最大化约束密度单类分类器

赵加敏[1] 冯爱民[1] 陈松灿[1,2] 潘志松[3]
(南京航空航天大学计算机科学与技术学院 南京 210016)[1]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]
(解放军理工大学指挥自动化学院 南京 210007)[3]

摘 要 针对单类分类器设计中的密度方法,采用以任务为导向的设计思想,通过人为指定核密度估计的密度函数上界,增强了边界低密度区域数据敏感性,同时也有效降低了密度估计的计算复杂度。进一步最大化全体样本的核密度估计函数并采用线性规划,可快速得到相应的稀疏解,因而称之为最大化约束密度单类分类器(Maximum constrained density based one-class classifier,MCDOCC)。为充分利用单类数据中可能出现的极少量异常数据,进一步提出了带负类的最大化约束密度分类器(MCDOCC with negative data,NMCDOCC),通过挖掘异常数据的先验信息来修正仅有正常类的数据描述边界,可提高分类器泛化能力。UCI数据集上的实验结果表明,MCDOCC的泛化能力与单类支持向量机相当,NMCDOCC较之则有所提高,从而能够更高效地估计目标类数据概率密度。

关键词 单类分类器,概率密度估计,最大化约束密度,先验信息
中图法分类号 TP391.4 **文献标识码** A

Maximum Constrained Density One-class Classifier

ZHAO Jia-min[1] FENG Ai-min[1] CHEN Song-can[1,2] PAN Zhi-song[3]
(College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China)[1]
(National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[2]
(Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007, China)[3]

Abstract A novel One-Class Classifier (OCC) was proposed within the framework of probability density estimation called Maximum constrained density based OCC, MCDOCC. By constraining the upper bound of the kernel density estimators with the introduced parameter, MCDOCC is more sensitive in the low-density region located on boundary, alleviates the computation cost at the same time. Then, through maximizing the average constrained density of the target data, MCDOCC optimizes the object function with linear programming and the sparse solution can be reached finally. To further improve the generalization ability, two ways for MCDOCC with Negative data (NMCDOCC) were developed for full utilizing the prior knowledge existed in outliers. Experimental results on UCI data sets show that the generalization ability of MCDOCC is comparable with one-class support vector machines, but NMCDOCC is better than it.

Keywords One-class classifier, Probability density estimation, Maximum constrained density, Prior knowledge

1 引言

密度方法和支持域方法是单类分类器设计的两类主要方法[1]。密度模型[2,3]主要通过估计训练数据密度并设置阈值来判定测试数据是否属于目标类[4]。以非参数化密度估计Parzen窗[5,6]为例,若样本足够多,通过选取相应核函数,理论上可逼近任意分布,然而由于相应的密度估计需要大量训练样本,因此影响了其测试效率。支持域方法OCSVM[7]及SVDD[8,9]则以任务为导向,采用最大化间隔或最小化超球的设计思想,并借助对偶理论和核技巧,得到稀疏解且可应用于非线性问题。不足的是因优化需采用二次规划,故时间复杂度较高[10]。针对上述情形,本文拟将以任务为导向的思想进一步引入到密度单类分类器设计中,通过限制高密度区为指定常数,不仅可有效提高计算效率,同时能够更为关注处于边界的低密度区样本点。为提高测试效率,采用了具有稀疏解的核密度估计模型;进一步,考虑到具体应用中少量异常样本的存在,设计出带负类的最大化约束密度分类器,挖掘异常数据的先验信息以获得更准确的密度估计。

2 基于最大化约束密度的单类分类器

分类器设计中应避免密度估计过程与分类任务脱离,强调密度估计与分类目标的整体性,使其为当前的分类任务服

到稿日期:2013-05-20 返修日期:2013-08-11 本文受国家自然科学基金重点项目(61035003)资助。
赵加敏(1986—),女,硕士生,主要研究方向为模式识别、异常检测;**冯爱民**(1971—),女,副教授,硕士生导师,主要研究方向为机器学习、数据挖掘及异常检测,E-mail:amfeng@nuaa.edu.cn;**陈松灿**(1962—),男,教授,博士生导师,主要研究方向为模式识别、机器学习、神经计算;**潘志松**(1973—),男,教授,主要研究方向为网络安全、模式识别。

务。基于此原则，提出了最大化约束密度单类分类器(Maximum Constrained Density One-Class Classifier：MCDOCC)，并进一步推广出含少量异常的情形(Negative Maximum Constrained Density One-Class Classifier：NMCDOCC)。

2.1 最大化约束密度单类分类器

密度估计需为分类任务服务，为使密度模型适合分类任务，则需类条件概率密度估计尽量可判别。对于单类问题而言，由于负类样本缺失，单类分类器仅能利用正常类样本，此可通过最大化全部训练样本的平均概率密度估计值优化目标类的条件概率密度估计。

进一步，为提高目标类条件概率密度估计的判别性，在目标类数据分布的低密度区，即不易分类的区域，目标类数据的密度估计均值应尽量大，而对于高密度区，即比较易于分类的区域，可将估计值简化为一个介于 0 和 1 间的常数，该常数须足够大以保证分类器具有良好的判别性能。故定义如下约束密度：

$$constrained\_den_x(x)=\min\{\hat{p}_x(x),\gamma_{\max}\} \tag{1}$$

式中，$\hat{p}_x(x)$为目标类条件概率密度估计，常数 $\gamma_{\max}>0$ 为约束密度上界。对单类问题，经验约束密度如下：

$$\hat{D}=\frac{1}{n}\sum_{i=1}^{n}constrained\_den_x(x_i) \tag{2}$$

若采用核密度估计[11,12]模型，且假设训练集由 n 个 d 维目标类样本组成，记为 $X_t=\{x_i\,|\,x_i\in R^d\}_{i=1}^n$，则上述目标类条件概率可表示为：

$$\begin{aligned}&\hat{p}_x(x)=\sum_{k=1}^{n}\alpha_k K(x,x_k)\\&\text{s. t.}\sum_{k=1}^{n}\alpha_k=1,\alpha_k\geqslant 0\end{aligned} \tag{3}$$

式中，权重因子 α_k 用于衡量训练样本 x_k 对密度函数的贡献，K 为归一化核函数。

最大化上述式(1)、式(2)所对应的经验约束密度可进一步描述为：

$$\begin{aligned}&\max\sum_i(\gamma_i+\varepsilon\delta_i)\\&\text{s. t.}\begin{matrix}\hat{p}_x(x_i)-\gamma_i-\delta_i\geqslant 0\\ \gamma_i\leqslant\gamma_{\max},\delta_i\geqslant 0\end{matrix},i=1,2,\cdots,n\end{aligned} \tag{4}$$

式中，γ_i 用于度量式(1)所对应的约束密度函数值 $constrained\_den_x(x_i)$，同时注意到当大量样本的约束密度等于 $\gamma_{\max}$ 时，最大化约束密度可能产生奇异解，因此增加参数 δ_i 来度量密度估计大于上界 $\gamma_{\max}$ 的部分，以避免奇异解。将式(3)代入其中，可知式(4)为关于参数 α 的线性规划，且优化结果表明仅有少数 α_k 值不为 0，即具有稀疏性。

对于未知的测试样本 z，计算其概率密度，然后根据低密度拒绝的策略判别其是否属于目标类，即大于指定阈值的样本视为目标类而被接受，而小于该阈值的样本则视为异常而被拒绝。测试样本 z 的密度函数如下：

$$p_x(z)=\sum_{k=1}^{n}\alpha_k K(z,x_k) \tag{5}$$

值得指出，虽然式(5)与 Parzen 窗判别形式相近，但由于 α_k 解的稀疏性，当训练样本数量较大时其计算时间复杂度优于 Parzen 窗。

2.2 带负类的最大化约束密度单类分类器(NMCDOCC)

上述最大化约束密度的单类分类器 MCDOCC 仅利用了正常样本，然而在包括垃圾邮件检测及医疗诊断在内的众多实际应用中，仍能获取到少量异常数据。若能充分挖掘其中所包含的先验知识，有望进一步提高分类器泛化能力[13]。然而，由于负类数据数量极少以致通常不足以反映其分布趋势，因此无法准确估计相应的概率密度函数。尽管如此，仍可以利用其来修正目标类概率密度估计，究其原因可采用图示说明之。

图 1 示意了仅考虑正类数据及存在少量异常数据两种情况下所获得的正常类概率密度估计。为方便表示，将密度估计简化为单峰形式，图中正类数据('*')呈单簇分布，少量异常类数据('·')均匀分布于单簇数据周围，且两类数据存在重叠区域。对 MCDOCC 而言，训练阶段仅有目标类样本可利用，然而通常由于样本数量有限或者抽样不均匀等因素，致使当前目标数据无法保证覆盖到目标类的真实分布空间，因而致使所获得的密度估计如图中$\hat{p}_2(x)$所示。考察图中某异常样本 y_j，可发现其相对该估计的概率值较大，因而易被错判为目标类。若利用少量异常数据可得到图中$\hat{p}_1(x)$所示的密度估计，异常样本 y_j 因对应的概率密度较之$\hat{p}_2(x)$显著降低而更易被正确的识别拒绝。进一步观察，可发现绝大多数正常类数据具有较高的估计值，且负类数据获得较小的概率估计，因而达到修正密度估计的目的。

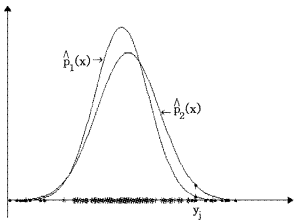


图 1 仅有目标类数据及存在少量异常数据的概率密度估计

受上述分析的启发，为取得更合理的正常类概率密度估计，可充分利用有限的异常数据所提供的先验知识，从而达到进一步提升分类器泛化能力的目的。这里，根据异常数据的利用方式，提出了两种不同的方案，分别为带负类的最大约束密度分类器 I(NMCDOCC_I)及带负类的最大约束密度分类器 II(NMCDOCC_II)，并分别介绍如下。

2.2.1 NMCDOCC_I

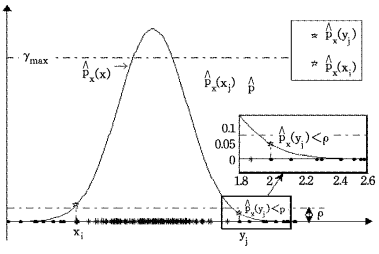


图 2 带负类的最大约束密度单类分类器 NMCDOCC_I 图示

假设训练阶段存在少量异常样本 $y_j\in R^d, j=1\cdots m$，$\hat{p}_x(x)$为目标类条件概率密度估计，则异常样本 y_j 相对目标类的密度函数值可表示为$\hat{p}_x(y_j)$，异常样本通常分布于正常类样本分布较稀疏的边界区域或者远离正常样本，如图 2 所示，其中'*'代表目标类数据，而'·'代表少量异常数据，简化起见，密度估计仍简化为单峰形式。为使分类器具有良好的判

别能力，相对正常类数据而言，一个合理的概率密度估计通常应满足如下前提：异常样本 y_j 相对目标类概率密度估计值 $\hat{p}_x(y_j)$ 较小甚至趋于 0。这里不妨令估计值 $\hat{p}_x(y_j)$ 小于一个正常数 ρ，通过合理地选择该常数，使得异常数据获得较小的密度估计值。

存在异常数据的约束密度定义仍如式(1)所示，NMCDOCC_Ⅰ优化目标为：

$$\begin{aligned}&\max \sum_i(\gamma_i+\varepsilon\delta_i)\\&\text{s.t. } \hat{p}_x(x_i)-\gamma_i-\delta_i\geqslant 0,\gamma_i\leqslant\gamma_{\max},\delta_i\geqslant 0,\\&\quad \hat{p}_x(y_j)\leqslant\rho,\\&\quad i=1,2,\cdots,n,j=1,2,\cdots,m\end{aligned}\tag{6}$$

对比仅含正常样本的式(4)，二者的差别在于式(6)考虑了异常信息且设置其相对正常类的概率密度估计值上限，即 $\hat{p}_x(y_j)\leqslant\rho$，进而达到利用异常数据的先验信息的目的。

2.2.2 NMCDOCC_II

由于异常样本通常分布于正常样本分布的低密度区，因此提高低密度区样本的判别性尤为重要。具体而言，异常样本通常分布于正常类样本分布较稀疏的边界区域或者远离正常类样本，因此一个“好”的正常类概率密度估计应该满足正常类数据的概率密度估计值相对较大，而异常数据相对此概率密度估计的值 $\hat{p}_x(y_j)$ 则应尽可能小，且前者大于后者。

分别考察分布于高密度区及低密度区的正常类样本，在正常数据分布的低密度区，正常数据与异常数据的概率密度估计值差别较小，即 $\hat{p}_x(x_i)-\hat{p}_x(y_j)$ 较小，此类区域相对难判别，因而应优化正常类概率密度估计，使得上述差值尽量大；在正常数据分布的高密度区，正常数据与异常数据的概率密度估计值相差较大，在此类区域可将二者密度差简化为一个常数，适当选取该常数，以保证分类器具有良好的判别性能，如图 3 所示。

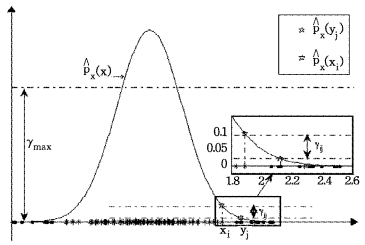


图 3 带负类的最大约束密度单类分类器 NMCDOCC_Ⅱ图示

值得指出，此处约束密度的概念已不同于 MCDOCC 算法，这里的约束密度不再针对正常类概率密度估计而言，而是相对于正常数据与异常数据概率密度估计差值。具体而言，在存在少量异常样本的情况下约束密度重新定义如下，为区别于 MCDOCC，将其记为 $constrained\_den_x'$：

$$constrained\_den_x'(x,y_j)=\min\{\hat{p}_x(x)-\hat{p}_x(y_j),\gamma_{\max}\}\tag{7}$$

式中的 $\gamma_{\max}$ 意义与式(1)有所不同，$\gamma_{\max}$ 转化为正常数据与异常数据概率密度差值的上界，如图 3 所示。相应平均约束密度定义为：

$$D'=\sum_j\int constrained\_den_x'(x,y_j)p_x(x)\mathrm{d}x\tag{8}$$

对于存在少量异常数据的单类问题，上式常通过经验约束密度代替，其计算式如下：

$$\hat{D}_x'=\frac{1}{N_1}\sum_{j=1}^{m}\sum_{i=1}^{n}constrained\_den_x'(x_i,y_j)\tag{9}$$

其中，$N_1=n+m$，即包括少量异常样本在内的全体训练样本数目。

最大化经验约束密度，目标函数如下：

$$\begin{aligned}&\max \sum_{ij}(\gamma_{ij}+\varepsilon\delta_{ij})\\&\text{s.t. } \hat{p}_x(x_i)-\hat{p}_x(y_j)-\gamma_{ij}-\delta_{ij}\geqslant 0,\gamma_{ij}\leqslant\gamma_{\max},\delta_{ij}\geqslant 0,\\&\quad i=1,2,\cdots,n,j=1,2,\cdots,m\end{aligned}\tag{10}$$

其中，参数 δ_{ij} 的作用同式(4)中的 δ_i，即度量概率密度差值大于上界 $\gamma_{\max}$ 的部分，防止奇异解。上式同样为关于参数 α_i 的线性规划问题，且具有稀疏解。

相较 NMCDOCC_I 而言，NMCDOCC_II 更充分地利用了数据的先验信息，然而，由于其优化参数相对较多，因此它在数据量较大的情况下，时间效率不如前者。

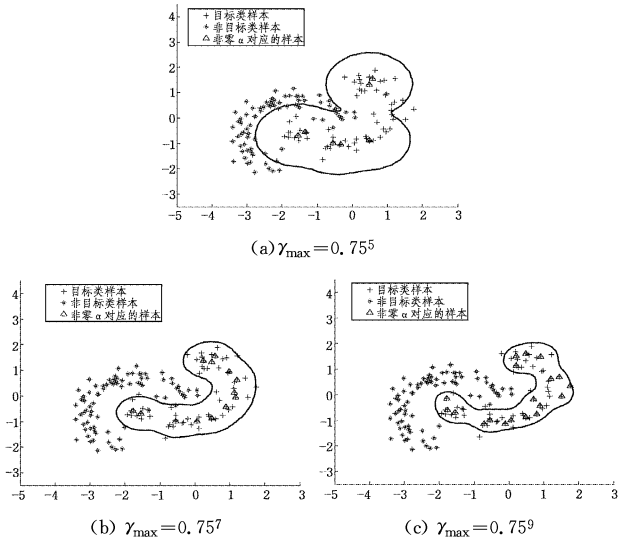
3 实验

实验针对参数对算法的影响及其在 UCI 真实数据集上的泛化能力两方面展开。实验中 Parzen 窗及 MCDOCC 的归一化核函数均采用高斯核。

3.1 参数影响

最大约束密度单类支持向量机有两个超参数，即核带宽参数及约束密度上界 $\gamma_{\max}$。现着重考察参数 $\gamma_{\max}$ 对 MCDOCC 的影响。

1)Toy 问题：实验选取香蕉型数据，其中训练集包含 100 个目标类样本，测试集包含 20 个目标类样本和 100 个非目标类样本。考察核参数 $\gamma_{\max}$ 对 MCDOCC 的影响，采用网格搜索确定，搜索范围为 $2^n(n=-2,-1,0,\cdots,8)$，选取最优值；由于非归一化核函数 $K(x,x)$ 值为 1，因此 $\gamma_i\leqslant 1$，不妨令参数 $\varepsilon=\gamma_{\max}^2$，而 $\gamma_{\max}$ 搜索范围为 $0.75^n(n=1,2,\cdots,9)$。图 4 各子图中实曲线为不同 $\gamma_{\max}$ 值的分类边界。



(a) $\gamma_{\max}=0.75^5$

(b) $\gamma_{\max}=0.75^7$ (c) $\gamma_{\max}=0.75^9$

图 4 参数 $\gamma_{\max}$ 对 MCDOCC 分类结果的影响

观察上述实验结果，当 $\gamma_{\max}$ 相对较大时(见图 4(a))，算法优化获得的非零混合因子较少(由样本点 Δ 支撑)，数据描述边界比较松；随着 $\gamma_{\max}$ 逐渐减小，非零混合因子数逐渐增多，数

据描述边界趋于精确；当 γ_{max} 继续减小时，非零混合因子比例继续增大，目标类数据描述边界更为复杂，出现过拟合趋势。

2)真实数据集：实验选取 UCI 医学检验标准数据集 Sonar 作为实验数据集，图 5 为采用 Fisher 降维后的数据分布可视化图[14]。随机选择 80%的正类数据作为训练样本，剩余的 20%作为测试样本，MCDOCC 全部异常类数据均为测试数据；对 NMCDOCC 算法而言，训练阶段仅有少量异常样本，故实验中异常类样本数量仅占全体训练样本的 5%。

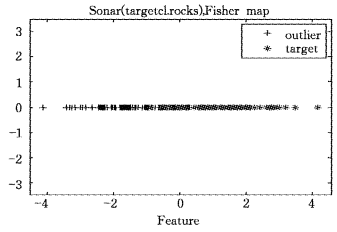


图 5　Fisher 降维后的 Sonar 数据分布可视化

考察 Sonar 数据集上参数 γ_{max} 对 MCDOCC、NMCDOCC_I 和 NMCDOCC_II的影响，核带宽参数采用网格搜索确定，搜索范围为 $2^n(n=-2,-1,0,\cdots,8)$，各自选取最优解作为 σ 值，而参数 γ_{max} 的考察范围为 $0.75^n(n=1,2,\cdots,10)$。图 6 显示了 MCDOCC、NMCDOCC_I 和 NMCDOCC_II 在 Sonar 数据集上的实验结果，展示了参数 γ_{max} 对上述算法的 AUC 值的影响。

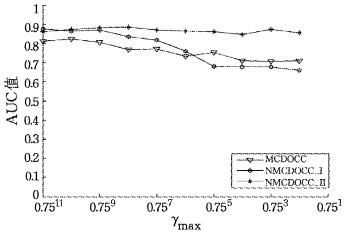


图 6　参数 γ_{max} 对 MCDOCC、NMCDOCC_I 及 NMCDOCC_II 的 AUC 实验结果影响

观察上述实验结果，当参数 γ_{max} 较大时 MCDOCC 及 NMCDOCC_I 算法的 AUC 值相对较小，随着参数的逐渐减小，两算法的 AUC 值增大且最后趋于平稳。NMCDOCC_I 受参数 γ_{max} 的影响较突出，即在核带宽 σ 确定的前提下通过合理地选取约束密度上界 γ_{max} 可使 NMCDOCC_I 获得较好的性能。相比之下，NMCDOCC_II 充分考察了少量异常样本的影响，因此具有较好的推广能力且表现更稳定。

3.2　实验设置与结果

1. 实验数据集

为验证算法的性能，在真实数据集上进行实验，实验中采用 8 个 UCI 医学检验标准数据集，依次为 Biomed、Import、Heart、Ionosphere、Sonar、Iris、Hepatitis 及 Liver。对于每个数据集，选取正常或健康模式作为目标类(正类)，异常或病态模式作为异常类(负类)。

为保证目标类数据具有代表性，随机选择 80%的正类数据作为训练样本，剩余的 20%的目标类数据作为正类测试数据，Parzen 窗/OCSVM/MCDOCC 选取全部异常类数据作为负类测试数据；对 NMCDOCC 算法而言，训练阶段仅有少量异常样本，故实验中负类样本数量仅占全体训练样本的 5%，即正常类训练样本数与异常类训练样本数之比为 95∶5。

具体如表 1 所列，表中从左到右各列信息依次表示实验数据集(Datasets)，数据维数(D)，目标类/异常类样本总数(＃Pos/＃Neg)，针对 MCDOCC 的目标类训练样本(＃TrPos)、目标类/异常类测试样本数(＃TePos/＃TeNeg)，针对 NMCDOCC 算法的目标类/异常类训练样本数(＃TrPos/TrNeg) 以及目标类/异常类测试样本数(＃TePos/＃TeNeg)。

表 1　实验中采用的数据集及其训练集/测试集数目分布

Datasets	D	＃Pos/＃Neg	Parzen 窗/OCSVM/MCDOCC		NMCDOCC	
			＃TrPos	＃TePos/＃TeNeg	＃TrPos/＃TrNeg	＃TePos/＃TeNeg
Biomed	5	127/67	102	25/67	102/5	25/62
Heart	13	164/139	123	41/139	123/7	41/132
Import	25	88/71	71	17/71	71/4	17/67
Ionosphere	34	225/126	180	45/126	180/10	45/116
Sonar	60	111/97	89	22/97	89/5	22/92
Iris	4	100/50	80	20/50	80/4	20/46
Hepatitis	8	123/32	98	25/32	98/5	25/27
Liver	6	200/145	160	40/145	160/8	40/137

2. 对比实验及相关参数设置

表 2　Parzen 窗/OCSVM/MCDOCC/NMCDOCC 的 gmeans±var 及 Sparseness(%)比较

数据集	Parzen 窗	OCSVM		MCDOCC		NMCDOCC_I		NMCDOCC_II	
	Gmeans ±var	Gmeans ±var	Sparse-ness(%)	Gmeans ±var	Sparse-ness(%)	Gmeans ±var	Sparse-ness(%)	Gmeans ±var	Sparse-ness(%)
Biomed	0.9425 ±0.0012	0.9133 ±0.0016	15.69	0.9454 ±0.0012	16.96	0.9527 ±0.0009	32.62	0.9544 ±0.0009	1.20
Heart	0.9169 ±0.0065	0.9003 ±0.0085	14.77	0.9325 ±0.0061	2.50	0.9362 ±0.0058	2.50	0.9306 ±0.0061	2.50
Import	0.8549 ±0.0176	0.8877 ±0.0120	13.94	0.8704 ±0.0129	9.72	0.8898 ±0.0103	34.85	0.9061 ±0.0059	3.96
Ionos-phere	0.9309 ±0.0002	0.9226 ±0.0010	16.78	0.9398 ±0.0006	10.56	0.9415 ±0.0002	16.06	0.9577 0.0003	10.06
Sonar	0.8687 ±0.0592	0.8878 ±0.0162	16.18	0.8802 ±0.0539	1.24	0.9077 ±0.0219	2.91	0.9188 ±0.0129	8.91
Iris	0.9267 ±0.0016	0.9225 ±0.0028	15.75	0.9415 ±0.0019	4.00	0.9511 ±0.0017	5.47	0.9477 ±0.0011	4.10
Hepa-titis	0.8825 ±0.0397	0.8846 ±0.0268	16.87	0.8878 ±0.0557	22.02	0.8953 ±0.0335	9.80	0.9019 ±0.0264	2.44
Liver	0.8712 ±0.0556	0.8708 ±0.0421	13.56	0.8732 ±0.0549	10.63	0.8939 ±0.0361	3.09	0.8980 0.0230	5.19

一方面，实验中分别以非参数密度估计 Parzen 窗、支持域方法 OCSVM 作为 MCDOCC 的对比算法，以 *gmeans* 评价指标作为衡量标准比较三者的泛化能力，$gmeans = \sqrt{TPR * TNR}$，主要考察分类结果中正负类数据被正确识别的比率，因而该指标值介于 0～1 之间且越接近 1 越好。考虑到 OCSVM 及 MCDOCC 具有稀疏解，实验中以支持向量(或非零权重数)占全部训练样本的比率作为稀疏性的衡量标准，考察二者稀疏性。另一方面，验证少量异常样本对算法的影响，进一步对比 MCDOCC 及带负类的最大约束密度分类器的推广能力。

表 2 所列为各算法重复 10 轮的 gmeans 指标的均值及方差结果。试验中各对比算法均针对实验数据集的目标类训练数据进行样本中心化处理，且通过 5-fold 交叉验证选择各参数值：

1)高斯核带宽参数 σ 搜索范围取 2^t，其中 $t\in[-2,10]$，搜索步长为 1。

2)MCDOCC、NMCDOCC_I 及 NMCDOCC_II 选取 $\gamma_{max}=0.75^n(n=1,2,\cdots,9)$，参数 $\varepsilon=\gamma_{max}^2$。

3) NMCDOCC□参数 ρ 和参数 γ_{max} 值存在密切联系，因而试验中设置 $\rho=\gamma_{max}/8$。

3.3 实验分析

(1)比较 Parzen 窗、OCSVM 和 MCDOCC 算法在实验数据集上的性能，MCDOCC 算法的 gmeans 指标在 Heart 和 Iris 数据集上优于 Parzen 窗、OCSVM，而在其余 6 个数据集上，其分类性能与 Parzen 窗及 OCSVM 相当。

(2)与 OCSVM 相似，MCDOCC 具有稀疏解。在 Biomed、Hepatitis 数据集上 OCSVM 的稀疏性占优，而在其余数 6 个实验据集上 MCDOCC 的稀疏性均优于 OCSVM，在 Heart、Sonar、Iris 上稀疏性尤为显著。

(3)与未考虑少量异常样本的 MCDOCC 算法相比，NMCDOCC_I、NMCDOCC_II 在除 Heart 数据集以外的 7 个数据集上泛化能力均优于前者。且 NMCDOCC_II 在全部实验数据集上具有显著的稀疏解。

(4)在 Heart 数据集上 NMCDOCC_I、NMCDOCC_II 与 MCDOCC 算法推广能力相当，分析其原因可能为：一方面，异常数据数量极少，因而先验信息不足；另一方面，由于缺乏针对单类问题的专用实验数据集，试验中采用经典的两类数据集，在某种程度上并不适合单类问题中未知异常类样本均匀分布于正常类样本以外的全部空间的要求，因此可能致使用于训练的少量异常数据采样自身不具代表性，进而降低修正能力。

结束语 在概率密度模型的框架内，坚持以任务为导向的分类器设计思想，提出了最大化约束密度的单类分类器 MCDOCC，该算法最大化经验约束密度函数，通过线性规划求解目标函数。实验结果表明，在 8 个 UCI 实验数据集上 MCDOCC 与 Parzen 窗及 OCSVM 推广性能相当，且具有稀疏解。上述算法进一步挖掘少量异常数据的先验信息，提出了两种存在少量异常数据的单类分类器算法，从而为异常检测提供了新方法。

参 考 文 献

[1] 陈斌，陈松灿，潘志松，等. 异常检测综述[J]. 山东大学学报，2009，39(6)：13－23

[2] Bishop. Neural networks for pattern recognition[M]. London：Oxford University Press，1995

[3] Sain S R，Gray H L，Woodward W A，et al. Outlier detection from a mixture distribution when training data are unlabeled [J]. Bulletin of the Seismological Society of America ，1999，89(1)：294-304

[4] Tarassenko L，Hayton P，Brady M. Novelty detection for the identification of masses in mammograms[C]// Fourth International Conference on Artificial Neural Networks. 1995. London：University of Cambridge，June 1995

[5] Duda R O，Hart P E，Stork D G. Pattern Classification(2nd Edition ed)[M]. New York：John Wiley & Sons，2001

[6] Yeung D Y，Chow C. Parzen-window network intrusion detectors [C]// Proceedings in the 16th International Conference on Pattern Recognition (ICPR'02). 2003

[7] Schölkopf B，Platt J C，Shawe-Taylor J. Estimating the support of a high-dimensional distribution [J]. Neural Computation，2001，13(7)：1443-1471

[8] Tax D，Duin R P. Support vector domain description[J]. Pattern Recognition Letters，1999，20(11/13)：1191-1199

[9] Dolia A，et al. Kernel ellipsoidal Trimming[J]. Computational Statistics and Data Analysis，2007，52(1)：309-324

[10] 冯爱民，陈松灿. 基于核的单类分类器研究综述[J]. 南京师范大学学报，2008，8(4)：1-6

[11] Xu Miao，Rahimi A，Rao R P N. Complementary kernel density estimation[J]. Pattern Recognition Letters，2012，33(10)：1381-1387

[12] Meinicke P，Twellmann T，Ritter H. Maximum contrast classifiers [C]// Proceedings of ICANN' 02 Proceedings of the International Conference on Artificial Neural Networks. London，UK：Notes In Computer Science (LNCS) Press，2002：745-750

[13] Tax D，Duin R P. Support vector domain description[J]. Machine Learning，2004，54：45-66

[14] http://homepage. tudelft. nl/n9d04/occ/index. html