炼数成金

标题: 分类指标准确率(Precision)和正确率(Accuracy)的区别

作者: cowryzhao **时间:** 2016-10-19 01:06
标题: 分类指标准确率(Precision)和正确率(Accuracy)的区别

一、引言

分类算法有很多，不同分类算法又用很多不同的变种。不同的分类算法有不同的特定，在不同的数据集上表现的效果也不同，我们需要根据特定的任务进行算法的选择，如何选择分类，如何评价一个分类算法的好坏，前面关于决策树的介绍，我们主要用的正确率（accuracy）来评价分类算法。

正确率确实是一个很好很直观的评价指标，但是有时候正确率高并不能代表一个算法就好。比如某个地区某天地震的预测，假设我们有一堆的特征作为地震分类的属性，类别只有两个：0：不发生地震、1：发生地震。一个不加思考的分类器，对每一个测试用例都将类别划分为0，那那么它就可能达到99%的正确率，但真的地震来临时，这个分类器毫无察觉，这个分类带来的损失是巨大的。为什么99%的正确率的分类器却不是我们想要的，因为这里数据分布不均衡，类别1的数据太少，完全错分类别1依然可以达到很高的正确率却忽视了我们关注的东西。接下来详细介绍一下分类算法的评价指标。

二、评价指标

1、几个常用的术语

这里首先介绍几个常见的模型评价术语，现在假设我们的分类目标只有两类，计为正例（positive）和负例（negtive）分别是：

1）True positives(TP): 被正确地划分为正例的个数，即实际为正例且被分类器划分为正例的实例数（样本数）；

2）False positives(FP): 被错误地划分为正例的个数，即实际为负例但被分类器划分为正例的实例数；

3）False negatives(FN):被错误地划分为负例的个数，即实际为正例但被分类器划分为负例的实例数；

4）True negatives(TN): 被正确地划分为负例的个数，即实际为负例且被分类器划分为负例的实例数。

		预测类别		
		Yes	**No**	**总计**
实际类别	**Yes**	**TP**	**FN**	**P（实际为Yes）**
	No	**FP**	**TN**	**N（实际为No）**
	总计	**P′（被分为Yes）**	**N′（被分为No）**	**P+N**

上图是这四个术语的混淆矩阵，我只知道FP叫伪阳率，其他的怎么称呼就不详了。注意P=TP+FN表示实际为正例的样本个数，我曾经误以为实际为正例的样本数应该为TP+FP，这里只要记住True、False描述的是分类器是否判断正确，Positive、Negative是分类器的分类结果。如果正例计为1、负例计为-1，即positive=1、negtive=-1，用1表示True，-1表示False，那么实际的类标=TF*PN，TF为true或false，PN为positive或negtive。例如True positives(TP)的实际类标=1*1=1为正例，False positives(FP)的实际类标=（-1）*1=-1为负例，False negatives(FN)的实际类标=（-1）*（-1）=1为正例，True negatives(TN)的实际类标=1*（-1）=-1为负例。

2、评价指标

1）正确率（accuracy）

正确率是我们最常见的评价指标，accuracy = （TP+TN）/(P+N)，这个很容易理解，就是被分对的样本数除以所有的样本数，通常来说，正确率越高，分类器越好；

2）错误率（error rate)

错误率则与正确率相反，描述被分类器错分的比例，error rate = (FP+FN)/(P+N)，对某一个实例来说，分对与分错是互斥事件，所以accuracy =1 - error rate；

3）灵敏度（sensitive）

sensitive = TP/P，表示的是所有正例中被分对的比例，衡量了分类器对正例的识别能力；

4）特效度（specificity)

specificity = TN/N，表示的是所有负例中被分对的比例，衡量了分类器对负例的识别能力；

5）精度（precision）

精度是精确性的度量，表示被分为正例的示例中实际为正例的比例，precision=TP/（TP+FP）；

6）召回率（recall）

召回率是覆盖面的度量，度量有多个正例被分为正例，recall=TP/(TP+FN)=TP/P=sensitive，可以看到召回率与灵敏度是一样的。

7）其他评价指标

- 计算速度：分类器训练和预测需要的时间；
- 鲁棒性：处理缺失值和异常值的能力；

- 可扩展性：处理大数据集的能力；
- 可解释性：分类器的预测标准的可理解性，像决策树产生的规则就是很容易理解的，而神经网络的一堆参数就不好理解，我们只好把它看成一个黑盒子。

对于某个具体的分类器而言，我们不可能同时提高所有上面介绍的指标，当然，如果一个分类器能正确分对所有的实例，那么各项指标都已经达到最优，但这样的分类器往往不存在。比如我们开头说的地震预测，没有谁能准确预测地震的发生，但我们能容忍一定程度的误报，假设1000次预测中，有5次预测为发现地震，其中一次真的发生了地震，而其他4次为误报，那么正确率从原来的999/1000=99.9%下降到996/1000=99.6，但召回率从0/1=0%上升为1/1=100%，这样虽然谎报了几次地震，但真的地震来临时，我们没有错过，这样的分类器才是我们想要的，在一定正确率的前提下，我们要求分类器的召回率尽可能的高。

http://blog.sciencenet.cn/blog-460603-785098.html
分类是一种重要的数据挖掘算法。分类的目的是构造一个分类函数或分类模型（即分类器），通过分类器将数据对象映射到某一个给定的类别中。分类器的主要评价指标有准确率(Precision)、召回率(Recall)、*Fb*-score、ROC、AOC等。在研究中也有采用Accuracy（正确率）来评价分类器的。但准确率和正确率这两个概念经常有人混了。【没有耐心看下面内容的博友请看最后的结论】
准确率(Precision) 和召回率(Recall)是信息检索领域两个最基本的指标。准确率也称为查准率，召回率也称为查全率。它们的定义如下：
Precision＝系统检索到的相关文件数量/系统检索到的文件总数量
Recall=系统检索到的相关文件数量/系统所有相关文件数量
Fb-score是准确率和召回率的调和平均：*Fb*=[(1+*b2*)*P*R]/（*b2**P+R），比较常用的是*F*1。
在信息检索中，准确率和召回率是互相影响的，虽然两者都高是一种期望的理想情况，然而实际中常常是准确率高、召回率就低，或者召回率低、但准确率高。所以在实际中常常需要根据具体情况做出取舍，例如对一般搜索的情况是在保证召回率的情况下提升准确率，而如果是疾病监测、反垃圾邮件等，则是在保证准确率的条件下，提升召回率。但有时候，需要兼顾两者，那么就可以用*F*-score指标。
ROC和AUC是评价分类器的指标。ROC是受试者工作特征曲线 receiver operating characteristic curve）的简写，又称为感受性曲线(sensitivity curve)。得此名的原因在于曲线上各点反映着相同的感受性，它们都是对同一信号刺激的反应，只不过是在几种不同的判定标准下所得的结果而已[1]。ROC是反映敏感性和特异性连续变量的综合指标,是用构图法揭示敏感性和特异性的相互关系，它通过将连续变量设定出多个不同的临界值，从而计算出一系列敏感性和特异性，再以敏感性为纵坐标、（1-特异性）为横坐标绘制成曲线。AUC是ROC曲线下面积(Area Under roc Curve)的简称，顾名思义，AUC的值就是处于ROC curve下方的那部分面积的大小。通常，AUC的值介于0.5到1.0之间，AUC越大，诊断准确性越高。在ROC曲线上，最靠近坐标图左上方的点为敏感性和特异性均较高的临界值。
为了解释ROC的概念，让我们考虑一个二分类问题，即将实例分成正类（positive）或负类（negative）。对一个二分问题来说，会出现四种情况。如果一个实例是正类并且也被预测成正类，即为真正类（True positive）,如果实例是负类被预测成正类，称之为假正类（False positive）。相应地，如果实例是负类被预测成负类，称之为真负类（Truenegative）,正类被预测成负类则为假负类（falsenegative）。列联表或混淆矩阵如下表所示，1代表正类，0代表负类。

		实际	
		1	0
预测	1	True Positive (TP) 真正	False Positive (FP) 假正
	0	False Negative (FN) 假负	True Negative TN 真负

基于该列联表，定义敏感性指标为：sensitivity=TP/(TP+FN)。敏感性指标又称为真正类率(truepositive rate ,TPR)，刻画的是分类器所识别出的正实例占所有正实例的比例。
另外定义负正类率(false positive rate, FPR),计算公式为：FPR=FP/(FP+TN).负正类率计算的是分类器错认为正类的负实例占所有负实例的比例
定义特异性指标为:Specificity=TN/(FP+TN)=1-FPR。特异性指标又称为真负类率（True Negative Rate，TNR）。
我们看，实际上，敏感性指标就是召回率，特异性指标=1-FPR。
ROC曲线由两个变量绘制。横坐标是1-specificity，即负正类率(FPR)，纵坐标是 Sensitivity，即真正类率(TPR)。
在此基础上，还可以定义正确率(Accuracy)和错误率(Error)。 Accuracy= (TP+TN)/(TP+FP+TN+FN)，Error= (FP+FN)/(TP+FP+TN+FN)。如果把预测为1看作检索结果，则准确率Precision= TP/(TP+FP)。
结论：
分类正确率（Accuracy），不管是哪个类别，只要预测正确，其数量都放在分子上，而分母是全部数据数量，这说明正确率是对全部数据的判断。而准确率在分类中对应的是某个类别，分子是预测该类别正确的数量，分母是预测为该类别的全部数据的数量。或者说，Accuracy是对分类器整体上的正确率的评价，而Precision是分类器预测为某一个类别的正确率的评价。

https://argcv.com/articles/1036.c

自然语言处理(ML),机器学习(NLP),信息检索(IR)等领域,评估(Evaluation)是一个必要的工作,而其评价指标往往有如下几点:准确率(Accuracy),精确率(Precision),召回率(Recall)和F1-Measure。
本文将简单介绍其中几个概念。中文中这几个评价指标翻译各有不同，所以一般情况下推荐使用英文。
现在我先假定一个具体场景作为例子。

> 假如某个班级有男生**80**人,女生**20**人,共计**100**人.目标是找出所有女生.
> 现在某人挑选出**50**个人,其中**20**人是女生,另外还错误的把30个男生也当作女生挑选出来了.
> 作为评估者的你需要来评估(**evaluation**)下他的工作

首先我们可以计算**准确率(accuracy)**,其定义是: 对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。也就是损失函数是0-1损失时测试数据集上的准确率[1].
这样说听起来有点抽象，简单说就是，前面的场景中，实际情况是那个班级有男的和女的两类，某人(也就是定义中所说的分类器)他又把班级中的人分为男女两类。accuracy需要得到的是此君**分正确的人**占**总人数**的比例。很容易，我们可以得到:他把其中70(20女+50男)人判定正确了,而总人数是100人，所以它的accuracy就是70 %(70 / 100).
由准确率，我们的确可以在一些场合，从某种意义上得到一个分类器是否有效，但它并不总是能有效的评价一个分类器的工作。举个例子,google抓取了argcv 100个页面，而它索引中共有10,000,000个页面,随机抽一个页面，分类下,这是不是argcv的页面呢?如果以accuracy来判断我的工作，那我会把所有的页面都判断为"不是argcv的页面",因为我这样效率非常高(return false,一句话),而accuracy已经到了99.999%(9,999,900/10,000,000),完爆其它很多分类器辛辛苦苦算的值,而我这个算法显然不是需求期待的,那怎么解决呢?这就是precision,recall和f1-measure出场的时间了.
在说precision,recall和f1-measure之前,我们需要先需要定义TP,FN,FP,TN四种分类情况.
按照前面例子,我们需要从一个班级中的人中寻找所有**女生**,如果把这个任务当成一个分类器的话,那么女生就是我们需要的,而男生不是,所以我们称女生为"正类",而男生为"负类".

	相关(Relevant),正类	**无关(NonRelevant),负类**
被检索到(Retrieved)	true positives(TP 正类判定为正类,例子中就是正确的判定"这位是女生")	false positives(FP 负类判定为正类,"存伪",例子中就是分明是男生却判断为女生,当下伪娘横行,这个错常有人犯)
未被检索到(Not Retrieved)	false negatives(FN 正类判定为负类,"去真",例子中就是,分明是女生,这哥们却判断为男生--梁山伯同学犯的错就是这个)	true negatives(TN 负类判定为负类,也就是一个男生被判断为男生,像我这样的纯爷们一准儿就会在此处)

通过这张表,我们可以很容易得到这几个值:
TP=20
FP=30
FN=0
TN=50
精确率(precision)的公式是P=TPTP+FP,它计算的是所有被检索到的item中,"应该被检索到"的item占的比例。
在例子中就是希望知道此君得到的所有人中,正确的人(也就是女生)占有的比例.所以其precision也就是40%(20女生/(20女生+30误判为女生的男生)).
召回率(recall)的公式是R=TPTP+FN,它计算的是所有检索到的item占所有"应该检索到的item"的比例。
在例子中就是希望知道此君得到的女生占本班中所有女生的比例,所以其recall也就是100%(20女生/(20女生+ 0 误判为男生的女生))
F1值就是精确值和召回率的调和均值,也就是
2F1=1P+1R
调整下也就是
F1=2PRP+R=2TP2TP+FP+FN
例子中 F1-measure 也就是约为 57.143%(2∗0.4∗10.4+1).
需要说明的是,有人[2]列了这样个公式
Fa=(a2+1)PRa2(P+R)
将F-measure一般化.
F1-measure认为精确率和召回率的权重是一样的,但有些场景下,我们可能认为精确率会更加重要,调整参数a,使用Fa-measure可以帮助我们更好的evaluate结果.
话虽然很多,其实实现非常轻松,点击此处可以看到我的一个简单的实现.

References

[1] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012.
[2] 准确率（Precision）、召回率（Recall）以及综合评价指标（F1-Measure）

作者: Miko_zhang **时间:** 2016-10-20 11:18
好详细呀！~多谢楼主分享，学习啦

作者: snailpt **时间:** 2016-10-20 20:09
不知道spark里有没有现成的函数可以直接用

作者: lgj573 **时间:** 2016-10-22 21:32
谢谢分享 学习了

作者: 隔壁老唐 **时间:** 2016-10-23 11:41
学习了，谢谢分享 ，很详细

作者: luotitan **时间:** 2016-10-23 21:38
不错的分享，学习了，谢谢！
