基于ROC曲线寻优的支持向量机性能研究

王旭辉[1] 舒 平[1] 曹 立[2]
(中国民用航空总局航空安全技术中心航空安全技术实验室 北京 100028)[1]
(南京航空航天大学民航学院 南京 210016)[2]

摘 要 支持向量机在小样本模式识别领域具有优势,但其性能评估及核参数、正则化参数的选择尚未有标准算法。将受试者操作特性曲线(Receiver Operating Characteristic,ROC)引入支持向量机分类性能分析和建模参数优化问题。在核参数及正则化参数所构成的二维空间中,调整模型参数阈值描绘ROC曲线,通过比较不同分类器ROC曲线下面积实现模型的性能分析,研究了基于ROC曲线最佳工作点的模型优化问题。工程实例表明,ROC曲线下面积有效地量化了模型的识别性能,并给出了一定寻优范围内的模型参数最优点,可以在SVM模型参数优化问题中推广应用。

关键词 模式识别,支持向量机,参数优化,受试者操作特性曲线
中图法分类号 TP274,TP391 **文献标识码** A

Performance Evaluation with Optimization Strategy for Support Vector Machine Based on ROC Curve

WANG Xu-hui[1] SHU Ping[1] CAO Li[2]
(Aviation Safety Institute, Center of Aviation Safety Technology, General Administration of Civil Aviation of China, Beijing 100028, China)[1]
(College of Civil Aviation, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China)[2]

Abstract Support vector machine (SVM) has become a popular tool in the area of pattern recognition, and parameters selection for SVM is an important issue to make it practically useful. In this paper, we introduced Receiver Operating Characteristic Curve into the performance evaluation and model optimization of SVM within the kernel parameters s and penalty factor c. Area under ROC curve was applied to the model evaluation, and model optimization was performed by seeking of optimal operating point of ROC. Pattern recognition experiment with UCI dataset shows that ROC curve is an effective approach for performance evaluation and optimization of SVM.

Keywords Pattern recognition, Support vector machine, Parameter optimize, ROC curve

以神经计算为代表的黑箱算法被广泛应用于模式识别领域,其中,基于统计学习理论(Statistical Learning theory, SLT)的支持向量机(Support Vector Machine, SVM)成为小样本机器学习领域的研究热点[1,2]。该理论提出结构风险最小化(Structural Risk Minimization, SRM)规则,指出在保证经验风险的同时要考虑模型的推广(泛化)能力,理论上保证了SVM算法的先进性,在此框架下所实现的SVM模型具有稀疏性[3]。

在SVM算法应用领域,算法性能评估指标尚未有金标准(Gold Standard)[4],常用指标有分类准确度(Classification Accuracy)、精确度(Precision)、检测概率(Probability of Detection)、混淆矩阵(Confusion Matrix)等,此类指标在数据样本类别分布不均或者分类错误代价分布不均情况下,评估效果不佳[5]。受试者操作特性曲线(Receiver Operating Characteristic, ROC)是目标算法性能的二维直观描述,其分析曲线不敏感于类别分布、类别先验概率和错误代价,具有直观性和可理解性,有效克服了上述指标的缺陷。ROC曲线评估方法在生物统计、决策分析和实验医学等领域得到广泛应用[6]。Sono[5],Bradley[7]通过理论和比较实验证明了ROC曲线下面积(Area Under the ROC, AUC)指标比正确率更适合作为模式识别性能的评价标准。本文通过研究ROC曲线在机器学习模型性能评估中的应用,探讨了SVM建模参数优化过程中,ROC曲线最佳工作点(Optimal Operating Point, OOP)[8]对于最优点的选择问题。

1 SVM分类算法

支持向量机方法的原理可以描述为:寻找一个满足分类条件的分类平面,并使训练集中的点距离该分类平面尽可能远,该平面定义为"最优超平面",算法实现过程中遵循结构风险最小化原则。

1.1 SVM分类算法

SVM分类算法基于二分类结构,考虑二分类样本集{+1,−1}的分类问题,给定数据集(x_i, y),$i=1,2,\cdots,N$,$x_i \in \mathbb{R}^N$,$y \in \{+1,-1\}$,分类超平面为:

到稿日期:2009-09-02 返修日期:2009-11-27 本文受国家高技术研究发展计划(863计划)(2006AA12A108)和国家自然科学基金(60879008)资助。

王旭辉(1979−),男,博士,主要研究方向为智能算法、海量数据挖掘、智能诊断等,E-mail:wangxh@mail.castc.org.cn;舒 平(1968−),男,研究员,主要研究方向为数据建模、事故决策系统;曹 立(1967−),男,博士,副教授,主要研究方向为进化算法、生物医电、智能模型等。

$$w \cdot x+b=0 \tag{1}$$

训练集的最优线性分类器是指能够将两类无误地分离，构造最优超平面的过程转化为求解如下问题：

$$\text{Min } \Phi(w)=\frac{1}{2}||w||^2+C\sum_{i=1}^{n}\xi_i \tag{2}$$

$$\text{Subject to } \begin{cases} y_i[w^T K(x_i)+b] \geqslant 1-\xi_i \\ i=1,2,\cdots,N \end{cases}$$

式中，误差变量 ξ 为允许错分的松弛变量，通过惩罚系数 c 控制对超出精度的样本的惩罚程度，核函数 $K(\cdot)$ 用以实现非线性样本的线性分类，对于该优化问题，定义 Lagrange 函数：

$$\text{Min } \Phi(w,b,a)=\frac{1}{2}||w||^2+\frac{1}{2}C\sum_{i=1}^{n}\xi_i^2-C\sum_{i=1}^{N}a_i\{y_i[(w \cdot K(x))+b]-1+\xi_i\} \tag{3}$$

$$\text{Subject to } \begin{cases} a_i \geqslant 0, i=1,2,\cdots,N \\ \sum_{i=1}^{N} a_i y_i=0 \end{cases}$$

a 为 Lagrange 乘子。对 w,b 求偏导，得到最优超平面，由下式给出：

$$\begin{cases} w^* = \sum_{i=1}^{N} a_i y_i x_i \\ b^* = -\frac{1}{2} w^* \cdot (x_r + x_s) \end{cases} \tag{4}$$

式中，x_r 和 x_s 为满足条件 $a_i \geqslant 0$ 的任何支持向量，从而分类器可表示为：

$$f(x)=\text{sign}(w^* \cdot x + b^*) \tag{5}$$

式中，w^* 是最优超平面的法向量，b^* 是该平面的偏移量，$f(x)$ 的正负即可判定未知样本集 x 所属类别。

1.2 建模参数

支持向量机建模过程中，其核函数的类型、核函数参数 σ 和惩罚系数 c 的选取决定模型的识别性能，建模过程中多采用基于启发式的交叉验证法。

核函数确定了数据在映射到高维空间之后的分布形态，核函数参数 σ 判定了特征空间中向量间归一化的欧氏距离，σ 的选择与学习样本输入空间的范围或宽度有关，输入空间范围越大，则 σ 取值越大，反之则越小。惩罚系数 c 是模型复杂度和训练错误率之间的折中。在建模过程中，无法得到推广能力估计值与这些参数的显式表达关系，且变化不连续，因此不能使用牛顿法、最陡梯度法等经典优化算法。

2 ROC 曲线及其评估应用

模式识别领域，对于所用算法性能的评估是衡量方法可用性和有效性的必经过程。选取 UCI 机器学习标准库中“Adult Database”，数据集意义是根据对象的 14 个特征属性来识别对象所属收入群体，群体类别以 50k/year 为阈值点，分为两类，分别选取人工神经网络与支持向量机作为识别算法进行评估比较。

2.1 ROC 曲线的绘制

ROC 曲线涉及 FPR 与 TPR 两个指标，FPR 即负例预测错误的数量与所有负例的比值，也叫错报率，反映模型的特异性；TPR 即正例预测正确的数量与所有正例的比值，也叫命中率，反映模型的灵敏度。

ROC 曲线以误检率（False Positive Rate，FPR）为 X 轴（即特异性 1-specificity），以检出率（True Positive Rate，TPR）为 Y 轴（即灵敏度 sensitivity），描绘了模型输出的收益和代价之间的关系。其输出正负例数据点的关系可以通过混淆矩阵来表示，图 1 描绘了混淆矩阵中数据点与 3 种评估指标的关系。

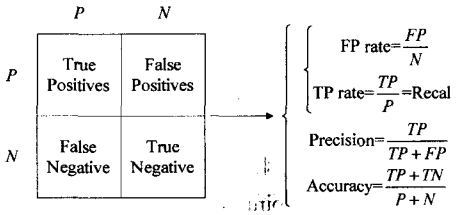

图 1 混淆矩阵及性能评估指标关系

混淆矩阵中，主对角线数值代表正确的分类（检出）输出点，反对角线数值代表分类识别过程中的错误（误检）输出点。

2.2 应用 ROC 曲线评估模型性能

对于指定的分类模型，其阈值的变化直接反映至模型的输出单元，由此可在 ROC 曲线坐标中描绘表征模型灵敏度与特异性的数据点，绘制 ROC 曲线，实现模型性能的比较。

针对所选 UCI 库中，收入群数据库的分类问题，分别选取训练集为 200，测试集为 80，建立 RBF 人工神经网络（RBFN）与支持向量机（SVM）两种模型。

RBFN 模型结构为 14-8-1，训练过程中，其可调阈值参数为径向基函数分布常数 c，从 1 至 5 逐步调整该参数值。建立模型后，获取模型对于 80 组测试数据集的分类输出，计算相应数据点的 TPR，FPR 数值，绘制出 $ROC_{\_RBFN}$ 曲线，如图 2 中虚线。

SVM 模型的可调阈值参数为核参数 σ，从 1 至 20 调整该参数范围，建立相应的模型，并计算当前阈值参数下 80 组测试集的分类输出，在同一坐标系中描绘出 $ROC_{\_SVM}$ 曲线，如图 2 中实线。

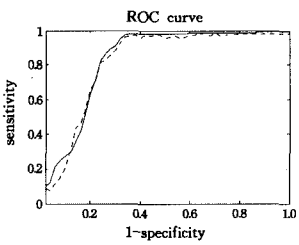

图 2 RBFN 及 SVM 模型的 ROC 曲线

从图 2 可知，$ROC_{\_SVM}$ 曲线具有更好的上凸性，其下面积要大于 $ROC_{\_RBFN}$ 曲线，用非参数法估计两种 ROC 曲线 AUC 值，作为量化指标衡量模型的性能，$ROC_{\_SVM}$ 曲线下面积为 0.89，$ROC_{\_RBFN}$ 曲线下面积为 0.83，SVM 模型对于 80 组测试样本具有更好的检出率和误检率。

利用 ROC 曲线评估模型性能，其实质是目标模型对于集数据的识别过程中，不同的模型参数所对应的模型结构相异，其识别输出不同，计算所得 TPR，FPR 也不同，而指标 FPR 和 TPR 对于模式识别问题的先验概率具有不敏感性，因此 ROC 曲线具备鲁棒性。可见，具有更好上凸特性的 ROC 曲线所对应的目标模型具有更好的识别性能，因此，ROC 曲线能够合理地评估模式识别模型对于未知集数据的识别性能。

3 建模参数最佳工作点

在评估两种模型性能的过程中，通过阈值参数变化所描绘的 ROC 曲线给出了性能比较结果，但并未给出阈值参数最

优点。因此，需要选择 ROC_{SVM} 曲线最佳工作点来确定 SVM 模型的阈值参数，进而确定 SVM 模型。

3.1 OOP 寻优原理

对于所研究的 SVM 模型，虽然从 ROC 曲线分析可知其性能优于 RBFN，但由于 ROC 曲线包含了选定阈值范围所有敏感度和特异度的组合，会存在 TPR_{SVM} 低于 TPR_{RBFN} 且 FPR_{SVM} 高于 FPR_{RBFN} 的阈值点，在此范围内 SVM 模型的性能要低于 RBFN 模型，因此，需要通过对 ROC 曲线最佳工作点(optimal operating point，OOP)的检测，来确定模型的阈值点。可见，基于 OOP 的寻优实质是将 OOP 的选择与模型的优化相结合，选定模型具有最佳 TPR 和 FPR 组合点的阈值。

ROC 曲线中确定 OOP 的方法主要有如下 3 种：

1)设置 FPR 最高限值，寻找低于该限值，而同时使 TPR 最大的切点(cutoff point)，即为 OOP。同理，若需要较高的 TPR，则可以通过设定 TPR 最低限值，检验 FPR 值来确定 OOP。

2)当识别系统正负例误检代价相近时，就要求敏感度和特异度都比较大的切点是最好的 OOP[9]，即 ROC 曲线最左上方的点。常选择灵敏度、特异度平方和最大的点作为 OOP[8,10]。

3)引入误检的代价、受益因子来计算 OOP 斜率[11,12]，该方法将 OOP 的统计学特性与实际应用相结合，选择具有特定斜率值的点为 OOP。

本实例所采取的 UCI 样本集对于误检代价相同，因此，采用第二种方法，通过选择 TPR 和 FPR 平方和最好的组合点来确定 OOP，以优化 SVM 模型的参数。

3.2 SVM 建模参数最佳工作点

对于 SVM 算法的建模识别过程，其可变阈值参数为核参数 σ 和惩罚因子 c，两个阈值参数变化独立于模型的输出，分别绘制其 ROC_σ 和 ROC_c 曲线，如图 3 所示。

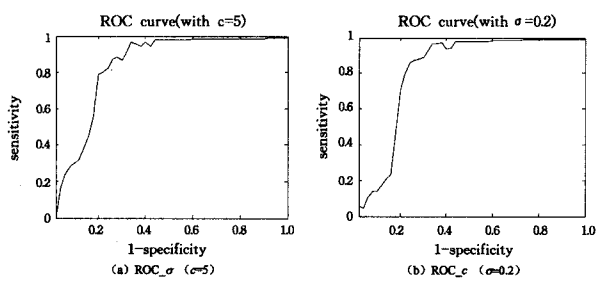

图 3 SVM 模型阈值参数 σ 及 c 的 ROC 曲线

从图 3 可知，核参数 σ 在其所选范围内，SVM 模型输出的敏感度与特异度变化波动较大，SVM 模型性能敏感于核参数的变化，由平方和法计算得 $\sigma=0.7$ 为该阈值 ROC 曲线的最佳工作点。惩罚因子 c 在其优化范围内，对于 SVM 模型误检率的影响较为平缓，由平方和法确定其最佳工作点 $c=11$，达到该点之后，阈值的增加对于模型检出率的贡献并不明显，却导致误检率的增加。两次 OOP 方法优化 SVM 模型后的参数如表 1 所列，γ 为 SVM 模型软间隔，Nsv 为支持向量数，CA 为分类精度。

表 1 基于 OOP 方法优化的 SVM 模型参数

	σ	c	γ	Nsv	TPR/%	FPR/%	CA/%
SVM	0.2	5	0.6	15	81.7	19.8	80.1
SVM_σ	0.7	5	0.8	13	85.3	17.4	83.8
SVM_c	0.2	11	1.1	12	89.6	16.3	84.5
SVM_σ,c	0.7	11	1.3	9	92.1	11.7	94.7

表 1 中，两个参数优化后所得 SVM_σ,c 模型对于测试样本集具有很好的灵敏度、特异度和识别率，其识别性能有明显的提高，具有最优性能。可见，对 SVM 模型的建模参数进行 OOP 优化，寻求其最佳工作点，获得的最终模型 SVM_σ,c 是可行且有效的。同时，由于 ROC 曲线的先验知识的不敏感性，使得由 OOP 方法优化所得的模型具备稳健性特征，具有统计学意义，这与 SVM 建模原理所强调的泛化性是一致的。

结束语 SVM 模型性能的评估和模型参数的优化是该算法领域内的研究热点，本文研究了 ROC 曲线在该研究领域的应用。

1. 利用 ROC 曲线评估 SVM 模型与 RBF 神经网络模型的识别性能，进行了曲线分布的直观分析和曲线下面积的定量分析，给出了 SVM 模型对于所选数据集识别能力优越性的结论，证明了 ROC 曲线在模型性能评估应用中的有效性。

2. 就 SVM 参数优化问题，尝试着利用 ROC 曲线中的最佳工作点进行优化，由工程实例的识别效果比较可知，该优化方法有效提高了 SVM 模型的识别性能，其统计性特征决定了该优化方法具有推广性。

参 考 文 献

[1] 许建华，张学工，李衍达. 支持向量机的新发展[J]. 控制与决策，2004，19(5)：481-484

[2] 刘向东，朱美琳，陈兆乾，等. 支持向量机及其在模式识别中的应用[J]. 计算机科学，2003，30(6)：113-117

[3] Vapnik V N. An Overview of Statistical Learning Theory[J]. IEEE Transactions on Neural Networks，1999，10(5)：988-999

[4] Cherkassky V，Ma Yunqian. Practical selection of SVM parameters and noise estimation for SVM regression [J]. Neural Networks，2004，17(1)：113-126

[5] Son H K，Yun M J，Jeon T J，et al. ROC analysis of ordered subset expectation maximization and filtered back projection technique for FDG-PET in lung cancer [J]. IEEE transactions on Nuclear Science，2003，50(1)：37-41

[6] 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(2)：171-186

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

[8] Robert J G，et al. Determination and interpretation of the Optimal Operating Point for ROC Curves derived through generalized linear models [J]. Understanding Statistics，2003，2(4)：219-242

[9] Riddle D L，Stratford P W. Interpreting validity indexed for diagnostic tests：An illustration using Berg Balance Test [J]. Physical Therapy，1999，7(9)：939-948

[10] Peng M S，So T S H. Logistic regression analysis and reporting：A primer [J]. Understanding Statistics，2002，1(3)：1-7

[11] Hampern E J，et al. Comparison of receiver operating characteristic curves on the basis of optimal operating point[J]. Academic Radiology，1996，7(3)：245-253

[12] Metz C E，Herman B A，et al. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data [J]. Statistics in Medicine，1998，17：1033-1053