软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(1):109−120 [doi: 10.3724/SP.J.1001.2013.04230]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于分辨粒度的 gROC 曲线分析方法*

董元方[1,2], 李雄飞[1], 李 军[1,3], 赵海英[4]

[1](符号计算与知识工程教育部重点实验室(吉林大学),吉林 长春 130012)

[2](长春理工大学 经济管理学院,吉林 长春 130022)

[3](长春理工大学 应用数学系,吉林 长春 130022)

[4](北京邮电大学 世纪学院,北京 100083)

通讯作者: 李雄飞, E-mail: lxf@jlu.edu.cn, http://cs.jlu.edu.cn

摘 要: ROC 曲线是模型选择的一种重要方法,但 ROC 曲线的不确定性影响了模型选择的准确性.基于分辨粒度,从反映得分的不确定性的角度提出 gROC 和 gAUC 的概念,从理论上讨论了 gROC 的若干性质.在给出其算法之后,利用双正态模型检验了 gROC 的合理性.在此基础上,提出了两个模型选择度量——λAUC 和ρAUC,并在 UCI 数据集上验证了该模型选择度量的高效性.实验结果表明,gROC 能够有效反映 ROC 曲线的不确定性,基于λAUC 和ρAUC 的模型选择方法优于基于 AUC 或 sAUC 的模型选择方法,在某些情况下,gROC 具有更强的对分类器性能的比较能力.

关键词: 机器学习;模型选择;分类;ROC 曲线;粒度

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 董元方,李雄飞,李军,赵海英.基于分辨粒度的 gROC 曲线分析方法.软件学报,2013,24(1):109−120. http://www.jos.org.cn/1000-9825/4230.htm

英文引用格式: Dong YF, Li XF, Li J, Zhao HY. gROC curve analysis method based on discernible granularity. Ruanjian Xuebao/Journal of Software, 2013,24(1):109−120 (in Chinese). http://www.jos.org.cn/1000-9825/4230.htm

gROC Curve Analysis Method Based on Discernible Granularity

DONG Yuan-Fang[1,2], LI Xiong-Fei[1], LI Jun[1,3], ZHAO Hai-Ying[4]

[1](Key Laboratory of Symbolic Computation and Knowledge Engineering for Ministry of Education (Jilin University), Changchun 130012, China)

[2](School of Economics and Management, Changchun University of Science and Technology, Changchun 130022, China)

[3](Department of Applied Mathematics, Changchun University of Science and Technology, Changchun 130022, China)

[4](Century College, Beijing University of Posts and Telecommunications, Beijing 100083, China)

Corresponding author: LI Xiong-Fei, E-mail: lxf@jlu.edu.cn, http://cs.jlu.edu.cn

Abstract: ROC Curve is an important method of model selection, but its uncertainty affects the accuracy of model selection. Based on discernible granularity and the view of reflecting the score's uncertainty, the study proposes the concept of gROC and gAUC, and discusses, theoretically, some properties of the gROC. The study also tests the reasonableness of gROC using binormal model after gave its algorithm. On this basis, the paper also proposes two model selection measures, λAUC and ρAUC. The effieciency of these measures is verified based on UCI data sets. Experimental results show that the gROC can effectively reflect the uncertainty of ROC curve, and the model selection methods based on λAUC and ρAUC are better than the method based on AUC or sAUC. In some cases, gROC has stronger capability on comparison of classifiers performance.

Key words: machine learning; model selection; classification; ROC curve; granularity

* 基金项目: 国家自然科学基金(60863010, 61163044); 国家重点基础研究发展计划(973)(2010CB334709); 吉林省科技发展计划(20090704)

收稿时间: 2011-05-19; 定稿时间: 2012-03-19

近年来,在机器学习研究中,分类器性能评估方法得到了广泛关注,模型选择度量的相关研究逐渐形成了新的研究热点.ROC 曲线(receiver operating characteristics curve,接收者操作特征曲线)分析是可视化地评估分类器性能,从而进行模型选择的方法[1].ROC 分析最早出现在信号探测研究中,其目的是描述击中率和误警率之间的权衡[2].1989 年,Spackman 将 ROC 分析引入到机器学习领域[3],用于评估和比较算法.2000 年,Swets 在《Scientific American》上发表的文章,引起了科学界对 ROC 分析的广泛关注[4].2004 年、2005 年和 2006 年,国际机器学习会议(Int'l Conf. on Machine Learning,简称 ICML)专门为 ROC 分析开辟主题讨论.欧洲人工智能会议(the European Conf. on Artificial Intelligence,简称 ECAI)在 2004 年也曾为 ROC 分析设立讨论专题.由于 ROC 分析具有直观、易懂、使用简单等优点,因而被广泛应用于医疗诊断[5]、数据挖掘、模式识别[6]和其他自然科学领域.但是,由于 AUC(area under the ROC curve,ROC 曲线下方面积)只是考虑了得分序列的序,没有考虑得分间"间隔"的绝对值,学者们针对 ROC 分析方法进行了很多改进和发展,并在其基础上提出了一些新的分析方法.

图案基因相关研究中,模型选择和对得分分类器的评估方法是实现高效织物基因检索的关键.而采用真实的织物基因数据训练得分分类器,分类器的分类结果存在得分排序不均匀现象,现有 ROC 方法不能很好地评估这样的分类器的性能.为此,提出基于分辨粒度的 gROC 方法.该方法不依赖采样技术,在获得 ROC 曲线的同时,能够依据得分之间内在的"粒度"特征,获得两条上下近似 ROC 曲线,真实 ROC 曲线以大概率落在上下近似 ROC 曲线之间.实验结果表明,在一定"粒度"下,基于该方法的分类器模型选择,能够更有效地区分分类器性能的优劣.

本文第 1 节介绍相关工作.第 2 节给出 ROC 分析的基本概念,并提出上下近似 ROC 和上下近似 AUC 的概念,以及模型选择度量.第 3 节给出 gROC 曲线分析相关度量的生成算法.第 4 节给出实验研究结果.第 5 节总结全文.

1 相关工作

分类算法的目的是通过训练创建分类器,使得所创建的分类器能够预测未知类别标签的数据的类别.准确率或错误率是最典型的评估分类算法预测性能的度量.朴素贝叶斯等分类器能够给出类别预测的置信度或概率估计,但是因为不考虑预测的概率而只关心类别,准确率会忽略这些信息.另外,当数据集不平衡时,准确率也不再是一个合适的分类器评估度量.

在许多数据挖掘的应用中,排序比分类更有实际意义.例如:在分析消费者购买意向时,对其有可能购买的商品按照可能性进行排序,比简单区分购买或不购买更有实际意义;在信息检索领域,通常用户只会关心最感兴趣的检索结果,因此需要对检索结果按照相关程度进行排序.贝叶斯分类、神经网络等算法能够给出样例分类的概率估计,是常见的排序算法.为了获得较优的排序算法,需要对各种算法进行评估,ROC 是最常用的机器学习算法性能评估工具.Bradley 使用 AUC 比较常见的机器学习算法[7],发现 AUC 比精确度更有优势.例如:AUC 增加了方差分析的灵敏度;AUC 直接度量排序优劣,使用 AUC 作为模型选择度量时,不仅保证较优的 AUC,而且保证较优的精确度.Huang 和 Ling 从理论和实验两方面论述了作为模型评估度量,AUC 优于精确度[8],并进一步构建了一种新的更有效的性能评估度量 AUC:acc.通过相关分析证明,这种度量与 RMS(root mean square error)的相关度更大,基于该度量构建分类器可以得到更好的预测性能[9].万柏坤等人应用 ROC 曲线进行人工神经网络参数优化和支持向量机性能比较,实验结果表明,ROC 曲线是优选特征参数及分类阈值的有效工具[10].

由于 ROC 分析只关注序而不考虑得分,一些学者分别提出了改进方法.Castanho 等人将模糊集合论与 ROC 组合,提出一种模糊 ROC 方法,模糊 ROC 分析用于评估基于模糊规则的系统性能[11].Ferri 等人指出,AUC 忽略了概率值而只关注序,并基于此分析提出 AUC 的概率版本 pAUC.pAUC 估计排序的性能,并将得分看成是真实得分的有噪声的观测,考虑了概率因素.该度量评估排序性能时,在考虑序的同时,也考虑概率值的大小[12].Wu 和 Flash 等人基于序和预测得分的原始值提出 sAUC,该度量避免了 AUC 不考虑预测得分的缺点[13,14].sAUC 考虑了得分间隔的绝对值,使得基于 sAUC 的分类器评估更具有鲁棒性.Calders 等人提出了 softAUC,并进一步给出一种有效的通过优化 AUC 构建分类器的方法[15].

之后,Vanderlooy等人给出了pAUC,sAUC,softAUC的一般形式,并通过理论分析和实验比较发现,这些变形并不比 AUC 更出色[16].Hand 指出 AUC 存在的潜在问题:AUC 相当于在依赖于得分分布的代价比分布上,平均误分类损失.由于得分分布依赖于分类器,因此在评估分类器性能时,使用 AUC 相当于使用一个依赖于分类器自身的度量.也就是说,AUC 是使用不同度量来评价不同分类器的[17].

在实际应用中,由于无法获得全部真实完整的数据,得到的 ROC 曲线只是经验 ROC 曲线,并不是真实的ROC 曲线.业已提出很多估计 ROC 曲线以及其不确定性的方法[18],其中最主要的是 ROC 曲线置信带方法,这些方法通常假定数据分布,或者依赖细致的采样方法[19−21].Macskassy 等人将医学界关于 ROC 曲线置信带的讨论引入到机器学习领域,讨论了垂直平均、阈值平均、联合置信区域、定宽带、Working-Hotelling 带等方法[19,20].在类不平衡情形下,ROC 置信带不再可靠,Elazmeh 等人给出了一种不平衡数据下计算 ROC 置信段的方法[21].Efron 和 Tibshirani 提出的 bootstrap 理论[22],经常被应用于生成 ROC 曲线的重采样过程.ROC 置信带在表现上一般呈现锯齿形,在小样本情况下,通常有效区域是不够准确的,特别是在较低真正率对应区域[23].而且,ROC 置信带需要多次重复生成 ROC 曲线才能获得结果,效率不高.

实际上,ROC 曲线的不确定性在得分序列中就有体现,当得分间距过小时,表明样例间可分辨性不够理想.本文提出的 gROC 方法基于分辨粒度的概念定义上下近似 ROC,估计 ROC 曲线的不确定性,可以避免大量的采样,提高了效率.模型选择实验及模型选择度量相关分析实验结果表明,基于 gROC 分析定义的模型选择度量 λAUC 和 ρAUC,优于基于 AUC 或 sAUC 的模型选择方法.

2 gROC 曲线分析方法

2.1 关于ROC和AUC

考虑两类别分类问题,每个样例被映射为正类和负类标签{*p*,*n*}中的一个元素.分类模型(或分类器)将样例映射到预测类别,有 4 种可能输出.假设样例是正例,如果被分为正例,则称其为真实正例;如果被分为负例,则称其为错误负例;假设样例是负例,如果被分为负例,则称其为真实负例;如果被分为正例,则称其为错误正例.给定一个分类器和作为测试集的一组样例,可以用混淆矩阵表示样例的分布情况,见表 1.

Table 1 Confusion matrix

表 1 混淆矩阵

	预测正类	预测负类
实际正类	True positives (TP)	False negatives (FN)
实际负类	False positives (FP)	True negatives (TN)

由混淆矩阵派生两个度量:

- 真正率: $TPrate = \dfrac{TP}{TP + FN}$;
- 假正率: $FPrate = \dfrac{FP}{TN + FP}$.

以 *FPrate* 为横坐标,以 *TPrate* 为纵坐标,所有可能的点(*FPrate*,*TPrate*)形成二维平面区域[0,1]×[0,1],称为ROC 空间.ROC 曲线由若干个点(*FPrate*,*TPrate*)连接形成,每个点对应一个分类器模型.点(0,0)表示把每个样例都预测成负类的模型;点(1,1)表示把每个样例都预测成正类的模型;点(1,0)是理想模型,将所有正例分类为正类,所有负例分类为负类.

接近 *x* 轴的点所代表的分类器相对保守,仅当具有足够依据时才将对象划分到正类;ROC 图右上边的分类器可以认为相对“宽泛”,即使判断依据较弱也会将对象划归到正类中.真实数据中有大量负类样例,所以,ROC 图中靠左边的分类器的性能更有吸引力.

许多分类器,例如决策树或规则集,对每个样例输出 Yes 或 No.当离散分类器应用于测试集时产生单一混淆矩阵,与 ROC 空间的一个点对应.另一些分类器,例如朴素贝叶斯分类器或神经网络,通常对一个样例生成概率

或得分(score),也即,一个表示样例隶属于某个类程度的数值.这些值可以是具有归一化特性的精确的概率,或者一般地,这些值是未经校准的得分,较高得分表示较高可能性,称为概率分类器,尽管实际上输出不一定满足概率归一化.由样例得分形成的序列称为算法生成的排序(rank).

排序或得分分类器可以通过设置阈值转化为离散(二值)分类器:如果分类器输出高于阈值,则分类器输出 Yes;否则,输出 No.每一个阈值对应 ROC 空间的一个点.从概念上说,可以通过从–∞到+∞改变阈值,描绘 ROC 空间中的一条 ROC 曲线.计算上,这是一种产生 ROC 曲线的低效方法.

ROC 曲线能够反映分类器的性能.称 ROC 曲线下方面积值为模型的 AUC,通常利用 AUC 作为衡量分类器性能的度量指标.

2.2 gROC及gAUC模型相关定义

对于模型所给定的得分序列,如果其正例得分 score 和负例得分 score 的间距比较大,说明该模型较好地区分了正例和负例;反之,如果正例得分和负例得分距离较近,实际上,由于分类的“不稳定性”,该分类器区分正例和负例的能力并不理想.为了刻画不同类别样例得分之间的间距所反映的分类器性能,建立了 gROC 和 gAUC 模型.

定义 1. 给定分类器得分值 s 和参数 δ,称 $(s,s+\delta)$ 为 s 的右 δ 邻域,记作 $U(s,\delta)$.

由于得分的不确定性,可以将得分位于定点的 δ 邻域内的样例视为不可分辨的,这些样例均视为同种类别.称邻域半径 δ 为分辨粒度,表示在 δ“粒度”水平上分类器不能明确分辨正例和反例.

定义 2. 将 $U(s,\delta)$ 中的样例均视为正例,计算 s 的假正率和真正率,分别记为 $FPrate^*$ 和 $TPrate^*$,称 $(FPrate^*, TPrate^*)$ 为上近似 ROC 点,记作 upROC(s,δ).

定义 3. 将 $U(s,\delta)$ 中的样例均视为负例,计算 s 的假正率和真正率,分别记为 $FPrate_*$ 和 $TPrate_*$,称 $(FPrate_*, TPrate_*)$ 为下近似 ROC 点,记作 lowROC(s,δ).

定义 4. 点 $(FPrate^*,TPrate^*)$ 的轨迹称为上近似 ROC 曲线,记作 upROC;由点 $(FPrate_*,TPrate_*)$ 构成的轨迹称为下近似 ROC 曲线,记作 lowROC.由 lowROC 和 upROC 形成的曲线对记为 gROC=⟨lowROC,upROC⟩,称为 gROC 曲线对.

定义 5. lowROC 下方面积称为下近似 AUC,记作 lowAUC;upROC 下方面积称为上近似 AUC,记作 upAUC.由 lowAUC 和 upAUC 形成的序偶记为 gAUC=⟨lowAUC,upAUC⟩.

设正例服从正态分布 $N(0.65,0.2^2)$,负例服从正态分布 $N(0.45,0.2^2)$,分别生成 2 000 个样例,即取 $N_p=N_n=2000$,仿真得分分类器见表 2(仅给出序号 2001~2050 的部分样例得分及类别信息).分辨粒度 δ 取为 0.02,ROC 曲线及上下近似 ROC 曲线如图 1 所示,其中,右侧为局部放大图.

Table 2 An example of scoring classifier (partial)

表 2 仿真得分分类器示例(部分样例)

序号	得分	类别	序号	得分	类别	序号	得分	类别	序号	得分	类别	序号	得分	类别
…	…	…	2011	0.483 96	*n*	2022	0.483 11	*p*	2033	0.482 47	*p*	2044	0.481 41	*n*
2001	0.485 71	*n*	2012	0.483 74	*p*	2023	0.483 04	*p*	2034	0.482 32	*n*	2045	0.481 24	*p*
2002	0.485 62	*n*	2013	0.483 72	*n*	2024	0.483 04	*n*	2035	0.482 30	*p*	2046	0.481 21	*n*
2003	0.485 08	*n*	2014	0.483 59	*p*	2025	0.482 98	*p*	2036	0.482 11	*p*	2047	0.481 12	*n*
2004	0.484 71	*n*	2015	0.483 58	*n*	2026	0.482 88	*p*	2037	0.482 07	*n*	2048	0.481 10	*n*
2005	0.484 70	*n*	2016	0.483 50	*n*	2027	0.482 78	*n*	2038	0.481 95	*n*	2049	0.481 09	*p*
2006	0.484 41	*p*	2017	0.483 38	*n*	2028	0.482 72	*p*	2039	0.481 89	*n*	2050	0.480 97	*p*
2007	0.484 19	*n*	2018	0.483 35	*p*	2029	0.482 61	*n*	2040	0.481 83	*n*	…	…	…
2008	0.484 17	*p*	2019	0.483 33	*n*	2030	0.482 58	*p*	2041	0.481 82	*p*			
2009	0.484 16	*n*	2020	0.483 30	*p*	2031	0.482 53	*n*	2042	0.481 68	*p*			
2010	0.484 14	*p*	2021	0.483 17	*n*	2032	0.482 52	*p*	2043	0.481 68	*p*			

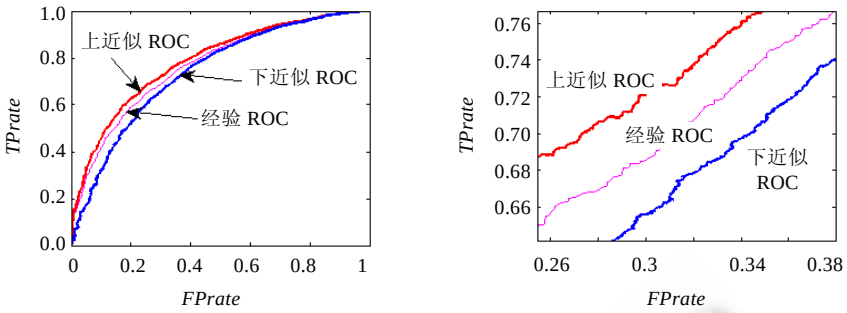


Fig.1 ROC curve and pair of the gROC curves

图 1 ROC 曲线和 gROC 曲线对

2.3 性质分析

考虑特定的类别,假设该类别样例总数为 N,样例分布密度函数为 $\varphi(x)$,分布函数为 $\Phi(x)$,令 ξ 表示 $U(s,\delta)$ 内的该类样例个数,则样例 $x\in U(s,\delta)$ 的概率为 $h(\delta)=\int_s^{s+\delta}\varphi(x)\mathrm{d}x=\Phi(s+\delta)-\Phi(s)$.由于随机变量 ξ 的取值服从二项分布,故有 $P\{\xi=k\}=C_N^k h^k(\delta)(1-h(\delta))^{N-k}$,其中,$k$=0,1,…,$N$.

随机变量 ξ 的数学期望为 $E(\xi)=Nh(\delta)=N(\Phi(s+\delta)-\Phi(s))$.上述讨论对于正例或负例均成立.

设 $U(s,\delta)$ 内正类样例数目为 ξ_δ^p,负类样例数目为 ξ_δ^n,可得如下结论:

性质 1. 对于给定的得分值 s 以及给定的分辨粒度 δ,upROC(s,δ)在 ROC 点的左上方;lowROC(s,δ)在 ROC 点的右下方.

证明:对于上近似 ROC 点有 $FPrate^*=\dfrac{FP-\xi_\delta^n}{FP-\xi_\delta^n+TN}$,$TPrate^*=\dfrac{TP+\xi_\delta^n}{TP+\xi_\delta^n+FN}$.

显然,$FPrate^*\leqslant FPrate$,$TPrate^*\geqslant Tprate$.因此,upROC(s,δ)位于 ROC 相应点的左上方.

同理有 $FPrate_*=\dfrac{FP+\xi_\delta^p}{FP+\xi_\delta^p+TN}$,$TPrate_*=\dfrac{TP-\xi_\delta^p}{TP-\xi_\delta^p+FN}$,$FPrate_*\geqslant FPrate$,$TPrate_*\leqslant Tprate$.

因此,lowROC(s,δ)位于 ROC 相应点的右下方. □

性质 2. 如果 $U(s,\delta)$内没有其他样例,则 upROC(s,δ),lowROC(s,δ)与 ROC 点重合.

证明:因为 $U(s,\delta)$内没有其他样例,则 $\xi_\delta^p=0,\xi_\delta^n=0$.所以,

$$FPrate^*=\frac{FP-\xi_\delta^n}{FP-\xi_\delta^n+TN}=FPrate,$$

$$TPrate^*=\frac{TP+\xi_\delta^n}{TP+\xi_\delta^n+FN}=TPrate,$$

$$FPrate_*=\frac{FP+\xi_\delta^p}{FP+\xi_\delta^p+TN}=FPrate,$$

$$TPrate_*=\frac{TP-\xi_\delta^p}{TP-\xi_\delta^p+FN}=TPrate.$$

因此,upROC(s,δ),lowROC(s,δ)与 ROC 点重合. □

性质 3. 当分辨粒度 δ 小于得分的最小间距时,upROC,lowROC 与 ROC 重合.

证明:由于 δ 小于得分的最小间距,故对于每个得分 s,$U(s,\delta)$内没有其他样例,由性质 2 可知结论成立. □

性质 4. 对于给定的得分值 s 以及两个分辨粒度 δ_1 和 δ_2,设 $\delta_1<\delta_2$,则 upROC(s,δ_2)位于 upROC(s,δ_1)之左上;lowROC(s,δ_2)位于 lowROC(s,δ_1)之右下.

证明:由于 $\delta_1<\delta_2$,故 $\xi_{\delta_2}^n\geqslant\xi_{\delta_1}^n$,$\xi_{\delta_2}^p\geqslant\xi_{\delta_1}^p$.从而有,

$$FPrate^*(s,\delta_2)=\frac{FP-\xi^n_{\delta_2}}{FP-\xi^n_{\delta_2}+TN}\leqslant\frac{FP-\xi^n_{\delta_1}}{FP-\xi^n_{\delta_1}+TN}=FPrate^*(s,\delta_1),$$

$$TPrate^*(s,\delta_2)=\frac{TP+\xi^n_{\delta_2}}{TP+\xi^n_{\delta_2}+FN}\geqslant\frac{TP+\xi^n_{\delta_1}}{TP+\xi^n_{\delta_1}+FN}=TPrate^*(s,\delta_1).$$

因此,$(FPrate^*(s,\delta_2),TPrate^*(s,\delta_2))$位于$(FPrate^*(s,\delta_1),TPrate^*(s,\delta_1))$左上方.

同理可证,$(FPrate_*(s,\delta_2),TPrate_*(s,\delta_2))$位于$(FPrate_*(s,\delta_1),TPrate_*(s,\delta_1))$右下方.

也即,upROC(s,δ_2)位于 upROC(s,δ_1)之左上;lowROC(s,δ_2)位于 lowROC(s,δ_1)之右下. □

由性质 4 可知,与δ_2对应的 upROC 曲线位于与δ_1对应的 upROC 曲线的上方;与δ_2对应的 lowROC 曲线位于与δ_1对应的 lowROC 曲线的下方.从而,upAUC 是关于δ的单调递增函数,lowAUC 是关于δ的单调递减函数.

2.4 基于gROC的模型选择

在利用 AUC 评价两个分类器性能时,只利用了 rank,而没有利用 score,因此不能很好地比较两个分类器的性能.表 3 的例子揭示出 AUC 的这一弱点,由于分类器 *A* 和分类器 *B* 对样例的排序完全相同,故其 ROC 曲线一样,从而 AUC 相同,利用 ROC 或 AUC 无法比较两个分类器的性能.但在 gROC 模型下,利用 gAUC 能够比较两个分类器的优劣.

Table 3 An illustrating example on AUC's shortcoming

表 3 说明 AUC 缺点的示例

得分分类器*A*						得分分类器*B*					
序号	得分	类别	序号	得分	类别	序号	得分	类别	序号	得分	类别
1	0.9	*p*	11	0.495	*n*	1	0.9	*p*	11	0.4	*n*
2	0.8	*p*	12	0.40	*n*	2	0.8	*p*	12	0.38	*n*
3	0.7	*n*	13	0.38	*p*	3	0.7	*n*	13	0.36	*p*
4	0.6	*p*	14	0.35	*n*	4	0.6	*p*	14	0.34	*n*
5	0.55	*p*	15	0.34	*n*	5	0.57	*p*	15	0.32	*n*
6	0.54	*p*	16	0.33	*n*	6	0.55	*p*	16	0.30	*n*
7	0.53	*n*	17	0.32	*p*	7	0.53	*n*	17	0.25	*p*
8	0.52	*n*	18	0.31	*n*	8	0.51	*n*	18	0.15	*n*
9	0.51	*p*	19	0.20	*p*	9	0.48	*p*	19	0.12	*p*
10	0.505	*p*	20	0.10	*n*	10	0.45	*p*	20	0.1	*n*

给定分辨粒度为δ=0.02,经计算可知,表 3 所示得分分类器 *A* 的 AUC 为 0.69,lowAUC 为 0.645,upAUC 为 0.735.而得分分类器 *B* 的 AUC 也为 0.69,但其 lowAUC 为 0.674,upAUC 为 0.691.根据 lowAUC,upAUC 的比较可知,分类器 *B* 的得分不确定性更小,具有更好的分类性能.

定义 6. 称 $\lambda=\frac{lowAUC}{upAUC}$ 为近似比率,用于刻画 lowROC 和 upROC 逼近 ROC 曲线的程度.

显然,$\lambda\in[0,1]$.当λ=1 时,lowROC 和 upROC 与 ROC 曲线重合.λ 的值越小,说明在给定的粒度之下,模型的 ROC 不确定性越大.

在比较两个模型的优劣时,如果它们的 AUC 值相同,可以利用近似比率 λ 比较模型优劣.在相同分辨粒度的情况下,具有较高近似比率的模型性能较好.一般情况下,两个模型的 AUC 值不一定相等,通常具有较高 AUC 的模型,其分类性能更好.由于分类器性能与 AUC 成正比,并且与近似比率成正比,因此可以定义如下的分类器性能度量.

定义 7. 称 λAUC=$\lambda\times$AUC 为模型的 λ 近似加权 AUC,记为 λAUC.

显然,当两个模型 AUC 相等时,判断两个模型性能的因素取决于近似比率 λ.而当两个模型 AUC 不同时,λAUC 能够反映模型 AUC 值以及 ROC 不确定性两个方面.因此,λAUC 比 AUC 更有效.

另外,如果 lowAUC=upAUC,则 λAUC=AUC,也即上下近似相同时,度量退化为 AUC.

下面考察 upAUC(lowAUC)随分辨粒度δ变化的情况.以分辨粒度δ为横轴,AUC/upAUC(lowAUC/AUC)为纵轴,绘制曲线图(如图 2 所示).

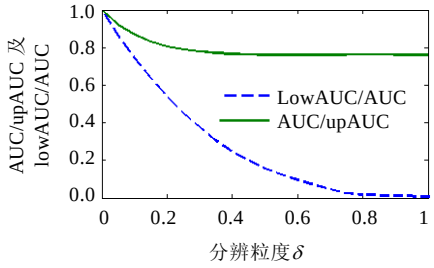


Fig.2 Trend of AUC/upAUC (lowAUC/AUC) along with the discernible granularity δ

图 2 AUC/upAUC(lowAUC/AUC)随分辨粒度δ的变化趋势

由性质 3 和性质 4 可知,随着分辨粒度δ由 0 变到 1,upAUC 由 AUC 单调上升到 1,而 lowAUC 由 AUC 单调下降到 0.

对于给定的分辨粒度δ,设 $f(\delta)=\frac{AUC}{upAUC}, g(\delta)=\frac{lowAUC}{AUC}$.如果得分序列分布存在局部稠密的情况,则 ROC 曲线的不确定性会更大.在分辨粒度不变的情况下,得分分布越稠密,upAUC 将越大,从而导致 $f(\delta)$越小.因此,$f(\delta)$曲线的位置可以作为反映得分分布不确定因素或 ROC 不确定性的一个度量.$f(\delta)$曲线的位置越靠近上方,说明 ROC 曲线不确定性越小,对应算法的性能越理想.同理,在分辨粒度不变的情况下,得分分布越稠密, lowAUC 将越小,从而导致 $g(\delta)$越小.因此,$g(\delta)$曲线的位置也可以作为反映得分分布不确定因素或 ROC 不确定性的一个度量.$g(\delta)$曲线的位置越靠近上方,说明 ROC 曲线不确定性越小,对应算法的性能越理想.

为了给出单一的度量,利用曲线下方面积来反映曲线位置是否靠近上方.曲线 $f(\delta)$及 $g(\delta)$下方面积分别为 $\int_0^1 f(\delta)\mathrm{d}\delta$ 和 $\int_0^1 g(\delta)\mathrm{d}\delta$. 采用面积值作为量化度量,取两者的几何均值,令 $\rho=\sqrt{\int_0^1 f(\delta)\mathrm{d}\delta\cdot\int_0^1 g(\delta)\mathrm{d}\delta}$, 如果两个模型的 AUC 值相同,则 ρ 值较大的算法具有更好的性能.兼顾 ρ 值及 AUC 值,定义如下度量:

定义 8. 称 ρAUC=ρ×AUC 为模型的 ρ 近似加权 AUC,记为 ρAUC.

近似比率 λ 与 ρ 都反映了 ROC 曲线的不确定性,但其角度不同.近似比率 λ 是分辨粒度δ取特定值时得到的,因此,λ 是在给定分辨粒度的情况下,ROC 不确定性的反映;ρ 值是在分辨粒度δ取所有可能取值的情况下得到的,因此,作为 ROC 不确定性的反映,ρ 值更适合大容量数据集.从而有,ρAUC 更适合作为大容量数据集上分类器模型选择度量.

3 gROC 分析相关度量的计算

根据 gROC 相关度量的定义,upAUC,lowAUC,λAUC,ρAUC 的计算均基于 gROC 曲线.因此,首先根据样例得分计算 gROC 曲线的 upROC 点和 lowROC 点的坐标值,然后基于梯形公式计算 upAUC 和 lowAUC.

算法. gROC 曲线生成.

输入:

- $Z=\{C_i,s(i)\}$;
- C_i:样例 i;
- $s(i)$:预测样例 C_i 为正例的得分;
- δ:分辨粒度.

输出:upROC 点的坐标值($FPrate^*$,$TPrate^*$)和 lowROC 点的坐标值($FPrate_*$,$TPrate_*$).

Begin

1. Z_d←按得分值 s 降序排列的 Z
2. 初始化:$upAUC$←0,$lowAUC$←0,n←0,p←0,ξ_n←0,ξ_p←0
3. for $C_i\in Z_d$ do

4. $n\leftarrow 0, p\leftarrow 0, \xi\leftarrow 0$
5. $left\leftarrow s_i+\delta$
6. $leftno\leftarrow$大于等于 $left$ 的样例个数
7. $p\leftarrow$由 1 到 $leftno$ 中正例个数
8. $n\leftarrow$由 1 到 $leftno$ 中负例个数
9. if $i==leftno+1$ then
10. if C_i 为正例 then $p\leftarrow p+1$
11. else $n\leftarrow n+1$
12. end if
13. $FPrate^*(i)\leftarrow n/N_n$
14. $TPrate^*(i)\leftarrow p/N_p$
15. $FPrate_*(i)\leftarrow n/N_n$
16. $TPrate_*(i)\leftarrow p/N_p$
17. else
18. $p\leftarrow p+(i-leftno)$;
19. for $k\in(leftno+1,...,i)$ do
20. if C_k 为负例 then
21. $\xi_n\leftarrow\xi_n+1$
22. else $\xi_p\leftarrow\xi_p+1$
23. end if
24. end for
25. $FPrate^*(i)\leftarrow n/(N_n-\xi_n)$
26. $TPrate^*(i)\leftarrow p/(N_p+\xi_n)$
27. $FPrate_*(i)\leftarrow n/(N_n+\xi_p)$
28. $TPrate_*(i)\leftarrow p/(N_p-\xi_p)$
29. end if
30. end for

End

基于 $FPrate^*$,$TPrate^*$,$FPrate_*$,$TPrate_*$,利用梯形公式计算可得模型的 upAUC 值和 lowAUC 值,从而可以按照 λAUC 及 ρAUC 定义计算相关度量.其中,λAUC 是在给定单一分辨粒度之下的计算结果;而 ρAUC 的计算需要迭代,令分辨粒度由 0 变化到 1,计算在不同分辨粒度之下的 $f(\delta)$及 $g(\delta)$.同样,利用梯形公式计算 $f(\delta)$及 $g(\delta)$曲线下方面积,然后,根据定义计算 ρAUC.

4 实验分析

4.1 ROC不确定性的分析

在估计 ROC 曲线不确定性的不同方法中,Metz 等人使用的双正态模型被广泛采用[24].在图 1 所示仿真实验的基础上,继续讨论 ROC 不确定性的分析.设正例服从正态分布 $N(0.65,0.2^2)$,负例服从正态分布 $N(0.45,0.2^2)$,分别生成 2 000 个样例,即,取 $N_p=N_n=2000$,分辨粒度取为 $\delta=0.02$.图 3 中标注了在上述参数下的经验 ROC 曲线,中间黑色曲线为真实 ROC 曲线.真实 ROC 曲线介于 lowROC 和 upROC 之间.同时生成 100 次经验 ROC 曲线,见图 3 中浅灰色曲线族(除特别标注曲线外).可以看出,lowROC 和 upROC 可以描述经验 ROC 曲线的不确定性.

经验 ROC 曲线对应的 AUC 值为 0.7628,upAUC 值为 0.7938,lowAUC 值为 0.7441,其近似比率 λ 为 0.9374.

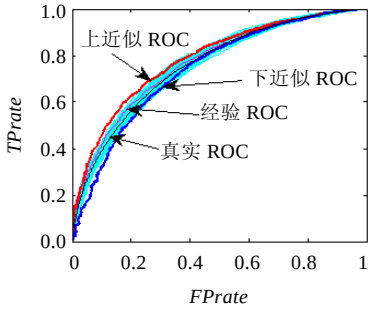


Fig.3 Uncertainty of ROC curve

图 3 ROC 曲线的不确定性

4.2 基于近似加权AUC的模型选择

4.2.1 数据集及实验设置

为评估算法的性能,选择 7 组具有不同实际应用背景的 UCI 数据[25].对于含有多个类别的数据,合并某些类别或只取两个类别.

表 4 是用于实验的数据信息,包括数据集大小、属性信息等.其中,Vowel 的 hed 类作为正类,其他类合并作为负类;Vehicle 的 van 类作为正类,其他类合并作为负类;取 Satimage 的 dampgrey soil 类为正类,其他类合并作为负类;取 Abalone 数据的第 18 类为正类,第 9 类为负类.

Table 4 UCI data sets

表 4 UCI 数据集

数据集	属性		样例数目
	连续	离散	
Monk2	0	6	169
sonar	60	0	208
Breast-W	9	0	699
Abalone	7	1	731
Vehicle	18	0	846
Vowel	10	3	990
satimage	33	0	6 435

为保持在训练集、验证集和测试集中类别分布一致,利用分层抽样将原始数据分成 3 部分:60%用于训练,20%用于验证,20%用于测试.实验基于 Matlab2010a 软件环境,选择朴素贝叶斯分类器.

4.2.2 模型选择度量比较

对作为训练的数据集利用 bootstrap 进行抽样[22],在每一个 bootstrap 抽样上构建朴素贝叶斯分类器.bootstrap 抽样次数取为 100,从而得到 100 个朴素贝叶斯分类器.在分类器验证阶段,比较 gAUC,AUC 和 sAUC 的模型选择能力,分别得到 AUC,sAUC,λAUC 和ρAUC 值最大的 4 个不同分类器.然后,对由不同模型选择度量选定的分类器,利用测试集上的 AUC 值比较分类器是否有显著差异.

表 5 列出了实验结果,表格中的数字为贝叶斯分类器在测试集上的 AUC 值.可以看出,一般情况下,基于 gAUC 的模型选择度量所确定的分类器,其测试集上的 AUC 值比直接使用 AUC 或 sAUC 作为模型选择度量所得分类器的测试集 AUC 值更高.实验结果表明,基于 gAUC 的模型选择度量是有效的.其原因是,λAUC 和ρAUC 兼顾了两个方面:AUC 值以及上下近似 ROC“逼近”真实 ROC 的程度.

另外,从实验结果可以看出,对于大样本数据集,ρAUC 比λAUC 的效果更好.其原因是,ρ值是在所有可能分辨粒度之下,ROC 不确定性的反映;而近似比率λ是在给定分辨粒度下,ROC 不确定性的反映.

Table 5 AUC values on the test set

表 5 测试集上的 AUC 值

数据集	AUC	sAUC	λAUC	ρAUC
Monk2	0.563 5	0.563 7	**0.567 4**	**0.567 4**
sonar	0.890 5	0.890 5	**0.897 1**	0.892 7
Breast-W	0.961 3	0.962 5	**0.965 9**	0.965 8
Abalone	0.857 2	0.857 2	**0.857 2**	**0.857 2**
Vehicle	0.886 0	0.912 6	0.922 9	**0.927 9**
Vowel	0.946 6	0.946 6	0.946 6	**0.947 8**
satimage	0.905 1	0.905 1	0.899 8	**0.907 5**

图 4 以 Sonar 数据集为例,说明不同模型选择度量的效果.各子图的每个数据点代表一个分类器(考虑图形效果,分类器数目取为 20),横坐标为模型选择度量值,纵坐标为对应的测试集 AUC 值.图中实线指示在不同模型选择度量下,最好的分类器所对应的测试集 AUC 值.由图 4 可知:当以λAUC 为模型选择度量时,仅有 1 个分类器在测试集上的 AUC 值比所选择的最优分类器要高;当以ρAUC 为模型选择度量时,有 2 个分类器在测试集上的 AUC 值比所选择的最优分类器要高;而当以 AUC 或 sAUC 为模型选择度量时,有 4 个分类器在测试集上的 AUC 值比所选择的最优分类器要高.实验结果表明,λAUC 和ρAUC 优于 AUC 和 sAUC.

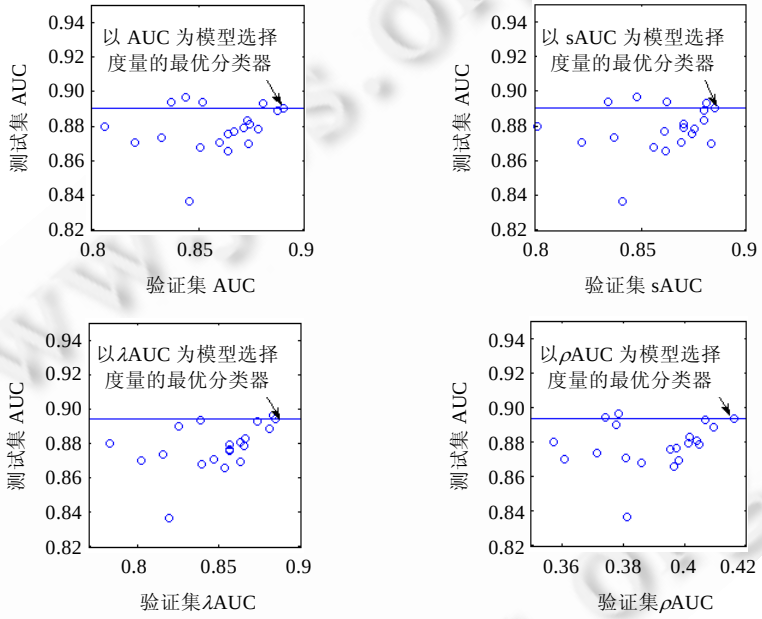


Fig.4 Model selection results of 20 classifiers under different metrics

图 4 不同度量下的 20 个分类器的模型选择结果

4.2.3 模型选择度量的相关分析

文献[9]利用 Pearson 相关分析比较了 AUC,accuracy 及 AUC:acc 与 RMS(root mean square error)的相关程度.受此启发,为了分析利用不同模型选择度量对若干分类器性能评估排序结果与测试 AUC 一致的程度,通过计算验证集上模型选择度量值与测试集 AUC 值之间的 Spearman 秩相关系数(秩相关系数又称为等级相关系数,或顺序相关系数),并进行秩相关系数检验,从而分析模型选择度量与测试集 AUC 值之间的相关程度.

仍然以 Sonar 数据集为例,采用第 4.2.2 节中的实验设置,计算在 100 个不同分类器上的 AUC,sAUC,λAUC 和ρAUC 度量值以及在测试集上的 AUC 值,取显著性水平α=0.05,求得 Spearman 秩相关系数分别如下:

r_s(AUC)=0.5154, r_s(sAUC)=0.5211, r_s(λAUC)=0.5214, r_s(ρAUC)=0.6470.

对应的 p 值分别为

p(AUC)=0.0207, p(sAUC)=0.0189, p(λAUC)=0.0186, p(ρAUC)=0.0021.

在显著性水平α=0.05 下,通过查临界值表可知,4 种度量都与测试集 AUC 相关,但根据具体相关系数值可知,λAUC 和ρAUC 与测试集 AUC 的相关程度大于验证集 AUC 和 sAUC.

5 结 论

ROC 曲线分析及 AUC 是高效的模型选择度量,但是,经验 ROC 曲线具有不确定性,从而造成 AUC 的不确定性,影响其作为模型选择度量的准确性.gROC 曲线分析方法基于分辨粒度的概念定义上下近似 ROC,讨论了 gROC 曲线的若干性质,分析表明,ROC 是 gROC 的特例.双正态模型实验结果表明,lowROC 和 upROC 能够有效反映 ROC 曲线的不确定性,与置信带方法相比,避免了大量的采样,大大提高了效率.

ROC 仅仅使用 rank,而 gROC 使用了 rank 和 score,并且充分考虑了 score 的不确定性.基于 gROC 曲线定义 gAUC 度量,也即上近似 AUC 和下近似 AUC,并在此基础上定义模型选择度量λAUC 和ρAUC.近似比率λ反映经验 ROC 曲线的内在不确定性,λAUC 融合了 AUC 值和近似比率λ两个方面,是在给定分辨粒度下的模型选择度量.ρAUC 同样兼顾了 AUC 值及 ROC 不确定性两个方面,不过,ρ是在所有分辨粒度下 ROC 不确定性的平均度量.实验结果表明,对于大样本数据集,ρAUC 比λAUC 更有效.

在 UCI 数据集上进行实验比较,由λAUC 和ρAUC 作为模型选择度量,所得分类器模型比使用 AUC 或 sAUC 所得分类器模型具有更高的 AUC 值,λAUC 和ρAUC 具有更强的对分类器性能进行比较的能力.通过 Spearman 秩相关分析实验,验证了λAUC 和ρAUC 与测试集 AUC 具有更大的相关度.

References:

[1] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006,27(8):861−874. [doi: 10.1016/j.patrec.2005.10.010]

[2] Egan JP. Signal detection theory and ROC analysis. In: Proc. of the Series in Cognition and Perception. New York: Academic Press, 1975.

[3] Spackman KA. Signal detection theory: Valuable tools for evaluating inductive learning. In: Segre AM, ed. Proc. of the 6th Int'l Workshop on Machine Learning (ML'89). San Francisco: Morgan Kaufman Publishers, 1989. 160−163.

[4] Swets J, Dawes R, Monahan J. Better Decisions Through Science. Scientific American, 2000,283(4):82−87. http://www.citeulike.org/user/rabio/article/3484365

[5] Zweig MH, Campbell G. Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clinical Chemistry, 1993,39(8):561−577.

[6] Adams NM, Hand DJ. Comparing classifiers when the misallocation costs are uncertain. Pattern Recognition, 1999,32(7): 1139−1147. [doi: 10.1016/S0031-3203(98)00154-X]

[7] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1997, 30:1145−1159. [doi: 10.1016/S0031-3203(96)00142-2]

[8] Huang J, Ling CX. Using AUC and accuray in evaluating learing algorithms. IEEE Trans. on Knowledge and Data Engineering, 2005,17(3):299−310. [doi: 10.1109/TKDE.2005.50]

[9] Huang J, Ling CX. Constructing new and better evaluation measures for machine learning. In: Proc. of the 20th Int'l Joint Conf. on Artifical Intelligence. 2007. 859−864. http://www.ijcai.org/papers07/Papers/IJCAI07-138.pdf

[10] Wan BK, Xue SJ, Li J, Wang RP. Application of ROC curve to select the pattern classification algorithms. Progress in Natural Science, 2006,16(11):1511−1516 (in Chinese with English abstract).

[11] Castanho MJP, Barros LC, Yamakami A, Vendite LL. Fuzzy receiver operating characteristic curve: An option to evaluate diagnostic tests. IEEE Trans. on Information Technology in Biomedicine, 2007,11(3):244−250. [doi: 10.1109/TITB.2006.879593]

[12] Ferri C, Flach P, Hernández-Orallo J, Senad A. Modifying ROC curves to incorporate predicted probabilities. In: Proc. of the ICML 2005 Workshop on ROC Analysis in Machine Learning. Bonn, 2005. http://www.dsic.upv.es/%7Eflip/ROCML2005/papers/ferriCRC.pdf

[13] Wu SM, Flach P. A scored AUC metric for classifier evaluation and selection. In: Proc. of the ICML 2005 Workshop on ROC Analysis in Machine Learning. Bonn, 2005. http://www.dsic.upv.es/%7Eflip/ROCML2005/papers/wuCRC.pdf

[14] Wu SM, Flach P, Ferri C. An improved model selection heuristic for AUC. In: Proc. of the 18th European Conf. on Machine Learning (ECML 2007). Berlin: Springer-Verlag, 2007. 478−489. [doi: 10.1007/978-3-540-74958-5_44]

[15] Calders T, Jaroszewicz S. Efficient AUC optimization for classification. In: Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007). Berlin: Springer-Verlag, 2007. 42−53. [doi: 10.1007/978-3-540-74976-9_8]

[16] Vanderlooy S, Hüllermeier E. A critical analysis of variants of the AUC. Machine Learning, 2008,72(3):247−262. [doi: 10.1007/s10994-008-5070-x]

[17] Hand DJ. Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 2009,77(1):103−123. [doi: 10.1007/s10994-009-5119-5]

[18] David RP, Steven CG, Mark EO, Timothy DR. Development of a bayesian framework for determining uncertainty in receiver operating characteristic curve estimates. IEEE Trans. on Knowledge and Data Engineering, 2010,22(1):31−45. [doi: 10.1109/TKDE.2009.50]

[19] Macskassy S, Provost F. Confidence bands for ROC curves: Methods and an empirical study. In: Hernández-Orallo J, Ferri C, Lachiche N, Flash PA, eds. Proc. of the 1st Workshop ROC Analysis in AI (ROCAI 2004) at ECAI-2004. 2004. 61−70. http://www.dsic.upv.es/ecai2004/workshops/accepted.html#w19#w19

[20] Macskassy S, Provost F, Rosset S. ROC confidence bands: An empirical evaluation. In: De Raedt L, Wrobel S, eds. Proc. of the 22nd Int'l Conf. on Machine Learning (ICML 2005). New York: ACM Press, 2005. 537−544. [doi: 10.1145/1102351.1102419]

[21] Elazmeh W, Japkowicz N, Matwin S. A framework for comparative evaluation of classifiers in the presence of class imbalance. In: Proc. of the ICML 2006 Workshop on ROC Analysis in Machine Learning. Pittsburgh, 2006. http://users.dsic.upv.es/~flip/ROCML2006/Papers/elazmehROCML06.pdf

[22] Efron B, Tibshirani R. An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.

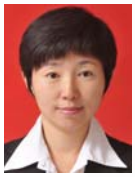
[23] David RP. Uncertainty estimation for target detection system discrimination and confidence performance metrics [Ph.D. Thesis]. Ohio: Air Force Institute of Technology, 2006.

[24] Metz C, Herman B, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Statistics in Medicine, 1998,17(9):1033−1053.

[25] Blake C, Keogh E, Merz CJ. UCI repository of machine learning databases. Irvine: Department of Information and Computer Science, University of California, 2011. http://www.ics.uci.edu/mlearn/MLRepository.html

附中文参考文献:

[10] 万柏坤,薛召军,李佳,王瑞平.应用 ROC 曲线优选模式分类算法.自然科学进展,2006,16(11):1511−1516.



董元方(1975－),女,内蒙古通辽人,博士,讲师,主要研究领域为机器学习,数据挖掘.
E-mail: yf.dong@163.com

李军(1974－),男,博士,副教授,CCF 会员,主要研究领域为机器学习,数据挖掘.
E-mail: Lijun.yq@163.com

李雄飞(1963－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为机器学习,知识发现,数据挖掘.
E-mail: lxf@jlu.edu.cn

赵海英(1969－),女,博士,副教授,主要研究领域为模式识别,图像图形处理.
E-mail: Zhaohaiying2008@gmail.com