度量指标的性能，他们主要关注的是这些度量指标对于混淆矩阵的变化不变性。

本文把文献中常用的算法性能度量指标按错误率、混淆矩阵和统计检验分为了三大类，详细地讨论了这 3 类分类学习算法性能度量的适用范围、背景和意义，分别从横向(每类度量中各方法之间的类内差异)和纵向(3 类度量之间的类间差异)分析和模拟对照了各性能度量指标之间的差异。进一步，在 UCI 数据库的 5 个常用分类数据集上，基于 3 个广泛使用的分类学习算法考查了上述提到的 3 类性能度量指标在分类算法选择上的一致性。

2 错误率的定义及其错误率估计

2.1 错误率的定义

在分类问题中，人们往往希望尽量减少分类的错误，基于此目标就可以进行分类算法性能的度量。假设分类问题有 c 个类别，各类别状态用 $\omega_i(i=1,\cdots,c)$ 表示，$p(\omega_i)$ 和 $p(x|\omega_i)$ 分别表示类别 ω_i 出现的先验概率和类条件概率密度函数，$\boldsymbol{X}=(X_1,\cdots,X_d)$ 表示 d 维特征向量，分类的目标就是基于此特征向量来预测它属于的类别[1,4]。

所谓错误率指平均错误率，其定义为：

$$P(e)=\int_{-\infty}^{\infty}P(e,x)\mathrm{d}x=\int_{-\infty}^{\infty}P(e\mid x)P(x)\mathrm{d}x \tag{1}$$

其中针对两类分类情形有：

$$P(e|x)=\begin{cases}P(\omega_1|x), & \text{当 } x\in\omega_2\\ P(\omega_2|x), & \text{当 } x\in\omega_1\end{cases}$$

如果通过判别函数把特征空间分为两类，则有：

$$\begin{aligned}P(e)&=\int_{R_1}P(\omega_2|x)P(x)\mathrm{d}x+\int_{R_2}P(\omega_1|x)P(x)\mathrm{d}x\\&=\int_{R_1}P(x|\omega_2)P(\omega_2)\mathrm{d}x+\int_{R_2}P(x|\omega_1)P(\omega_1)\mathrm{d}x\\&=P(x\in R_1|\omega_2)P(\omega_2)+P(x\in R_2|\omega_1)P(\omega_1)\end{aligned} \tag{2}$$

其中，R_1，R_2 分别是判别为 ω_1，ω_2 类的两个区域。相应地，对于多类分类问题，如果把特征空间划分为 c 个区域，则基于这 c 个区域的平均错误率为：

$$P(e)=\sum_{i=1}^{c}\sum_{\substack{j=1\\j\neq i}}^{c}[P(x\in R_j|\omega_i)]P(\omega_i) \tag{3}$$

由式(1)－式(3)可以看到，错误率的计算依赖于数据的分布(包括先验分布)，然而现实中数据的分布往往无法获得或者非常复杂，因此理论错误率的计算非常困难，以至于直接基于它进行算法性能的评价更不可能。

即使在正态分布情形下，错误率的计算也是比较困难的，无法给出完全的解析解的形式，只能通过查表的方式来进行计算，类协方差相等的两类正态情形的结果如下：

$$P_1(e)=\int_{(t+\eta)/\sigma}^{\infty}(2\pi)^{-1/2}\exp(-\xi^2/2)\mathrm{d}\xi$$

$$P_2(e)=\int_{-\infty}^{(t-\eta)/\sigma}(2\pi)^{-1/2}\exp(-\xi^2/2)\mathrm{d}\xi$$

为此，基于数据估计的多种常用的错误率估计方法被提出。

注记 1：最小错误率度量实质上是 0-1 损失下的最小风险度量：

$$EPE=E\sum_{k=1}^{K}L(G_k,\hat{G}(X))P(G_k|X)$$

其中，L 表示 0-1 损失，$\hat{G}(X)$ 为预测的类别。这个问题的最优解称为贝叶斯分类器，它是理论最优解，贝叶斯分类的错误率被称为贝叶斯率。它们同样存在由于无法获取数据分布而得到理论解析解的问题。

注记 2：正确率等于 1 减去错误率，也是实际应用中一个

很重要的度量。

2.2 错误率的估计

2.2.1 训练错误率

最简单的错误率估计为训练错误率,用训练数据上的分类错误比例(分类错误的样本占总训练样本的比例)作为错误率的估计,记为:

$$\hat{P}_1(e)=\frac{1}{n}\sum_{i=1}^{n}I_{\{y_i\neq f^D(x_i)\}} \tag{4}$$

其中,n 为总训练样本个数;$I_{\{y_i\neq f^D(x_i)\}}$ 表示示性函数,当分类错误时它为 1,正确时为 0;f 为类别的预测函数。

然而,遗憾的是训练错误率显然是错误率的一个偏乐观的估计,甚至在极端情形可以通过让分类器记住每个训练样本的类别而使训练错误率严格为 0,但是这样的分类算法的性能并不能代表它在新数据上的表现,且它在没有观测到的新样本上往往表现很差。

2.2.2 测试错误率

在计算训练错误率时将所有数据既用于训练又用于测试往往容易导致过拟合,从而乐观估计真实错误率,因此提出了称为测试错误率的错误率估计,它通过用独立测试集上的数据来估计分类器的性能。具体地,训练数据集被划分为两部分,即训练集 D_1 和测试集 D_2,测试集中的数据不用于算法的拟合,只用于算法性能的评价,记为:

$$\hat{P}_2(e)=\frac{1}{n_2}\sum_{i=1}^{n_2}I_{\{y_i\neq f^{D_1}(x_i)\}} \tag{5}$$

其中,n_2 为测试样本个数。

注记 3:对照等式(4)和等式(5),它们在形式上非常相似,但存在本质上的区别。等式(4)的预测函数的训练和错误率的计算中使用了相同的数据,但等式(5)中它们采用的是不同的数据。测试错误率更加接近于真实的错误率。

2.2.3 基于交叉验证的错误率估计

然而,当样本量较小时,若专门拿出一部分数据来作为独立测试集,训练样本数据将大大减少,从而影响分类器训练的性能以及测试错误率估计的方差,因此提出了基于交叉验证和 Bootstrap 方法的错误率估计。首先,我们给出基于最常用的标准 K 折交叉验证的结果[1,4,44-46]。

具体地,数据集 D 被分成 K 个大小大致相同的数据子集,记为 $D_k^{(v)}$,$k=1,\cdots,K$,令 $D_k^{(t)}$ 表示从数据集 D 中移走 $D_k^{(v)}$ 中的元素得到的第 k 个训练集,那么,基于子集 $D_k^{(t)}$ 训练,$D_k^{(v)}$ 测试 K 个错误率估计而得到的平均标准 K 折交叉验证错误率估计为:

$$\hat{P}_3(e)=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_v}\sum_{(x_i,y_i)\in D_k^{(v)}}I_{\{y_i\neq f^{D_k^{(t)}}(x_i)\}} \tag{6}$$

其中,$n_k\approx n/K$。当 $K=n$ 时,K 折交叉验证即为留一交叉验证。

注记 4:基于其他交叉验证变形(如 RLT(Repeated Learning-Testing)交叉验证、蒙特卡罗(monte-calo)交叉验证、随机 5×2 交叉验证和组块 3×2 交叉验证)的错误率估计在形式上与式(6)类似,只是采用的数据划分方式不同。

注记 5:交叉验证方法通过重采样的方式获得了多个数据集,基于多次重复的平均得到的错误率估计相对于一次实验来说更加准确、可靠,且可以提供它的估计方差。

2.2.4 基于 Bootstrap 的错误率估计

Bootstrap 抽样[12]是另一种通过重采样的方式来模拟多个数据集进行错误率估计的方法,如果记 $D_1,D_2,\cdots,D_B$ 为 B 次有放回地 Bootstrap 随机抽取的数据集,那么基于这 B 次 Bootstrap 抽样结果给出的错误率估计为:

$$\hat{P}_4(e)=\frac{1}{B}\sum_{i=1}^{B}\frac{1}{n}\sum_{i=1}^{n}I_{\{y_i\neq f^{D_i}(x_i)\}} \tag{7}$$

注记 6:对照等式(6)和等式(7),它们虽然都是基于多次重复平均得到的错误率估计,但是它们的重采样方式完全不同,分别采用的是无放回和有放回的抽样。

2.2.5 基于 AIC 和 BIC 的错误率估计

训练错误率是错误率的一个乐观估计,即训练错误率总是低估真实的错误率,如果能估计出它们之间的差,则可以提供错误率的一个较好估计,因此基于赤池信息准则(Akaike Information Criterion,AIC)和贝叶斯信息准则(Bayesian Information Criterion,BIC)的错误率估计方法被提出[47-48]。

乐观性则定义为真实错误率与训练错误率期望的差,表达式如下:

$$op=P(e)-E(\hat{P}_1(e)) \tag{8}$$

AIC 和 BIC 方法是通过估计乐观性来提供真实错误率的一个比训练错误率更好的估计,它们一般用于传统的线性模型情形。

$$\begin{aligned}AIC&=\hat{P}_1(e)+2d\hat{\sigma}^2/n\\&=\hat{\sigma}^2/n(-2\log lik+2d)\end{aligned} \tag{9}$$

$$BIC=-2\log lik+(\log n)d \tag{10}$$

2.3 与错误率相关的其他度量

与错误率具有单调关系的一些判据也常常用于分类器性能的度量,如后验概率分布。一般来说,后验概率分布愈集中,分类错误概率愈小;后验概率分布愈平缓(接近均匀分布),分类错误率就愈大。如在一个极端情形下,如果有 $P(\omega_i|x)=1$ 且 $P(\omega_j|x)=0,\forall j\neq i$,则此时 x 分类为 ω_i,而错误率为 0。

另外,鉴于一般情形下错误率的计算较困难,文献[1]也研究了理论上估算错误率的上界的方法。

3 基于混淆矩阵的分类算法性能度量

Yildiz 等[14]于 2011 年指出,在两类分类问题中,错误率度量无法区分两类样本分错的程度。为了更好地理解这个问题,错误率可以通过混淆矩阵来重新描述。为此,基于混淆矩阵的多个分类算法性能度量指标被提出,包括准确率、召回率、F-度量、敏感度、特异度、ROC 曲线、基于 ROC 曲线的 AUC 度量,以及与 ROC 曲线和 AUC 度量相关的多种变形度量。下文首先介绍混淆矩阵。

3.1 混淆矩阵

不失一般性,仅考虑如下两类分类问题:每个类都和一个二元标签 $l=\{+,-\}$ 相关,用于表示所考虑任务分类的类别。分类算法产生一个预测 z,表示预测得到的类别标签。对于一个具体的二类分类问题,实验结果能被归结为一个 2×2 的矩阵。

表 1 混淆矩阵

Table 1 Confusion matrix

真正类别	预测类别	
	+	−
+	TP	FN
−	FP	TN

表 1 中，TP(True Positives)表示真实正例样本被正确分类为正例样本的数目，TN(True Negatives)表示真实负例样本被正确分类为负例样本的数目，FP(False Positives)表示真实负例样本被错误分类为正例样本的数目，FN(False Negatives)表示真实正例样本被错误分类为负例样本的数目。显然，基于表 1，错误率的表达式为：

$$\hat{p}(e)=\frac{FP+FN}{TP+FP+FN+TN}$$

就像 Yildiz 等[14]指出的那样，经验错误率无法分清楚 FP 和 FN 分错的程度，即可能第一类分错的样本占总错误率的 90%，而第二类只占 10%，这时把它们同等对待显然不合适。因此，基于得到的 TP，TN，FP 和 FN，我们可以计算多种分类算法的性能度量。

注记 7：对于多类分类问题，同样可以得到一个基于多类标签的混淆矩阵，如表 2 所列，错误率的计算同样是非对角线元素个数除以总样本个数。但它同样无法分出错误的细节，如第一类是错分为第二类还是第三类无法得知。当然，这得到的同样是错误率的数据估计结果，理论错误率的度量很难获取。

表 2 多类混淆矩阵

Table 2 Confusion matrix for multiple classes

实际	预测		
	类 1	类 2	类 3
类 1	43	5	2
类 2	2	45	3
类 3	0	1	49

3.2 准确率和召回率

基于表 1 中获得的 TP，TN，FP 和 FN，正类的准确率(precision)和召回率(recall)可分别表示为：

$$p=\frac{TP}{TP+FP} \tag{11}$$

$$r=\frac{TP}{TP+FN} \tag{12}$$

负类的准确率和召回率可类似计算。事实上，理论的准确率定义为当类别预测为正(+)时它的真实类别也为正时的概率，即：

$$p=P(l=+\,|\,z=+) \tag{13}$$

相应地，理论的召回率为：

$$r=P(z=+\,|\,l=+) \tag{14}$$

显然，基于 TP，TN，FP 和 FN 的经验准确率和召回率是理论准确率和召回率的估计。

注记 8：召回率也被称为灵敏度(S_N)和真阳性率，相应地也有与之相对的特异性(S_p)和假阳性率(α)的定义：

$$S_p=\frac{TN}{TN+FP}$$

$$\alpha=\frac{FP}{TN+FP}$$

显然，$S_p=1-\alpha$。

注记 9：对于多类分类问题，准确率和召回率可基于其理论的定义和式(13)、式(14)来计算，即它们等于当预测类别和真实类别都为+时的联合概率除以它们相应的边际概率($P(z+)$或$P(l+)$)。

注记 10：式(11)、式(12)计算的都是单个类别的准确率和召回率，现实中我们常常也需要一个整体的准确率和召回率：

$$p_{ave}=\frac{\sum_{j=1}^{m}\omega(j)p(j)}{m}$$

$$r_{ave}=\frac{\sum_{j=1}^{m}\omega(j)r(j)}{m}$$

其中，$\omega(j)$为类别的权重，$p(j)$和 $r(j)$分别为类别 j 的准确率和召回率。

3.3 F 得分

通过计算准确率和召回率的加权调和平均，可以得到如下的 F 得分：

$$F_\beta=(1+\beta^2)\frac{pr}{r+\beta^2 p}=\frac{(1+\beta^2)TP}{(1+\beta^2)TP+\beta^2 FN+FP} \tag{15}$$

其中，β 是加权参数。特别地，当 $\beta=1$ 时，F 得分为自然语言处理领域广泛使用的 F_1 得分，即：

$$F_1=\frac{2pr}{p+r}=\frac{2TP}{2TP+FN+FP} \tag{16}$$

注记 11：当在实际应用中对照两种分类算法时，可能算法 A 具有高的准确率、低的召回率，而算法 B 具有高的召回率、低的准确率，此时显然基于准确率和召回率度量无法直接识别出两种算法的优劣。基于此，F 度量通过调和它们之间的折中可以提供一个合适的度量指标。

3.4 ROC 曲线

ROC 曲线是根据多个不同的二分类分界值(阈值)，以真阳性率(灵敏度，TPR)为纵坐标，假阳性率(1−特异度，FPR)为横坐标绘制的曲线，如图 1 所示。ROC 曲线与传统方法的最大差异在于：传统方法往往基于某一临界值做出分类的判断，然而 ROC 曲线是基于多个临界值的综合判断来给出一个更加直观、准确的评价，与其他需要给出临界值的度量相比，其不存在临界值的选择问题。图 1 中，曲线越靠近左上角，算法的性能就越高。

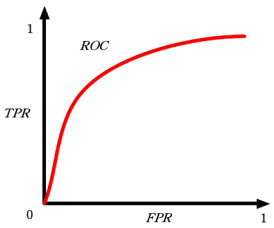


图 1 ROC 曲线

Fig. 1 ROC curve

注记 12：ROC 度量的提出是具有里程碑意义的，它颠覆了传统的只基于某个决策来提供算法性能的度量，它联合多个决策来提供一个更加公平、准确的度量指标。并且，当测试集中正负样本的分布变化时，ROC 曲线保持不变，这对实际

数据集中经常出现的类别不均衡现象有很好的鲁棒性,即在类别分布发生明显改变时依然能客观地评价分类器的性能。ROC 曲线现在已经成为计算机各相关领域广泛使用的算法性能度量指标[29,39,41]。

注记 13:当然,如果为了进行对照,只基于某个阈值也可以画出 ROC 曲线,并且此时可精确计算第 3.5 节中的 AUC 度量值。

3.5 AUC 度量

ROC 曲线提供了算法性能的一种新的图形化度量方式,曲线越靠近左上角,其对应算法的性能越高。但是,当两种算法的性能曲线交叉时,此度量将无法辨别这两种算法的差异。为此,一个新的基于 ROC 曲线的数值度量指标被提供。AUC,即 ROC 曲线下的面积,是一个介于 0 和 1 之间的数值度量指标,它可以直观准确地评价分类算法性能的好坏,其值越大,算法的性能就越好。

从图 1 可以看到,很难计算出精确的 AUC 值,一般有 3 种近似计算方法:1)类似于图 2 的梯形近似法,如果划分得足够细,它将趋于真实的 ROC 曲线;2)通过 AUC 与 Wilcoxon-Mann-Witney 检验的等价性进行 AUC 值的计算;3)通过如下近似公式进行计算:

$$AUC=\frac{\sum_i rank_i-M(1+M)/2}{M\times N}$$

其中,M 和 N 分别表示正负样本的个数。

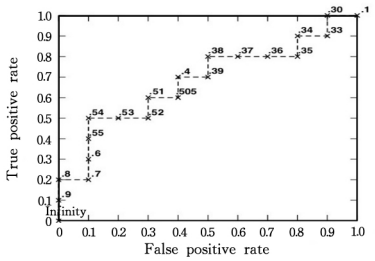


图 2 ROC 近似

Fig. 2 ROC approximation

特别地,对于两类分类问题,当类分布为均匀分布时,AUC 度量可表示为:

$$AUC(j,k)=\frac{\sum_{i=1}^{m} f(i,j)\sum_{t=1}^{m} f(t,k)I(p(i,j),p(t,j))}{m_j\times m_k} \quad (17)$$

其中,$f(i,j)$ 和 $p(i,j)$ 分别表示第 i 个样本被分类为第 j 类的真实和估计概率,m_j 是类样本个数,I 是一个对照函数。当 $a>b$ 时,$I(a,b)=1$,当 $a<b$ 时,$I(a,b)=0$,当 $a=b$ 时,$I(a,b)=0.5$。

把两类分类问题扩展到多类分类,则有如下 4 种 AUC 性能度量指标[19,49-50]:

$$AUNU=\frac{\sum_{i=1}^{c} AUC(j,rest_j)}{c} \quad (18)$$

$$AUNP=\sum_{j=1}^{c} p(j)AUC(j,rest_j) \quad (19)$$

$$AULU=\frac{1}{c(c-1)}\sum_{j=1}^{c}\sum_{k\neq j}^{c} AUC(j,k) \quad (20)$$

$$AULP=\frac{1}{c(c-1)}\sum_{j=1}^{c}\sum_{k\neq j}^{c} p(j)AUC(j,k) \quad (21)$$

其中,c 为类别个数,$p(j)$ 为类的先验概率。

注记 14:如图 3 所示,如果我们只基于一个阈值来绘制 ROC 曲线,AUC 值则很容易基于 TPR 和 FPR 值来计算,$AUC=(1+TPR-FPR)/2$[51]。

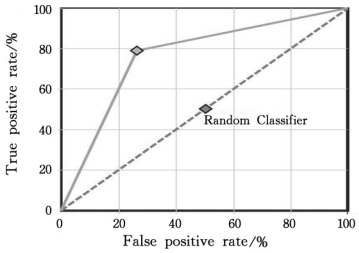


图 3 单个阈值的 ROC 曲线

Fig. 3 ROC curve for a threshold value

3.6 其他度量

除了上述的几个度量外,基于混淆矩阵的度量还有 KS(Kolmogorov-Smirnov)曲线、提升(lift)值和增益(gain)值,以及经常在多分类问题中使用的 Kappa 系数等。其中,KS 曲线和 AUC 曲线相同,都使用 TPR 和 FPR 两个指标来衡量模型的好坏,KS 曲线指以样本数为横坐标、以 TPR 和 FPR 为纵坐标,绘制的正样本曲线和负样本曲线,这两条曲线的最大间隔距离即为 KS 值,该值体现了模型的分类能力,其值越大越好。KS 值的表达式为:

$$KS=\max\{TPR-FPR\}$$

Lift 值是正例的准确率与正例所占比例之间的比值,它表示与不使用模型相比,使用模型后预测能力提升的程度,而 Gain 值本质上是正例的准确率,它们的表达式为:

$$Lift=\frac{TP/(TP+FP)}{(TP+FN)/n},Gain=\frac{TP}{TP+FP}$$

而经常用于多分类问题的 Kappa 系数为实际一致性与非机遇一致性的比值,评价模型的分类结果与随机分类的结果相比误差减少的比例,该值越高,模型的分类准确度就越高,其计算方法为:

$$Kappa=\frac{P_0-P_e}{1-P_e}$$

其中,P_0 表示模型的准确率,即:

$$P_0=\frac{\text{混淆矩阵的对角线元素和}}{\text{整个矩阵元素之和}}$$

P_e 表示模型随机猜对的概率,即:

$$P_e=\frac{\sum_i \text{第 } i \text{ 行元素}*\text{第 } i \text{ 列元素之和}}{(\text{整个矩阵元素之和})^2}$$

4 基于统计检验的性能度量

第 2 节、第 3 节介绍的两大类算法性能度量方法皆是直接基于某个度量指标的大小来进行学习算法性能的度量,然而两种分类学习算法之间的性能差异极有可能是由数据的随机性引起的,即它们之间的差异完全淹没在方差中,因此此时得到的算法性能差异的结论就极有可能是错误的。为此,基于统计显著性检验或置信区间的许多性能度量方法被提出。

对于两种分类算法 A 和 B,如果记 PMI(Performance Measure Index)为某个性能的度量指标,则统计显著性

检验具有如下形式。

原假设 $H0: PMI_A = PMI_B$

对立假设 $H1: PMI_A \neq PMI_B$

4.1 McNemar 检验

McNemar 检验是一种很古老的两种算法性能对照的统计检验方法，由 Everitt[52] 于 1977 年提出，它通过考虑两种算法分类正确和错误的样本个数来构造检验统计量。

如果记 $n_{00}, n_{01}, n_{10}, n_{11}$ 分别为两种算法 A 和 B 都误分类的样本个数、算法 A 误分类但算法 B 正确分类的样本个数、算法 A 正确分类但算法 B 误分类的样本个数、算法 A 和 B 都正确分类的样本个数，且 $n_{00}+n_{01}+n_{10}+n_{11}=n$，则在原假设成立的条件下两种算法应该具有相同的错误率，可以建立如下的列联表。

n_{00}	$\frac{n_{01}+n_{10}}{2}$
$\frac{n_{01}+n_{10}}{2}$	n_{11}

基于此列联表可构造如下的卡方分布检验统计量：

$$(|n_{01}-n_{10}|-1)^2/(n_{01}+n_{10}) \sim \chi^2(1) \tag{22}$$

注记 15：McNemar 检验方法虽然简单，但是它在实际应用中非常有效。缺陷是此检验方法只依赖于一次数据划分，划分的好坏直接影响最后的对照结果。

4.2 *K* 折交叉验证 t 检验

K 折交叉验证 t 检验是目前在机器学习领域最流行的检验方法之一，它随机将数据集划分为 K 个不相交的子集，使用 $K-1$ 个子集来训练，剩下的 1 个子集来测试，将所有的 K 个子集轮流测试，模拟出 K 个训练集和测试集，通过联合这 K 个训练和测试结果构造检验统计量。K 折交叉验证 t 检验方法可以直接解决或缓解上述 McNemar 检验过分依赖于某次数据划分结果的问题。

具体地，如果记 $T_1, T_2, \cdots, T_K$ 为由数据集 D 随机划分的 K 个子集，$D_1, D_2, \cdots, D_K$ 为由任意 $K-1$ 个子集组成的 K 个不同的训练集，则基于此 K 个训练和测试集构造的 K 折交叉验证 t 检验统计量具有如下形式：

$$t_{KCV}=\frac{\hat{\mu}_K(A)-\hat{\mu}_K(B)}{\sqrt{S^2_{\hat{\mu}_K(A)-\hat{\mu}_K(B)}/K}} \sim t(K-1) \tag{23}$$

其中，$\hat{\mu}_K(A)=\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_k(A)$，$\hat{\mu}_K(B)=\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_k(B)$ 分别为算法 A 和 B 基于某一性能度量指标的平均估计值，$S^2_{\hat{\mu}_K(A)-\hat{\mu}_K(B)}$ 为方差估计。

注记 16：K 折交叉验证 t 检验方法的优劣主要取决于它的方差估计部分，如常用的样本方差估计将因低估真实方差而导致其检验是激进的，进而可能得出原本没有显著差异的两种算法被检验出有显著差异的错误结论。另外，K 折交叉验证中 K 值的选择是一个值得研究的问题，常用的有 2，5，10 等[22-24]。

4.3 5×2 交叉验证 t 检验

5×2 交叉验证 t 检验是由 Dietterich[22] 于 1998 年提出的能解决 K 折交叉验证方差低估问题的一种有效的统计检验方法，在分类算法的性能对照中被广泛使用。它基于 5 次重复的二折交叉验证来构造检验统计量，如果记 $\hat{\mu}_j^{(i)}(A)$，$\hat{\mu}_j^{(i)}(B)$，$i=1,\cdots,5$，$j=1,2$ 分别为算法 A 和 B 在第 i 次重复第 j 折交叉验证上基于某一性能度量指标的估计值，则 5×2 交叉验证 t 检验统计量可以表示为：

$$t_{5\times 2CV}=\frac{\hat{\mu}_1^{(1)}(A)-\hat{\mu}_1^{(1)}(B)}{\sqrt{\sum_{i=1}^{5}S_i^2/5}} \sim t(5) \tag{24}$$

其中，$\hat{\mu}_1^{(1)}(A)-\hat{\mu}_1^{(1)}(B)$ 表示第 1 次重复第 1 折交叉验证上算法 A 和 B 性能的差，S_i^2 表示第 i 次重复中算法性能差的样本方差[22-24]。

注记 17：5×2 交叉验证方法的提出为算法性能的对照提供了一个新的思路，此方法的引用率达到了 3 000 多次。随后，还有 5×2 交叉验证 F 检验等多种版本的检验方法被提出，但此方法中 5 次重复的选择带有一定的随机性，只是对于一个经验的选择，更多次的重复可能会得到更好的结果。

4.4 组块 3×2 交叉验证 t 检验

Wang 等[24] 注意到 5×2 交叉验证的 5 个训练集和测试集中由于样本重叠个数不同而无法准确估计其方差，因此于 2014 年提出了具有相同重叠样本个数的组块 3×2 交叉验证以及相应的组块 3×2 交叉验证 t 检验方法。基于 3 次特殊重复的二折交叉验证得到的组块 3×2 交叉验证 t 检验可以表示为[24]：

$$t_{3\times 2CV}=\frac{\hat{\mu}_{3\times 2}}{\sqrt{S_{3\times 2}}} \sim t(5) \tag{25}$$

其中：

$$\hat{\mu}_{3\times 2}=\sum_{i=1}^{3}\sum_{j=1}^{2}(\hat{\mu}_j^{(i)}(A)-\hat{\mu}_j^{(i)}(B))/6$$

$$S_{3\times 2}=\sum_{i=1}^{3}\sum_{j=1}^{2}(\hat{\mu}_j^{(i)}(A)-\hat{\mu}_j^{(i)}(B)-\hat{\mu}_{3\times 2})^2/6$$

注记 18：Wang 等[53] 于 2017 年把组块 3×2 交叉验证扩展到了组块 $m\times 2$ 交叉验证，并提出了更广泛的组块 $m\times 2$ 交叉验证检验方法。

4.5 其他检验

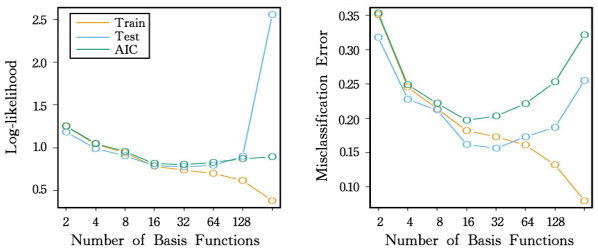
上述检验方法是对多个度量指标都通用的检验方法，还有一些检验是专门适用于某些度量指标的，如 Wang 等于 2015 年[17] 和 2016 年[25] 提出的只专门适用于准确率、召回率、F 得分的置信区间（统计显著性检验）方法，他们证明了基于 K 折交叉验证准确率和召回率服从 Beta 分布，从而给出了基于 Beta 分布的准确率和召回率置信区间度量，而基于组块 3×2 交叉验证的 F 得分的分布可以由 Beta 分布导出，但它不能精确地服从 Beta 分布[17,25]。

5 实验对照

本节将通过实验数据来对照这 3 大类方法的性能，包括各类方法之间的对照以及每类方法内部的对照。

5.1 错误率相关度量性能对照

例 1[4] 通过拟合逻辑斯蒂回归模型来区分两个音素“aa”和“ao”的训练和测试错误率，以及 AIC 错误率估计，如图 4 所示。



(a)Log-likelihood Loss (b)0-1 Loss

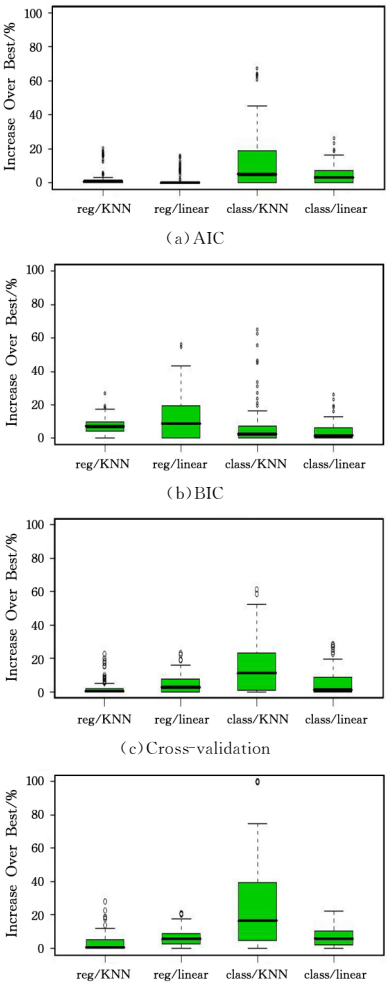
图 4 训练和测试错误率

Fig. 4 Training and test error rate

从图 4 可以看到,无论是对数似然损失还是 0-1 损失,训练错误率都在最下方,即训练错误率低估真实的测试错误率,且在一定条件下它可以趋于零,但真实错误率不可能为零。AIC 方法通过估计乐观性一定程度上缓解了训练错误率的低估问题。

例 2[4] 考虑包含 50 个观测、20 个预测变量的数据集,这些预测变量均匀地分布在超立方体$[0,1]^{20}$上,采用最近邻(KNN)和最优子集回归两种方法进行回归和分类实验。其中,图 5 给出了以下这个量的分布。

$$100\cdot\frac{Err(\hat{\alpha})-\min\limits_{\alpha}Err(\alpha)}{Err(\alpha)-\min\limits_{\alpha}Err(\alpha)}$$



(a)AIC (b)BIC (c)Cross-validation (d)Bootstrap

图 5 4 种方法的盒图对照

Fig. 5 Boxplots for four methods

其中,横坐标表示的是回归还是分类问题,以及使用的是 KNN 模型还是线性模型,如 class/KNN 表示分类问题中采用 KNN 模型。通过对照 AIC、BIC、交叉验证和 Bootstrap 4 种方法在图 5 所示的 4 种情况下的表现得出如下结论:由于图 5 给出了所选择的模型相对于最佳模型的误差,即它越接近于 0 越好,因此总的来说 AIC 和交叉验证方法优于 BIC 和 Bootstrap 方法。但另一方面,AIC 方法在 4 种情况下对预测误差的过高估计分别为 38%,37%,51%和 30%,BIC 方法与 AIC 方法的表现类似。比较而言,交叉验证方法对误差的估计则降低到了 1%,4%,0%和 4%,Bootstrap 方法的性能与交叉验证方法大致相同。因此,交叉验证方法和 Bootstrap 方法具有较高的估计精度,然而它们的计算开销相比 AIC 和 BIC 方法大很多。

5.2 基于混淆矩阵相关度量的性能对照

文献[6]从另一个角度对照了基于混淆矩阵的 6 个度量指标,即错误率(Err)、F 得分,以及 4 种 AUC 度量(ANU,ANP,ALU,ALP)的性能。通过考察这些度量指标在 30 个数据集上的相关性来讨论这些指标之间的共性和差异。

秩相关和线性相关的结果分别如表 3 和表 4 所列。首先,可以明确地看到秩相关和线性相关的结果几乎没有差异,它们之间最大的差异也仅有 0.01。这说明在实际使用相关性进行度量指标差异的度量时,两种相关性指标只使用一种即可,因为它们得出的结论是一致的。

表 3 6 种度量方法的秩相关结果

Table 3 Rank correlation for six measures

	Err	*F*	*ANU*	*ANP*	*ALU*	*ALP*
Err	—	0.95	0.70	0.72	0.70	0.72
F	0.95	—	0.70	0.69	0.71	0.71
ANU	0.70	0.70	—	0.98	1.00	0.99
ANP	0.72	0.69	0.98	—	0.97	1.00
ALU	0.70	0.71	1.00	0.97	—	0.99
ALP	0.72	0.71	0.99	1.00	0.99	—

表 4 6 种度量方法的线性相关结果

Table 4 Linear correlation for six measures

	Err	*F*	*ANU*	*ANP*	*ALU*	*ALP*
Err	—	0.95	0.69	0.71	0.69	0.71
F	0.95	—	0.70	0.69	0.70	0.70
ANU	0.69	0.70	—	0.98	1.00	0.99
ANP	0.71	0.69	0.98	—	0.97	0.99
ALU	0.69	0.70	1.00	0.97	—	0.98
ALP	0.71	0.70	0.99	0.99	0.98	—

其次,错误率和 *F* 得分度量有很强的相关性,在两种相关性指标下它们的相关性都是 0.95,这表明在整体数据集上这两种度量指标是没有显著差异的。然而,在不平衡数据集下这两个度量指标的相关性明显下降,大约只有 0.75。错误率与 AUC 的 4 个度量指标之间的相关性都较小,大约为 0.70,这表明了 AUC 与错误率是两类完全不同的性能度量,进一步从数值上验证了这两类方法的差异。F 得分与错误率有很大的相关性,因此它与 AUC 度量的相关性与错误率类似,只有 0.70。

进一步来看 4 个 AUC 性能度量指标之间的差异,它们两两之间的相关性(无论是线性相关还是秩相关)几乎都是

100%，这也说明了虽然这些 AUC 的变形是从不同角度提出的，但它们在算法性能的度量上几乎是相同的。特别是 ALU 和 ANU 度量，它们在 30 个数据集上对于线性相关和秩相关都是完全相关的，即这两个度量是完全相同的。因此，在使用 AUC 进行算法性能度量时，只需要考虑其中的一个即可。

文献[7]则从错误率、准确率、召回率、F 得分等基于混淆矩阵的度量，在对于混淆矩阵的变化不变性的角度，对照了这些度量的差异，如表 5 和表 6 所列。

表 5 两类分类情形度量方法的变化不变性

Table 5 Invariance properties of performance measures for binary classification

	I1	I2	I3	I4	I5	I6	I7	I8
Err	+	−	−	−	−	+	−	−
p	−	+	−	+	−	+	+	−
r	−	+	−	−	+	+	−	+
F	−	+	−	−	−	+	−	−
AUC	−	−	−	−	−	+	−	+

表 5 中，Err，p，r，F 分别表示错误率、准确率、召回率和 F 得分度量，I1，I2，I3，I4，I5，I6，I7，I8 分别表示混淆矩阵中对角线元素交换、TN 值变化、TP 值变化、FN 值变化、FP 值变化、所有元素乘以相同的倍数、列乘以的倍数不同、行乘以的倍数不同这 8 种不同的变化。

表 6 多类分类情形度量方法的变化不变性

Table 6 Invariance properties of performance measures for multi-class classification

	I1	I2	I3	I4	I5	I6	I7	I8
Err	+	−	−	−	−	+	−	−
p	−	+	−	+	−	+	+	−
r	−	+	−	−	+	+	−	+
F	−	+	−	−	−	+	−	−

从表 5 和表 6 可以看到，对于两类和多类分类情形，这些度量的混淆矩阵变化不变的结果是一致的，即此性质不受类别多少变化的影响。具体来看，只有 I6 所有的度量都有变化不变性，这也是显然的，因为 I6 的变化只是对整个矩阵乘以一个相同的倍数，由这些度量公式很容易推出这个性质。其次是 I2，它对于准确率、召回率和 F 得分具有不变性，最差的是 I3，它对于所有度量都不能保持不变性。对各个度量进行对照，准确率和召回率保持变化不变性的情形是最多的，共有 4 种情形，而其他度量只有 2 种情形。

5.3 基于统计检验的性能度量指标对照

首先，Dietterich[22]于 1998 年通过实验对照了 McNemar 检验、K 折交叉验证 t 检验和 5×2 交叉验证 t 检验的性能，结果如图 6 所示。

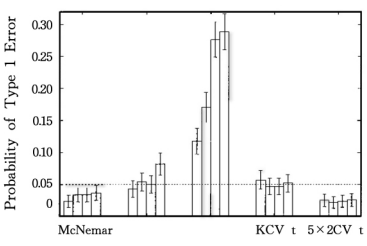


图 6 第一类错误的概率

Fig. 6 Probability of Type I error

图 6 给出了常用的 K 折交叉验证 t 检验方法在设定的 4 种模拟情形下第一类错误都接近或略高于显著性水平 0.05，而 McNemar 检验和 5×2 交叉验证 t 检验的第一类错误明显低于 0.05，尤其是 5×2 交叉验证 t 检验方法。

Wang 等[24]于 2014 年对照了 K 折交叉验证 t 检验、5×2 交叉验证 t 检验和组块 3×2 交叉验证 t 检验的性能，他们通过实验验证了组块 3×2 交叉验证 t 检验具有更小的犯第一类错误的概率和更大的势函数，如表 7 和图 7 所示。

表 7 第一类错误的概率

Table 7 Probability of type I error

	情形 1	情形 2
KCVt 检验	0.08	0.07
校正 KCVt 检验	0.07	0.05
5×2CVt 检验	0.05	0.05
组块 3×2CVt 检验	0.05	0.05

表 7 中，情形 1 和情形 2 分别对应具有不同均值和方差的正态数据分布情形，情形 1 的均值和方差为(0，1)和(1，1/6)，情形 2 的均值和方差为(0，1)和(1，0.173)。

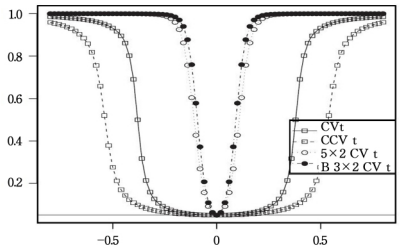


图 7 4 种检验的势函数

Fig. 7 Power functions for four tests

5.4 不同类的性能度量指标对照

例 3[27]考虑一个单变量多项式分类问题，自变量由[−1，1]上的均匀分布生成，类别变量通过比较后验概率

$$P(C_0|x)=\frac{P(C_0)P(x|C_0)}{P(C_0)P(x|C_0)+P(C_1)P(x|C_1)}$$

和

$$P(C_1|x)=\frac{P(C_1)P(x|C_1)}{P(C_0)P(x|C_0)+P(C_1)P(x|C_1)}$$

的大小得到，若 $P(C_0|x)>P(C_1|x)$，则 r' 取值为 0，否则为 1。类条件密度通过如下的两类高斯混合模型

$$P(x|C_0)=0.5N(-0.58,0.17)+0.5N(0.32,0.13)$$

和

$$P(x|C_1)=0.6N(-0.10,0.15)+0.4N(0.65,0.09)$$

生成。这里，多项式的阶数 m 取值为 1～9，将其作为模型选择的候选变量。

表 8 最优模型的次数

Table 8 Number of the best model

样本	1	2	3	4	5	6	7	8	9
100	0	0	155	158	165	151	150	118	103
1000	0	0	90	146	167	174	160	141	122

从表 8 可以看到，如果直接由错误率来进行模型选择，则选到的最优模型在两种样本量下分别为阶数为 5 和 6 的多项式，然而，无论是哪种样本量情形，选取的 3，4，5，6，7 阶多项式的次数都很相近，虽然 5 阶(6 阶)是最优的，但其极有可能

是由方差的波动导致的。进一步,通过使用组块 3×2 交叉验证 t 检验也验证了这一点,3,4,5,6,7 阶多项式之间确实是没有统计显著性差异的,从而应该选择低阶的 3 阶多项式。由此可见,仅使用性能度量指标估计本身来进行模型选择,有可能导致选择到错误的模型。

例 4[25] 考虑两类分类问题,数据 $Z=(X,Y)$,$P(Y=0)=P(Y=1)=1/2$,$X|Y=0\sim N(\mu_0,\Sigma_0)$,$X|Y=1\sim N(\mu_1,\Sigma_1)$,其中 $\mu_0=0_5$,$\Sigma_0=I_5$,$\mu_1=\beta_1 1_5$,$\Sigma_1=\beta_2\Sigma_0$,对照两个机器学习中常用分类器分类树和线性判别分类器的性能。

在$(\beta_1,\beta_2)=(1,2)$情形下,分类树的准确率和召回率分别为 0.768 和 0.758,而线性判别分类器的准确率和召回率分别为 0.820 和 0.800,如果仅考虑单点性能度量,则线性判别分类器的性能显著优于分类树分类器,但是实际上它们的方差波动很大。Wang 等[25] 于 2016 年给出的分类树的准确率和召回率的置信区间分别为(0.689,0.846)和(0.578,0.836),显然线性判别分类器的准确率和召回率落入了分类树的置信区间中,在准确率和召回率性能度量下这两个分类器在统计上是没有显著差异的,单点的性能度量结果可能导致错误的结论。

5.5 一致性分析

很多文献都提到这样一个事实,对于给定的一个数据集,依据于某个性能度量,算法 A 在这个数据上得到了最好的性能,然而,在另一个度量指标下,该算法的性能并非是最好的。例如,Huang 等[54] 于 2003 年指出朴素 Bayes 和决策树算法具有非常相似的预测精度,然而,在他们的另一篇文献[41] 中,他们却说在 AUC 度量下朴素 Bayes 的性能显著优于决策树。因此,为了检验多种算法在不同性能度量下是否具有一致性,本文在 UCI 数据库中 5 个常用的数据集 Bupa,Ecoli,Flag,Glass 和 Ionosphere 上,基于 3 个广泛使用的分类算法(线性判别分析、C4.5 决策树和最近邻)进行了实验分析[55-58]。

表 9 6 种性能度量的一致性对照

Table 9 Comparison of consistency for six performance measures

	LDA					C4.5					NN				
	Bupa	Ecoli	Flag	Glass	Iono.	Bupa	Ecoli	Flag	Glass	Iono.	Bupa	Ecoli	Flag	Glass	Iono.
5CV	0.326	0.042	0.147	0.365	0.134	0.351	0.050	0.208	0.447	0.112	0.361	0.043	0.396	0.201	0.133
Boots.	0.319	0.043	0.113	0.336	0.113	0.330	0.038	0.127	0.326	0.079	0.280	0.046	0.306	0.153	0.106
P	0.688	0.971	0.884	0.580	0.829	0.657	0.956	0.808	0.475	0.855	0.709	0.968	0.664	0.751	0.833
R	0.915	0.961	0.859	0.545	0.977	0.969	0.965	0.844	0.388	0.979	0.759	0.959	0.648	0.797	0.973
F	0.786	0.966	0.871	0.561	0.897	0.783	0.961	0.824	0.414	0.913	0.733	0.964	0.655	0.773	0.897
AUC	0.600	0.954	0.812	0.669	0.852	0.658	0.939	0.809	0.743	0.891	0.658	0.957	0.595	0.782	0.811

表 9 中,5CV,Boots.,P,R,F 和 AUC 分别表示基于五折交叉验证的错误率估计度量、基于 Bootstrap 的错误率估计度量、准确率、召回率、F 度量和 AUC 度量。LDA,C4.5 和 NN 分别表示线性分类算法、C4.5 树算法和最近邻算法。

从表 9 可以看到,当仅使用错误率的估计进行算法性能度量时,在 Flag,Glass 和 Ionosphere 数据集上,错误率的五折交叉验证估计和 Bootstrap 估计的度量结果是一致的,但是在 Bupa 和 Ecoli 数据集上这两个度量得到的结果是不一致的。例如,在 Bupa 数据集上,五折交叉验证估计度量下的最优算法是线性判别分类器,但是在 Bootstrap 估计下最优的算法是最近邻。

对于准确率、召回率和 F 度量,它们的值越大,算法的性能就越高。同样,在 Flag,Glass 和 Ionosphere 数据集上,准确率、召回率和 F 度量具有最优算法(模型)选择的一致性,它们得到的最优算法分别是线性判别分类器、最近邻和 C4.5。然而,在 Bupa 数据集上,具有最大准确率的是最近邻方法,具有最大召回率的却是 C4.5,二者的调和平均最大的(即 F 值最大)又是线性判别分类器。

对于 AUC 度量,其值越大越好。在数据集 Bupa,Ecoli 和 Glass 上选择的最优算法都是最近邻,在 Flag 上选择的是线性判别分类器,在 Ionosphere 上选择的是 C4.5。

表 10 列出了在 LDA vs C4.5,LDA vs NN 和 C4.5 vs NN 两两算法一致性对照的结果。大多数情形下,基于统计检验的各种性能度量具有较好的一致性。实际上,基于统计显著性检验的度量可以为只基于错误率的度量的可信度提供确信保证。例如,在 Flag 数据集上,在基于错误率的五折交叉验证估计度量下,最近邻算法的分类性能优于 C4.5 方法,但是若使用五折交叉验证 t 检验方法进行分析,则二者之间没有统计显著性差异,因为它们的 p 值为 0.048。这是因为,虽然 NN 的五折交叉验证的错误率 0.396 大于 C4.5 的五折交叉验证的错误率 0.208,但是它们的标准差也较大,达到了 0.1。

表 10 基于统计检验的度量的一致性对照

Table 10 Comparison of consistency for statistical test based measures

	LDA vs C4.5					LDA vs NN					C4.5 vs NN				
	Bupa	Ecoli	Flag	Glass	Iono.	Bupa	Ecoli	Flag	Glass	Iono.	Bupa	Ecoli	Flag	Glass	Iono.
5CVt	0.573	0.769	0.773	0.421	0.592	0.577	0.774	0.052	0.124	0.792	0.737	0.777	0.048	0.082	0.595
52CVt	0.575	0.531	0.811	0.609	0.707	0.652	0.873	0.040	0.271	0.845	0.837	0.607	0.040	0.199	0.686
32CVt	0.487	0.363	0.626	0.415	0.595	0.537	0.757	0.011	0.099	0.694	0.775	0.411	0.012	0.080	0.547
Mcne.	0.064	0.471	0.181	0.202	0.420	0.003	0.805	0.689	0.336	0.714	0.067	0.478	0.272	0.167	0.543

因此,在进行算法性能度量时,应根据实际的任务需求来选择适用的性能度量指标,且应该综合分析多种性能度量指标的结果,给出一个相对准确的结论。如果可能,应该进行算法性能对照的统计显著性检验。

结束语 在机器学习的实际应用中,算法性能的评价尤为重要,因为它可以用于指导整个学习算法的选择与应用过

程，为最终学习算法或模型的选定提供确信保证。为此，本文讨论了3类常用的分类学习算法性能度量指标的优缺点和适用性，并进行了实验对照与分析。

分析结果表明，不同的性能度量指标都有它所适用的任务，但也存在着相应的问题，在实际应用中，我们应根据实际的任务需求来选择适用的性能度量指标。它们也有很多的地方需要进行完善和深入的研究：1）提供接近于真实错误率的简单准确估计方法一直是错误率度量研究努力的方向；2）不仅需要提供分类学习算法的AUC度量，还需要提供它的方差或置信区间，以给出更加准确的算法性能度量结果；3）各统计检验中的方差估计问题、划分重复问题等都是特别需要深入探讨的课题；4）如何综合各类性能度量指标，给出一个更加准确、全面的算法性能评价结果，将是一个非常值得研究的问题。

参 考 文 献

[1] ZHANG X G. Pattern Recognition[M]. Beijing: Tsinghua University Press, 2010.

[2] BIAN Z Q. Pattern Recognition[M]. Beijing: Tsinghua University Press, 1988.

[3] DUDA R O, HART P E, STORK D G. Pattern Classification [M]. New York: Springer, 2001.

[4] HASTIE T, TIBSHRANI R, FRIEDMAN J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [M]. New York: Springer, 2001.

[5] VAPNIK V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1999.

[6] FERRI C, HERNANDEZ-ORALLO J, MODROIU R. An Experimental Comparison of Performance Measures for Classification[J]. Pattern Recognition Letters, 2009, 30(1): 27-38.

[7] SOKOLOVA M, LAPALME G. A Systematic Analysis of Performance Measures for Classification Tasks [J]. Information Processing & Management, 2009, 45(4): 427-437.

[8] WEBB A R, COPSEY K D. Introduction to Statistical Pattern Recognition [M]. Academic Press, 1972: 2133-2143.

[9] TURNER K, GHOSH J. Estimating the Bayes Error Rate through Classifier Combining[C]// International Conference on Pattern Recognition. IEEE Computer Society, 1996: 695-699.

[10] BREIMAN L. The Little Bootstrap and other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error [J]. Journal of the American Statistical Association, 1992, 87(419): 738-754.

[11] SHAO J. Bootstrap Model Selection[J]. Publications of the American Statistical Association, 1996, 91(434): 655-665.

[12] LOPES M E, WANG S, MAHONEY M W. A Bootstrap Method for Error Estimation in Randomized Matrix Multiplication [J]. Journal of Machine Learning Research, 2019, 20: 1-40.

[13] BRADLEY E. Prediction, Estimation, and Attribution[J]. Journal of the American Statistical Association, 2020, 115(530): 636-655.

[14] YILDIZ O T, ÖZLEM A, AIPAYDIN E. Multivariate Statistical Tests for Comparing Classification Algorithms[C]// International Conference on Learning and Intelligent Optimization. Springer-Verlag, 2011: 1-15.

[15] GOUTTE C, GAUSSIER E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation [J]. International Journal of Radiation Biology & Related Studies in Physics Chemistry & Medicine, 2005, 51(5): 345-359.

[16] POWERS D M W. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation [J]. Journal of Machine Learning Technology, 2011, 2: 37-63.

[17] WANG Y, LI J H, LI Y F, et al. Confidence Interval for F1 Measure of Algorithm Performance based on Blocked 3×2 Cross-validation [J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(3): 651-659.

[18] MUSCHELLI J. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric [J]. Journal of Classification, 2020, 37(3): 696-708.

[19] FAWCETT T. An Introduction to ROC Analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[20] FLACH P A. The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics[C]// Machine Learning, Proceedings of the Twentieth International Conference. DBLP, 2003: 194-201.

[21] LOBO J M, JIMENEZ-VALVERDE A, REAL R. AUC: a Misleading Measure of the Performance of Predictive Distribution Models [J]. Global Ecology & Biogeography, 2008, 17(2): 145-151.

[22] DIETTERICH T G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms [J]. Neural Computation, 1998, 10(7): 1895-1923.

[23] YANG L, WANG Y. Analysis of Variance of F1 Measure based on Blocked 3×2 Cross Validation [J]. Journal of Frontiers of Computer Science and Technology, 2016, 10(8): 1176-1183.

[24] WANG Y, WANG R B, JIA H C, et al. Blocked 3×2 Cross-validated t-test for Comparing Supervised Classification Learning Algorithms [J]. Neural Computation, 2014, 26(1): 208-235.

[25] WANG Y, LI J H. Credible Intervals for Precision and Recall Based on a K-Fold Cross-Validated Beta Distribution [J]. Neural Computation, 2016, 28(8): 1694-1722.

[26] BISANI M, NEY H. Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.

[27] LIU Y Q, WANG Y, LI J H. Model Selection Algorithm based on Blocked 3×2 Cross-validated t-test [J]. Journal of Shanxi University of Science & Technology, 2015, 33(1): 179-183.

[28] ZADROZNY B, ELKAN C. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers [C]// Proceedings of the 18th International Conference on Machine Learning (ICML). 2001: 609-616.

[29] CORTES C, MOHRI M. AUC Optimization vs. Error Rate Minimization [C]// Advances in Neural Information Processing Systems 16(NIPS 2003). 2003: 313-320.

[30] ROSSET S. Model Selection via the AUC [C]// Machine Learning Proceedings of the 21st International Conference (ICML). 2004: 89-96.

[31] FLACH P A. The Geometry of ROC Space:Understanding Machine Learning Metrics through ROC Isometrics[C]// Machine Learning,Proceedings of the 20th International Conference (ICML). 2003:194-201.

[32] FUERNKRANZ J,FLACH P A. ROC 'n' Rule Learning-towards a Better Understanding of Covering Algorithms [J]. Mach. Learn. ,2005,58(1):39-77.

[33] BUJA A,STUETZLE W,SHEN Y. Loss Functions for Binary Class Probability Estimation:Structure and Applications [EB/OL]. [2005-11-03]. http://stat. wharton. upenn. edu/~buja/PAPERS/paper-proper-scoring. pdf.

[34] COSTA E P,LORENA A C,CARVALHO A C P L F,et al. A Review of Performance Evaluation Measures for Hierarchical Classifiers[C]// Proceedings of the AAAI 2007 Workshop Evaluation Methods for Machine Learning. 2007.

[35] DEMSAR J. Statistical Comparisons of Classifiers over Multiple Data Sets [J]. Journal of Machine Learning Research,2007,7:1-30.

[36] FERRI C,FLACH P A,HERNANDEZ-ORALLO J. Improving the AUC of Probabilistic Estimation Trees[C]// 14th European Conference on Machine Learning,Proceedings,Lecture Notes in Computer Science(ECML 2003). Springer,2003:121-132.

[37] ARIS F H,WENCESLAO G M. A Comparative Study of Methods for Testing the Equality of Two or More ROC Curves [J]. Comput Stat,2018,33:357-377.

[38] MALACH T,POMENKOVA J. Comparing Classifier's Performance Based on Confidence Interval of the ROC [J]. Radioengineering,2018,27(3):827-834.

[39] DAVIS J,GOADRICH M. The Relationship between Precision-recall and ROC Curves[C]// Proceedings of the 23rd International Conference on Machine Learning(ICML'06). 2006:233-240.

[40] WU S,FLACH P A,FERRI C. An Improved Model Selection Heuristic for AUC[C]// 18th European Conference on Machine Learning. 2007:478-489.

[41] HUANG J,LING C X. Using AUC and Accuracy in Evaluating Learning Algorithms[J]. IEEE Trans. Knowl. Data Eng. (TKDE),2005,17(3):299-310.

[42] CARUANA R,NICULESCU-MIZIL A. Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria[C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2004:69-78.

[43] FERRI C,FLACH P A,HERNANDEZ-ORALLO J ,et al. Modifying ROC Curves to Incorporate Predicted Probabilities [C]// Second Workshop on ROC Analysis in ML. 2004:33-40.

[44] BENGIO Y,GR Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation [J]. Journal of Machine Learning Research,2004,5:1089-1105.

[45] MARKATOU M,TIAN H,BISWAS S,et al. Analysis of Variance of Cross-Validation Estimators of the Generalization Error[J]. Journal of Machine Learning Research, 2005, 6(1): 1127-1168.

[46] MORENOTORRES J G,SAEZ J A,HERRERA F. Study on the Impact of Partition-induced Dataset Shift on k-fold Cross-validation [J]. IEEE Transactions on Neural Networks & Learning Systems,2012,23(8):1304-1312.

[47] AKAIKE H. Information Theory and an Extension of the Maximum Likelihood Principle[M]// Breakthroughs in Statistics. New York:Springer,1992:610-624.

[48] SCHWARZ G. Estimating the Dimension of a Model [J]. Annals of Statistics,1978,6(2):15-18.

[49] FAWCETT T. Using Rule Sets to Maximize ROC Performance [C]// IEEE International Conference on Data Mining. IEEE Computer Society,2001:131-138.

[50] HAND D J,TILL R J. A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems [J]. Machine Learning,2001,45(2):171-186.

[51] LOPEZ V,FERNANDEZ A,HERRERA F. On the Importance of the Validation Technique for Classification with Imbalanced Datasets:Addressing Covariate Shift When Data is Skewed [J]. Information Sciences,2014,257(2):1-13.

[52] EVERITT B S. The Analysis of Contingency Tables [M]. London:Chapman and Hall,1977.

[53] WANG R,WANG Y,LI J,et al. Block-Regularized m×2 Cross-validated Estimator of the Generalization Error [J]. Neural Computation,2017,29(2):519-554.

[54] HUANG J,LU J,LING C X. Comparing Naive Bayes,Decision Trees,and SVM with AUC and Accuracy[C]// Proceedings of the Third IEEE International Conference on Data Mining (ICDM). IEEE Computer Society,2003:553-556.

[55] DUA D,KARRA TANISKIDOU E. UCI Machine Learning Repository[EB/OL]. Irvine,CA: University of California,School of Information and Computer Science,2017. http://archive. ics. uci. edu/ml/index. php.

[56] PAUL H,KENTA N. A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins [J]. Intelligent Systems in Molecular Biology,1996,4:109-115.

[57] EVETT I W,SPIEHLER E J. Rule Induction in Forensic Science [J]. Knowledge Based Systems,1989,1:152-160.

[58] SIGILLITO V G,WING S P,HUTTON L V,et al. Classification of Radar Returns from the Ionosphere Using Neural Networks [J]. Johns Hopkins APL Technical Digest,1989,10:262-266.

YANG Xing-li, born in 1986, Ph.D candidate, lecturer, is a member of China Computer Federation. Her main research interest includes statistical machine learning.