基于代价信息的二类分类器性能评估方法

姜 鹏[1],秦 锋[2],罗 慧[2]
(1.安徽工业大学 电气工程学院,安徽 马鞍山 243002;
2.安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:基于ROC曲线的AUC评估方法能有效评估二类分类器的性能,但是该方法只能评估分类器的总体性能,对代价信息不敏感。基于AUC方法提出用AUCCH方法评估二类分类器性能,该方法在具体代价信息下能分辨出最优分类器,在代价信息未知时能分辨出潜在最优分类器。在MBNC实验平台下编程实现,通过对AUC方法和AUCCH方法实验结果的比较,表明该方法具有有效性和健壮性。

关键词:AUC;二类分类器;代价信息;AUCCH;最优分类器;潜在最优分类器
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)12-0063-04

Method for Appraising Performance of Two-Classifier Based on Cost Information

JIANG Peng[1], QIN Feng[2], LUO Hui[2]
(1. Sch. of Electronics and Info. Eng., Anhui University of Technology, Ma'anshan 243002, China;
2. School of Computer Science, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: The AUC method can effectively appraise performance of two-classifier, but it only can appraises overall performance of classifier and is insensitive to cost information. In this paper, a new method of appraising performance of two-classifier which is based on AUC is referred. The new method is called AUCCH. With cost information, the method can identify the most optimal classifier, With no cost information, it can identify the potential optimal classifier. Making experiment in MBNC experiment platform, through comparing the results between the AUC method and the AUCCH method, the results show that the new method is effective and robust.

Key words: AUC; two-classifier; cost information; AUCCH; optimal classifier; potential optimal classifier

0 引 言

机器学习领域中分类算法是众多研究学者关注的热点,如何得到性能优越的分类器更是热中之热的问题,优秀的评估标准非常重要。一种优异的分类评估标准是有效性和健壮性的结合。而现在的众多评估标准考虑有效性的较多,对健壮性却有忽视。

定义1:健壮性(Robust) 如果分类器的性能评估方法对类的不同错误分类代价敏感,则该评估方法具有健壮性。

近几年来,基于ROC曲线(Receiver Operating characteristics Curve)分析的AUC(the Area Under the ROC Curve)评估方法被认为是一种优越的评估尺度越来越多地应用到分类器评估中[1]。AUC方法优于传统的准确性评估标准,但是如果考虑具体的代价信息,该方法就有一定的局限性。如得到的两条ROC曲线相互交叉,其中某一条曲线可能在某些范围内高于另一条曲线,此时,AUC方法只能在总体性能上评估分类器的性能,对不同的代价信息不能做出具体的判断。文中提出一种基于AUC和代价信息的二类分类器性能评估方法AUCCH(AUC Convex Hull)[2]来评估分类器性能,AUCCH方法是通过绘制出不同分类器产生的ROC曲线簇的凸弧,根据错误代价比产生的同等性能线与AUCCH曲线的切点来判断出该代价比下的最优分类器。该方法不仅能找出某一具体代价信息下的最优的分类器,还能在代价信息未知时找出潜在的最优分类器。

收稿日期:2008-03-16
基金项目:安徽省自然科学研究重点项目(KJ2007A051)
作者简介:姜 鹏(1968-),男,江苏江都人,高级工程师,研究方向为自动化控制和人工智能;秦 锋,教授,硕士生导师,研究方向为人工智能、数据挖掘、机器学习。

1 同等性能线

设{e,n}分别代表专家评估的正例和反例,{E,

$N\}$ 分别代表分类器预测的正例和反例，$P(e)$ 是专家评估为正例占的比例，则 $P(n) = 1 - P(e)$ 是专家评估为反例占的比例，$P(E \mid n)$ 为分类器预测的正例占所有的反例的比例，$P(N \mid e)$ 为分类器预测的反例占所有的正例的比例。根据决策分析理论[3]，基于代价敏感的二类分类器的分类代价可由式(1)得出：

$$C_1 = c(E,n)P(E \mid n) + c(N,e)P(N \mid e) \quad (1)$$

其中 $c(E, n)$ 表示错误的正例分类代价，$c(N,e)$ 表示错误的反例分类代价，$c(E,n)/c(N,e)$ 表示错误分类的代价比。

表1列出了二类分类模型的混淆矩阵，其中 $FP = P(E \mid n)$，$FN = 1 - TP = P(N \mid e)$。

表1 二类分类模型的混淆矩阵

	预测正例(E)	预测反例(N)
正例(e)	正确分类的正例(TP)	错误分类的反例(FN)
反例(n)	错误分类的正例(FP)	正确分类的反例(TN)

考虑混淆矩阵和测试例的先验概率可得式(2)：

$$C_2 = P(n)FPc(E,n) + P(e)(1 - TP)c(N,e) \quad (2)$$

在 ROC 曲线空间中连接两个点坐标(FP_1, TP_1)和(FP_2, TP_2)，得到的直线称为同等性能线(Iso - Performance)，位于同一条同等性能线上的分类器具有相同的代价。进而计算同等性能线的斜率：

$$k = \frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{P(n)c(E,n)}{P(e)c(N,e)} \quad (3)$$

由式(3)知，同等性能线的斜率 k 越大，分类器的分类代价越小[3]。

2 AUC 评估方法

AUC 评估方法是通过计算 ROC 曲线下的面积得到分类器的性能，其值的大小介于 0 到 1 之间。AUC 值等于 0.5 的分类器，近似于随意猜测的结果；AUC 值小于 0.5 的分类器性能很差，没有任何现实意义；只有 AUC 值大于 0.5 时评估分类器性能才有意义，且 AUC 值越大则分类器性能越好。

文中采用 Hand 和 Till 提出的一种较为简便的计算方法[4]：先按后验概率的大小将数据集记录重排成新的排序表，再按如下公式计算：$AUC = \frac{s_0 - n_0(n_0+1)/2}{n_0 n_1}$，其中，$n_0$ 和 n_1 分别是测试数据集中正例个数和负例个数，$s_0 = \sum r_i$，r_i 是第 i 个正例在排序表中的序号。

3 一种新分类器评估方法——AUCCH 方法

AUCCH 方法是在 ROC 曲线空间中绘制出多个分类器产生的 ROC 曲线簇的凸弧，凸弧称为 AUCCH 曲线，在代价信息确定的情况下，不同的代价信息得到不同的同等性能直线，根据同等性能直线与 AUCCH 曲线的切点来判定此类代价信息下的最优分类器。在代价信息不确定的情况下，只要某一分类器的 ROC 曲线段与 AUCCH 曲线有重合部分，该分类器在某种分类代价下就可能成为最优分类器称为潜在最优的分类器[3,5]；同理，当 ROC 曲线位于 AUCCH 曲线下方时，所对应的分类器在任何代价信息下都不会成为最优分类器，该分类器的性能较差，故可事先排除。

用 AUCCH 方法评估多个二类分类器的步骤如下：

Step 1. 用 AUC 评估方法，绘制不同分类器的 ROC 曲线。

Step 2. 绘制 ROC 曲线簇的最外轮廓 AUCCH 曲线，算法遵照 AUCCH 曲线必须拥有单调递减的斜率原则[6]。AUCCH 曲线是由多个 ROC 曲线的最外凸弧构成，从(0,0)点开始连接不同 ROC 曲线上的点，如果新 ROC 点的连接线段的斜率大于前面线段的斜率，则放弃前面连接的 ROC 点，重复此操作直到 ROC 曲线簇中所有的点的斜率都被比较，这样就找到了(0,0)点到下一个点的最大斜率线段，以上一个斜率线段的末端为起点连接其后的 ROC 点，重复上述操作找到此点的最大斜率线段，如此进行下去，就得到一个斜率单调递减的凸弧曲线。

Step 3. 根据不同分类器的 ROC 曲线是否与 AUCCH 曲线相切，判别出潜在最优分类器，位于 AUCCH 曲线之下的分类器可事先删除。

Step 4. 根据不同的正例和负例的错误分类的代价比，得到不同的同等性能线，分析出在不同代价下的最优分类器。

如图 1 所示，其中横轴 FPR(False Positive rate)是错误的正例率，数值从 0～1，纵轴 TPR(True Positive rate)是正确的正例率，数值从 0～1，$TPR = \frac{TP}{P}$，$FPR = \frac{FP}{N}$，P 和 N 分别代表正例总数和负例总数，评估分类器 A、B、C 得到三条 ROC 曲线，由三条曲线的最外轮廓得到 AUCCH 曲线，k_1、k_2 是同等性能线。分类器 C 对应的 ROC 曲线明显位于 AUCCH 曲线下，在任何代价下都不会成为最优分类器，表明其性能较差，在选择最优分类器时可以事先拒绝这个分类器。分类器 A、B 对应的 ROC 曲线段与 AUCCH 曲线有重合部分，分类器 A、B 在某些代价条件下可能成为最优分类器，是潜在的最优分类器。

不同的分类代价对应产生一条同等性能线，同等

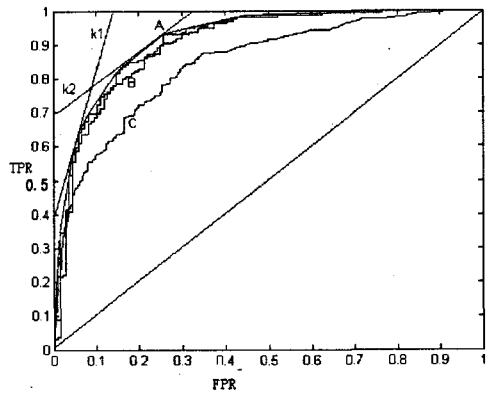

图 1　AUCCH 曲线

性能线与 AUCCH 曲线相切时,切点所在的 ROC 曲线对应的分类器就是该分类代价下的最优分类器。例如,测试数据中负例数目是正例数目的 10 倍,假设有两种不同的错误分类代价:情况 a,负例的错误分类代价是正例的错误分类代价的 10 倍;情况 b,负例的错误分类代价是正例的错误分类代价的 2.5 倍,根据式(3)可得到不同代价下的同等性能线,如图 2 所示。

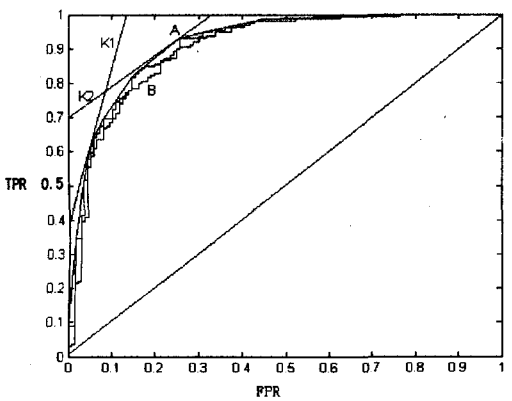

图 2　潜在最优的分类器

计算同等性能线的斜率:

(a) $k_1 = \frac{10 \times 1}{1 \times 10} = 1$

(b) $k_2 = \frac{10 \times 1}{1 \times 2.5} = 4$

同等性能线 k_1 和 k_2 的斜率分别为 1 和 4,显而易见,在情况 a 下,k_1 所切的 AUCCH 曲线中与分类器 A 的 ROC 曲线重合,故分类器 A 的性能为最优;而在情况 b 下,k_2 所切的 AUCCH 曲线中与分类器 B 的 ROC 曲线重合,故分类器 B 的性能为最优。

4　实验设计与结果分析

4.1　实验设计

用 AUCCH 方法评估分类器性能在 MBNC(Bayesian Networks Classifier using Matlab)实验平台[7]下编程完成,MBNC 实验平台集成了多种贝叶斯分类器,文中选用朴素贝叶斯分类器 NBC(Naive Bayes Classifier)和树扩展朴素贝叶斯分类器 TANC(Tree Augmented Naive Bayes Classifier)两种分类器作为评估对象。

测试用的标准数据集是从 UCI(University of California in Irvine)上下载,下载网址是 http://www.ics.uci.edu/mlearn/MLRepository.html。文中选用 4 个 2 类别的标准数据集,经过预处理后的数据集概况见表 2。

表 2　数据集的概况

数据集	属性数	例数
Pima	6	768
Corral	7	128
Glass2	6	163
Cleve	11	296

4.2　实验结果分析

用 AUC 方法评估分类器得到的实验结果见表 3。其中,NBC－AUC 表示用 AUC 方法评估 NBC 得到的数值,TANC－AUC 表示用 AUC 方法评估 TANC 得到的数值。

表 3　用 AUC 方法评估的实验结果

数据集	NBC－AUC	TANC－AUC	最优分类器
Pima	0.84	0.85	TANC
Corral	0.94	1	TANC
Glass	0.90	0.89	NBC
Cleve	0.90	0.91	TANC

用 AUCCH 方法分别评估 NBC 和 TANC 得到的曲线如图 3 至图 6 所示,随机在曲线上选择 2 个点,进而绘制同等性能线。实验结果见表 4。第 2 列表示随机选取点的错误分类代价比 $c(E,n)/c(N,e)$,同等性能线的斜率由式(3)计算可得。

通过对表 3 和表 4 实验数据的分析可知,AUCCH 方法较之 AUC 方法评估分类器有更大的健壮性:

(1)在不同代价信息下 AUCCH 方法能分辨出最

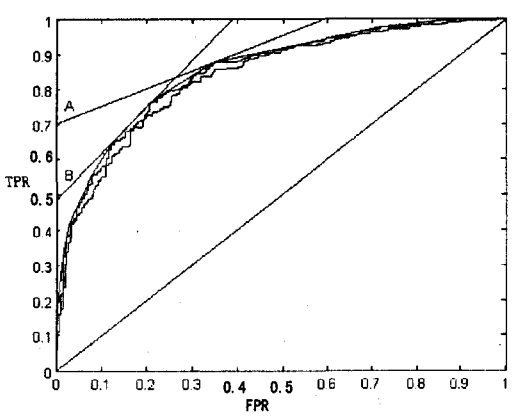

图 3　Pima 数据集的 AUCCH 曲线

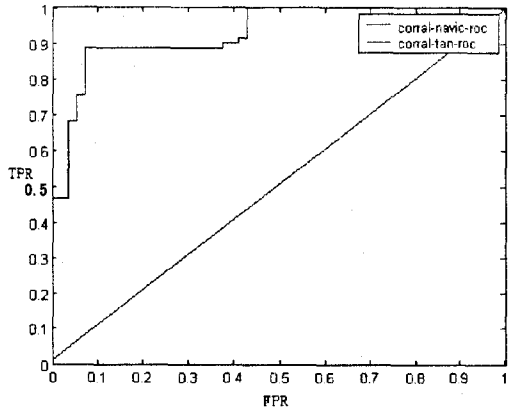

图 4 Corral 数据集的 AUCCH 曲线

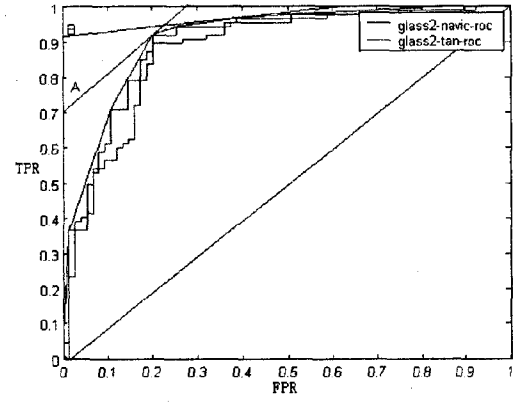

图 5 Glass 数据集的 AUCCH 曲线

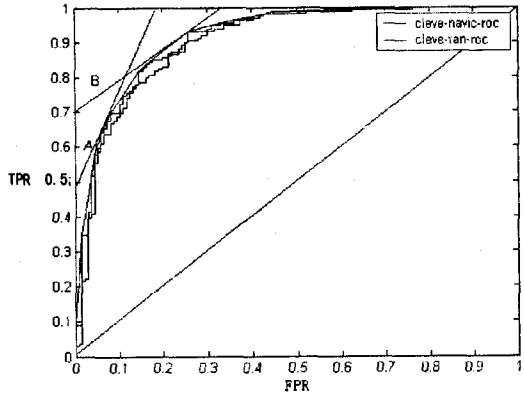

图 6 Cleve 数据集的 AUCCH 曲线

表 4 用 AUCCH 方法评估的实验结果

数据集	$C(E,n)/C(N,e)$	同等性能线 k	最优分类器
Pima	0.95	$A=0.51$	NBC
Pima	2.67	$B=1.44$	TANC
Corral	0.69	$A=0.49$	TANC
Corral	2.35	$B=1.67$	TANC
Glass	0.13	$A=0.11$	NBC
Glass	0.17	$B=0.15$	TANC
Cleve	3.4	$A=2.89$	TANC
Cleve	1.07	$B=0.91$	NBC

优分类器。表 3 中数据集 Pima 和 Cleve 的 TANC - AUC 值高，表明 TANC 的性能优于 NBC；数据集 Glass 的 NBC - AUC 值高，表明 NBC 的性能优于 TANC，AUC 评估方法只能总体上评估分类器的性能。而由表 4 可知，不同的错误分类代价比，AUCCH 方法得到的不同最优分类器，NBC 和 TANC 分类器在不同分类环境都可能是最优分类器。

(2)在代价信息未知的情况下 AUCCH 方法能分辨出潜在的最优分类器。如图 2 所示，对于数据集 Pima、Corral 和 Cleve，评估 NBC 得到的 ROC 曲线和评估 TANC 得到的 ROC 曲线都相切于 AUCCH 曲线，表明 NBC 和 TANC 分类器都是潜在的最优分类器；对于数据集 Corral，只有评估 TANC 得到的 ROC 曲线都切于 AUCCH 曲线，而评估 NBC 得到的 ROC 曲线完全位于 AUCCH 曲线之下，因此 TANC 是潜在最优分类器，NBC 分类器在任何情况下都不可能是潜在最优分类器，在这种数据分布下 NBC 分类器可以事先拒绝。

综上可知，用 AUCCH 方法评估分类器性能要优于用 AUC 方法，能根据具体代价信息作出具体的判断，更适合现实领域的应用，充分体现了该方法的健壮性，实验结果也表明该方法是有效的和健壮的。

5 结束语

AUCCH 方法能根据错误代价信息评估出最优分类器和分辨出潜在最优分类器，有较强的健壮性，更加适用于现实环境，有着广泛的应用前景。进一步的工作是如何将该方法从二类分类推广到多类分类的评估中。

参考文献：

[1] Ling C X, Huang Jin, Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy[C]//In proceeding of 18th International conference on Artificial Intelligence. [s.l.]: [s.n.], 2003: 329 - 341.

[2] Provost F, Fawcett T. Robust Classification for Imprecise Environments[J]. Machine Learning, 2001, 42(3): 203 - 208.

[3] Provost F, Fawcett T. Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions[C]//In Proc. Third Intl. Conf. Knowledge Discovery and Data Mining (KDD - 97). Menlo Park, CA: AAAI Press, 1997: 43 - 48.

[4] Hand D J, Till R J. A simple generalization of the area under the ROC curve for multiple class classification problems[J]. Machine Learning, 2001(45): 171 - 186.

[5] Bettinger R. Cost - Sensitive Classifier Selection Using the ROC Convex Hull Method Release 4.1, SAS Institute[M]. Cary, NC: [s.n.], 2000.

[6] 骆名剑. 基于 ROC 的分类算法评估方法[D]. 武汉：武汉科技大学，2005.

[7] 程泽凯，林士敏，陆玉昌，等. 基于 Matlab 的贝叶斯分类器平台 MBNC[J]. 复旦学报，2004，43(5)：729 - 732.