二类分类器的ROC曲线生成算法

邹洪侠,秦 锋,程泽凯,王晓宇
(安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:ROC曲线分析技术越来越多地被应用在机器学习和数据挖掘领域中,用来全面度量分类器的性能。ROC曲线分析是对分类器性能的二维描述,它对类分布和不同错误分类代价不敏感,但直观、可理解等特性使它在类分布未知的领域和代价敏感学习中变得越来越重要。准确高效地绘制出分类器的ROC曲线是使用ROC曲线分析技术及其AUC方法全面度量分类器性能的基础,也是进行代价敏感学习的关键。文中将从理论和具体实现两方面分别对二类分类器的ROC曲线生成算法及绘制ROC曲线的具体过程做详细阐述,基于MBNC实验平台,使用MATLAB语言构建该算法,进而比较不同分类器在不同类分布下的分类性能。通过观察实验结果可知,提出的ROC曲线生成算法准确可行,符合实际。

关键词:分类器评估;ROC曲线;MATLAB
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2009)06-0109-04

Algorithm for Generating ROC Curve of Two-Classifier

ZOU Hong-xia, QIN Feng, CHENG Ze-kai, WANG Xiao-yu
(School of Computer Science, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: The ROC curve analysis is applied more and more in the machine learning and the data mining domain, which is used to measure classifier's performance comprehensively. The ROC curve analysis is a two-dimensional description for classifier's performance. It is insensitive to the class distribution and the different misclassification costs, direct-viewing as well as being understood and so on, these characteristics make it more and more important in the distribution unknown domain and the cost sensitive learning. You must draw ROC curve efficiently and accurately so that you could measure the performance of the classifier using the ROC curve analysis technology and its AUC approach, and it is also the key to sensitive learning. On drawing ROC curve, there is not yet a tool on hand at present. It will elaborate the algorithm of generating two-classifier's ROC curve and the concrete process of drawing ROC curve from the theory and the experiment separately in detail, and construct the algorithm using the MATLAB language based on MBNC experiment platform, then compare different classifiers' performance under different kind of distribution. By observing the resulits of the experiment, can see that the algorithm generating ROC curve in this paper is accurate, feasible and realistic.

Key words: classifier appraisal; ROC curve; MATLAB

0 引 言

准确率评估方法和ROC曲线(Receiver Operating Characteristic,受试者工作特征曲线)分析及AUC评估方法是2种主要的分类器性能评估方法。近几年来,ROC曲线分析方法得到了更广泛的使用。理论上,与准确率评估方法相比,ROC曲线分析方法有以下优点:

(1)充分利用了预测得到的概率值。

(2)给出不同类的不同分布情况差别,即当是不平衡数据时,不同的数据分布,将得到不同的分类结果,而准确率评估则默认所有的数据集都是平衡数据集。

(3)考虑了不同种类错误分类代价的不同,而准确率评估默认所有的错误代价都是相同的,这在现实生活中是不切实际的。

(4)二类分类的ROC曲线通过斜率反映了正例和反例之间的重要关系,同时也反映出类的分布和代价之间的关系。

(5)可以使分类器的评估结果用曲线的形式更直观地展示在二维空间中[1]。

分类器分为离散分类器(Discrete Classifier)和概率分类器(Probabilistic Classifier),前者如决策树,后者如

收稿日期:2008-09-24;**修回日期**:2008-12-27
基金项目:安徽省自然科学研究重点项目(KJ2007A051);安徽省自然科学研究项目(2006KJ061B)
作者简介:邹洪侠(1982-),女,吉林松原人,硕士研究生,研究方向为人工智能、数据挖掘、机器学习;秦 锋,教授,研究方向为人工智能、数据挖掘、机器学习;程泽凯,副教授,研究方向为人工智能、数据挖掘、机器学习。

贝叶斯分类器和神经元网络。离散分类器只预测类别,在 ROC 空间中只产生一个单独的点,而概率分类器则对每个实例产生一个概率值,该值表示实例属于正例的可能性大小。ROC 分析正是利用这些概率值产生代表不同分类器性能的 ROC 曲线。

1 有关 ROC 曲线的几个基本概念

(1) 混淆矩阵(Confusion Matrix)。

在二类分类问题中,令 $\{p,n\}$ 是专家预测的类标签集合,其中 p 代表正例,n 代表负例。测试集合中每个实例 i 的类别都是该集合中的一个元素;令 $\{T,F\}$ 为分类器预测的类标签集合,每个实例 i 将被分类器预测为 $\{T,F\}$ 中的一个类别。那么给定一个分类器和实例,将会出现 4 种可能的结果:如果专家判断是正例,被预测的也是正例,则真正的正例 TP(True Positive)加 1;如果专家判断是反例,被预测成正例,则错误的正例 FP(False Positive)加 1;如果专家判断是正例,被预测成反例,则错误的负例 FN(False Negative)加 1;如果专家判断是反例,被预测的也是反例,则真正的负例 TN(True Negative)加 1。可以用表 1 所示的混淆矩阵来表示测试集合中实例的分布情况。

表 1 二类别混淆矩阵

分类器预测	专家预测	
	P	n
T	TP	FP
F	FN	TN

据此,可以得到一个对应于混淆矩阵的概率矩阵 $\boldsymbol{M}$,如表 2。

表 2 二类别概率矩阵

	P	N
T	TPR	FPR
F	FNR	TNR

在逻辑上,显然可以得出:TPR = 1 - FNR ,FPR = 1 - TNR[2]。

(2) ROC 曲线。

ROC 曲线是以 FPR 为横轴,TPR 为纵轴,横轴与纵轴长度相等,为单位 1,形成一个正方形的二维空间,在此二维空间中将各个(FPR,TPR)点标出,用直线连接各相邻两点构建而成的一条曲线,如图 1。

2 绘制 ROC 曲线的理论方法

如前所述,概率分类器对每个实例都会产生一个概率值,通过设置一个阈值 t 可以将概率分类器用做一个离散分类器:如果概率分类器对实例 i 的预测结果大于阈值 t 时,实例 i 被划分为正例,否则被划分为负例,当所有实例被处理完后,便会在 ROC 空间中产生一个点。那么如果让阈值 t 不断变化,就会产生无数的点,将这些点连接起来就形成一条曲线。理论上,阈值 t 可以在 $-\infty$ 和 $+\infty$ 之间取值,但概率分类器对实例的预测值在 0 到 1 之间,显然 $-\infty$ 到 $+\infty$ 这个取值范围过大,所以不用,而是取实例 i 属于正例的分类器预测概率值 $f(i)$ 的最小值和最大值作为阈值 t 取值范围的下限和上限。以上便是绘制 ROC 曲线的基本思想。

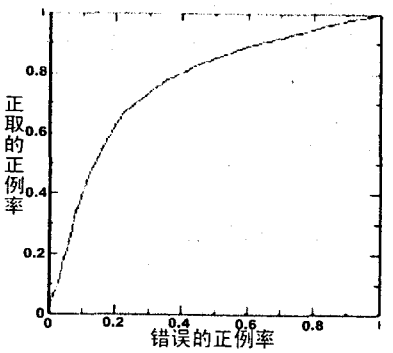

图 1 一个 ROC 曲线的实例

给定一个分类器和测试集合,研究学者们往往想要高效地绘制出相应的 ROC 曲线。但根据以上方法所设计出来的算法绘制 ROC 曲线效率和实用性都很差:它要求知道概率值中的最小值和最大值,以及循环变量从最小值变化到最大值的步长,而这些都需要从测试集合和 f 值中比较得出;并且它包括了 2 层循环,算法的复杂度为 $O(n^2)$(n 为测试集合中的实例数目)[1]。所以实际应用并不采用上述算法,而是使用文中所讨论并实现的一种简化的算法。

3 ROC 曲线实现方法

结合 ROC 曲线的理论算法思想,并考虑这样一个事实:如果一个实例 i 在阈值为 T 时被划分为正例,那么当给定阈值 $t \leqslant T$ 时,实例 i 也将被划分为正例。可以得到一个绘制 ROC 曲线的简化方法:将由分类器预测得到实例 i 属于正例的概率值 $f(i,+)$ 和属于负例的概率值 $f(i,-)$ 做比较,并将比值按递减顺序排列成一个表,从上到下扫描该表。每次扫描该表中的一个值时处理一个实例并将 FP 和 TP 的值做相应调整,保存并描点。具体的实现步骤见图 2。

4 绘制 ROC 曲线的算法伪码

为叙述方便,文中将以下算法记为算法 1。

算法 1 绘制 ROC 曲线的实现算法。

输入:实例集合 L,所有实例的概率值 $f(i,+)$ 和

$f(i, -)$，正例数目 p，反例数目 n。

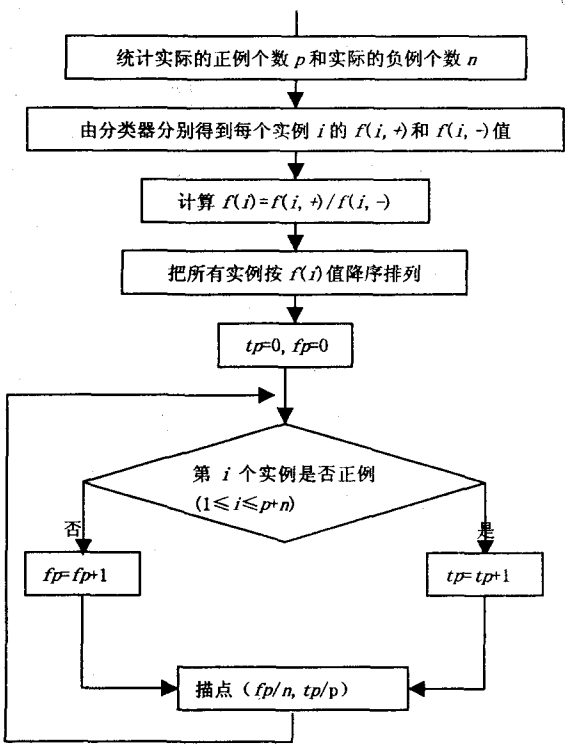

图 2　绘制 ROC 曲线的流程图

输出：ROC 曲线。

算法：对于每一个实例，计算 $f(i)$：$f(i) = f(i, +)/f(i, -)$

把所有的 $f(i)$ 值按降序排列

初始绘图点为(0,0)

for 每个不同的值 $f(i)$

if 下一个实例的值为正例 then 绘图点向上移动 $1/p$ 个单位

else 绘图点向右移动 $1/n$ 个单位

end if

end for

对于算法 1，给定如表 3 所示的测试集合，其中包含 10 个正例和 10 个负例，由概率分类器预测的概率值按递减顺序排序。图 3 给出了算法 1 利用表 3 数据绘制得到的 ROC 曲线。其中 20 个概率值就是 20 个决策阈值，分类器根据不同的决策阈值进行分类，得到从点(0,0)向点(1,1)方向延伸形成一条 ROC 曲线。实际上，任何从有限数据集合生成的 ROC 曲线都是阶梯形状的，测试数据集合中实例数目趋近无限的时候，ROC 曲线接近真实的曲线。

5　实验结果及分析

根据算法 1，基于 MBNC(Bayesian Networks Classifier using Matlab)实验平台[3]采用 5 叠交叉验证

表 3　一个概率分类器在 20 个测试样例上的概率表

序号	类别	概率	序号	类别	概率
1	+	0.90	11	+	0.40
2	+	0.80	12	-	0.39
3	-	0.70	13	+	0.38
4	+	0.60	14	-	0.37
5	+	0.55	15	-	0.36
6	+	0.54	16	-	0.35
7	-	0.53	17	+	0.34
8	-	0.52	18	-	0.33
9	+	0.51	19	+	0.30
10	-	0.50	20	-	0.10

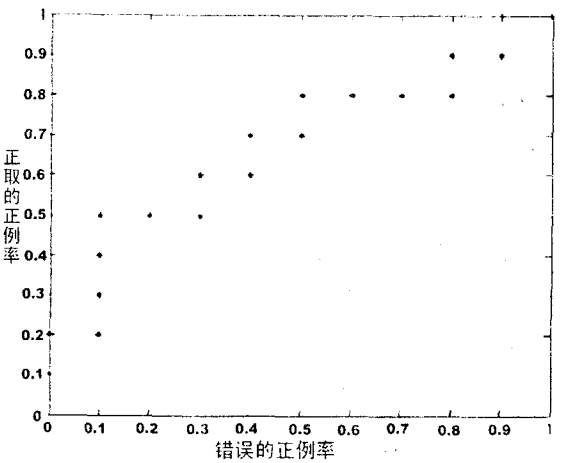

图 3　算法 1 在表 3 所示数据上的 ROC 曲线

法[4]，用 MATLAB 语言编写程序[5]绘制评估朴素贝叶斯分类器 NBC(Naive Bayes Classifier)和树扩展朴素贝叶斯分类器 TANC(Tree Augmented Naive Bayes Classifier)的分类器性能的 ROC 曲线，所用数据集为 UCI(University of California in Irvine)标准数据集 car[6]。

当数据集中正例和负例之比为 1:3 时，分别得到如组图 4 所示的 NBC 分类器的 ROC 曲线和组图 5 中的 TANC 分类器的 ROC 曲线。其中每组图中左边的曲线是第一叠验证中绘制的 ROC 曲线，右边的曲线是将所有数据集中实例重新合并成一个集合，并根据实例的概率值按降序排列后得到的平均 ROC 曲线；图 6 和图 7 中的曲线是改变数据集中正负例的比例为 10:1 后得到的 NBC 和 TANC 对应的五叠所有数据的 ROC 曲线。

一个优秀分类器对应的 ROC 曲线应该尽量靠近单位方形的左上角[1]。比较组图 4 中的 NBC 分类器和组图 5 中的 TANC 分类器的 ROC 曲线可以看出，TANC 分类器的分类性能优于 NBC 分类器；当类分布改变后，比较图 6 和图 7 中的 ROC 曲线，仍旧能够得出相同的结论。这表明算法 1 是正确和有效的。

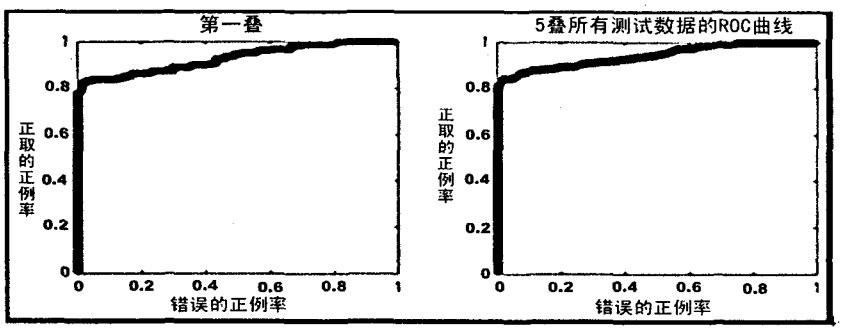

图 4 算法 1 在 car 数据集下的 NBC 分类器的 ROC 曲线(正负例比 1:3)

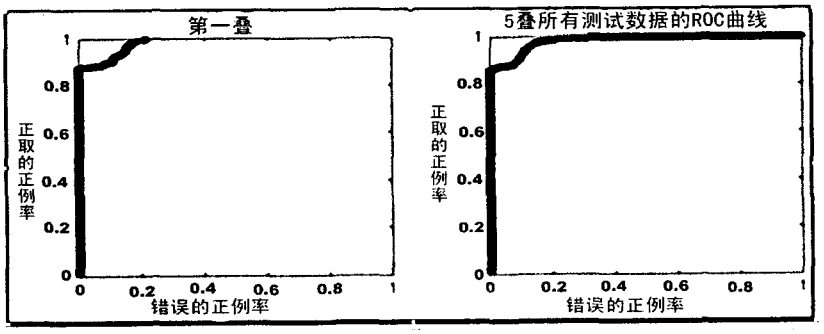

图 5 算法 1 在 car 数据集下的 TANC 分类器的 ROC 曲线(正负例比 1:3)

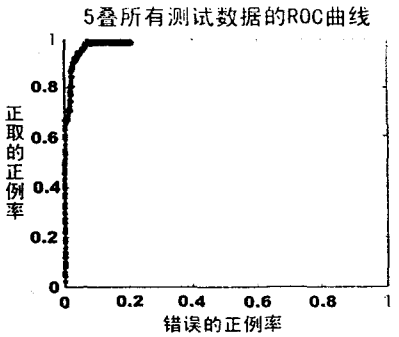

图 6 正负例比 10:1 时 NBC 对应的 ROC 曲线

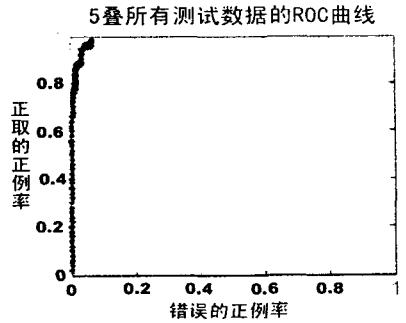

图 7 正负例比 10:1 时 TANC 对应的 ROC 曲线

6 结束语

用 MATLAB 语言设计并绘制了二类分类器的 ROC 曲线,完成了分类器评估结果的可视化要求,可以初步判断出两种分类器性能的优劣。虽然,在实际应用中,研究学者们并不采用直接观察 ROC 曲线的方法来判断不同分类器的性能。但是 ROC 曲线具有对类分布和不同错误代价不敏感等优越性能,为用 AUC 的方法[7]进一步精确评估不同分类器的总体性能提供了很好的理论基础,增加了最终评估结果的可信度。另外,AUC 评估方法是一种分类器性能评估的全局测度,所以不能区分具体类分布和错误分类代价下的最优分类器。使用 ROC 曲线与同等性能直线相结合的方法,在类分布和错误代价未知情况下可以找出潜在最优分类器;当类分布和错误代价已知时,可以找出最优分类器[7,8]。所以选择合适的算法,准确高效地绘制出 ROC 曲线具有重要意义。

采用了五叠交叉验证的方法,得到的 ROC 曲线是根据所有五叠验证时产生的所有概率值绘制的平均 ROC 曲线。到目前为止,有很多关于如何平均多条 ROC 分析曲线的讨论,如阈值平均法和垂直平均法。但如何最优地平均 ROC 曲线还有待进一步研究。

参考文献:

[1] Fawcett T. Roc Graphs: Notes and Practical Considerations for Researchers[R]. Palo Alto, CA: HP Laboratories, 2004.

[2] Han J, Kamber M. 数据挖掘概念与技术[M]. 范 明,孟小峰译. 北京:机械工业出版社,2001.

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

[4] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition Society, 1997, 30: 1145-1159.

[5] 张 铮,杨文平,石博强,等. MATLAB 程序设计与实例应用[M]. 北京:中国铁道出版社,2003.

[6] Bohanec M. UCI[DB/OL]. 1997-06-01. http://www.ics.uci.edu/mlearn/MLRepository.Html.

[7] 杨 波,程泽凯,秦 锋. 用 AUC 评估分类器的预测性能[J]. 情报学报,2007,26(2):275-279.

[8] 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.