基于核的单类分类器研究

冯爱民, 陈松灿

(南京航空航天大学 信息科学与技术学院, 江苏 南京 210016)

[摘要] 以统计学习理论为背景, 以核方法为基础的两类典型单类分类算法: 单类支持向量机(OCSVM)和支持向量数据域描述(SVDD), 均以降低 VC维为目标, 其中前者通过寻找一个远离原点的超平面, 使目标数据所在的正半空间尽量最小; 而后者通过寻找一个包含大部分目标数据的最小超球, 实现体积最小化. 围绕上述两算法, 已有大量改进形式出现. 本文以此为主线, 分别从模型构建、模型改进和数据预处理的角度, 进行了回顾和阐述, 并对各算法的特点给出了相应的总结.

[关键词] 核方法, 单类分类器, 单类支持向量机, 支持向量数据域描述

[中图分类号] TP 391 4 [文献标识码] A [文章编号] 1672-1292(2008) 04-0001-06

Study on One-Class Classifiers Based On Kernel Method

Feng Aimin, Chen Songcan

(College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract As state-of-the-art algorithms based on kernel method, one-class SVM (OCSVM) and Support Vector Data Description (SVDD) root into the sound theoretical basis of statistical learning theory. In order to decrease the VC dimension for promoting the generalization ability, OCSVM tries to find a hyperplane with the furthest distance to the origin for minimizing the positive half space lived by most of the target data; While SVDD tries to find the minimal volume hypersphere enclosing most of the given samples. Focusing on the two algorithms, some variants or improved versions are proposed to avoid some disadvantages of the above models. In this paper, we review most of these variants and give a detailed relation among the discussed algorithms to the original models.

Key words kernel method; one-class-classifier; one-class SVM (OCSVM); support vector data description (SVDD)

核方法[1,2]是近年来发展形成的一种新的机器学习方法, 其实质是通过核诱导的隐映射将低维输入空间的非线性问题变换至高维甚至无穷维特征空间中的(近似)线性问题来解决. 由于采用了对偶形式而使数据能以内积形式刻画, 因而可通过核代入最终在特征空间中获得对原非线性问题的解决, 既避免了维数灾难, 又能获得优越的性能. 核方法已广泛应用于监督学习如支持向量分类与回归[3,4]、核判别分析[5]和无监督学习如核主成分分析(KPCA)[6]中. 本文所论的是核方法在无监督学习中的另一重要应用, 即异常检测, 或称为单类问题上的应用.

单类问题, 是指训练样本中只有一类目标数据(也称为正常数据). 其它非目标数据(也称为异常数据或负类样本)由于缺失的原因不同而归为两类: 一类如故障诊断[7]、医疗诊断[8]、网络入侵检测[9]等, 这类问题的异常数据因为鲜有发生或获取的代价过高而造成样本缺失甚至根本没有, 如故障诊断, 不可能为了获取与正常类相当的异常数据而人为地破坏机器; 另一类如目标识别中的人脸检测[10]、图像检索[11]、文本检测[12]等, 这类问题的异常数据虽然容易获得, 但因为类型过多导致根本无法穷尽所有的异常数据而使现有负类样本不具代表性, 如人脸检测, 任何非人脸都可作为非目标数据. 更多的单类应用, 可参考文献[13].

针对单类问题, 单类分类器的设计分训练(学习)和测试两个阶段. 训练阶段仅用提供的目标数据来得到分类器; 测试阶段为比较各单类算法的性能, 一般情况下会提供负类样本. 由此可见, 单类分类器与两

收稿日期: 2008-06-18
基金项目: 国家自然科学基金(60603029和 60703016)资助项目.
通讯联系人: 陈松灿, 教授, 博士生导师, 研究方向: 人工智能、神经网络和模式识别. E-mail: s_chen@nuaa.edu.cn

类分类器最大的不同,就是训练过程没有负类样本,所以说单类分类器的主要目的是对目标数据高密度区的估计,主要用于描述数据而不是像两类那样判别数据[14].因此不难理解,单类分类器的主要任务是识别正常类和拒绝异常类[15],有时也称为数据描述[16]或异常检测[10].

1　基于核的单类分类器模型

在统计学习理论中,小的VC维意味着分类器好的推广能力,而VC维 h $w^2 r^2 + 1$[17]取决于两个因素,其一是分类超平面的法向量 w,由于其长度 $\|w\|$ 与间隔间成反比,最大化间隔就意味着最小化 $\|w\|$,从而减小VC维.其二,降低VC维的另一个途径也可以减小 r,这里的 r就是包含样本数据的最小超球.基于上述原理,自然出现了如下两类单类分类器模型,简洁起见,这里直接描述其核形式.

1.1　单类支持向量机(OCSVM)

Schölkopf等提出的One-Class SVM (OCSVM)[18],便是最大化间隔的单类学习算法.它巧妙地利用了原点作为负类的代表,通过最大化原点和目标数据间的最小欧氏距离 $\frac{\rho}{\|w\|}$ 来寻找最优超平面 $w^{\mathrm{T}}x - \rho = 0$,其中 w 是超平面的法向量,ρ 是超平面截距,使超平面尽量远离原点,从而最小化大部分目标数据所在的正半空间.图1所示的是采用高斯核的OCSVM在2维空间中所找到的最优超平面.

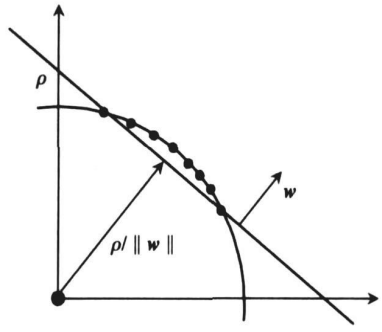

图1　OCSVM取高斯核时找到的最优超平面

Fig.1　The optimal hyperplane of OCSVM with the RBF kernel

具体而言,给定数据集 $X = \{x_i \mid x_i \in R^d\}_{i=1}^n$,为使算法具有一定的鲁棒性引入松弛因子 $\xi = [\xi_1, \ldots, \xi_n]^{\mathrm{T}}$,OCSVM目标函数如下:

$$\begin{aligned} &\min_{w,\xi,\rho} \ \frac{1}{2}w^{\mathrm{T}}w - \rho + \frac{1}{n\nu}\sum_{i=1}^{n}\xi_i \\ &\text{s.t.}\ w^{\mathrm{T}}\varphi(x_i) \geqslant \rho - \xi_i,\ \xi_i \geqslant 0,\ i = 1, \ldots, n. \end{aligned} \tag{1}$$

式中,$\nu \in (0, 1]$ 是所谓的百分比估计,和支持向量个数密切联系,即 ν 是边界支持向量的上界,是全部支持向量个数的下界,称为 ν 属性[19].

引入向量 $\alpha = [\alpha_1, \ldots, \alpha_n]^{\mathrm{T}}$,式(1)的对偶形式表示为:

$$\begin{aligned} &\min \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i, x_j), \\ &\text{s.t.}\ \sum_{i=1}^{n}\alpha_i = 1,\ \alpha_i \geqslant 0,\ i = 1, \ldots, n; \end{aligned} \tag{2}$$

这是一个二次规划问题,可采用经典的二次规划软件包或者序列最小优化算法[20]来优化.

通过上述过程可见,由于OCSVM套用了两类 ν-SVM[19]框架,从而保留了支持向量机解的全局最优和稀疏性.

1.2　支持向量数据域描述(SVDD)

如图2所示,支持向量数据域描述(Support Vector Data Description, SVDD)[21,22]通过寻找一个包含正常数据的最小超球来降低VC维以提高分类器性能.给定数据集,通过一个非线性映射将数据映射到一个高维甚至无穷维的特征空间中,核空间中软间隔的SVDD目标函数如下:

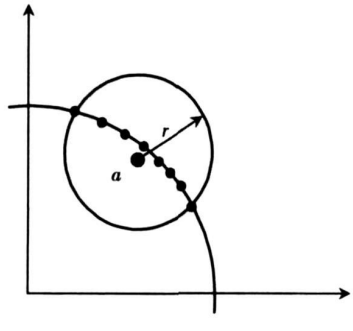

图2　SVDD取高斯核时找到的最小超球

Fig.2　The minimum hypersphere of SVDD with the RBF kernel

$$\begin{aligned} &\min\ r^2 + C\sum_{i=1}^{n}\ _i, \\ &\text{s.t.}\ \|\ (x_i) - a\|^2 \quad r^2 + \ _i,\ _i \quad 0, \quad i = 1, \ , n. \end{aligned} \tag{3}$$

式中, a 表示超球中心, C 是正则化因子.

对偶形式表示为:

$$\begin{aligned} &\max\ \sum_{i=1}^{n}\ _i\boldsymbol{K}(x_i, x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\ _i\ _j\boldsymbol{K}(x_i, x_j), \\ &\text{s.t.}\ \sum_{i=1}^{n}\ _i = 1, \ _i \quad 0, \ i = 1, \ , n; \end{aligned} \tag{4}$$

通过二次规划求得所有样本对应的 $_i$, 则球心可以由非零 $_i$ 的对应样本稀疏表示, 而半径可由任意一个支持向量到球心的距离获得.

需要说明, 尽管 OCSVM 和 SVDD所采用的模型不同, 但两者在高斯核下等价, 具体内容参见[16, 18].

2 改进或变异算法

图 3所示为本节阐述的全部改进算法.

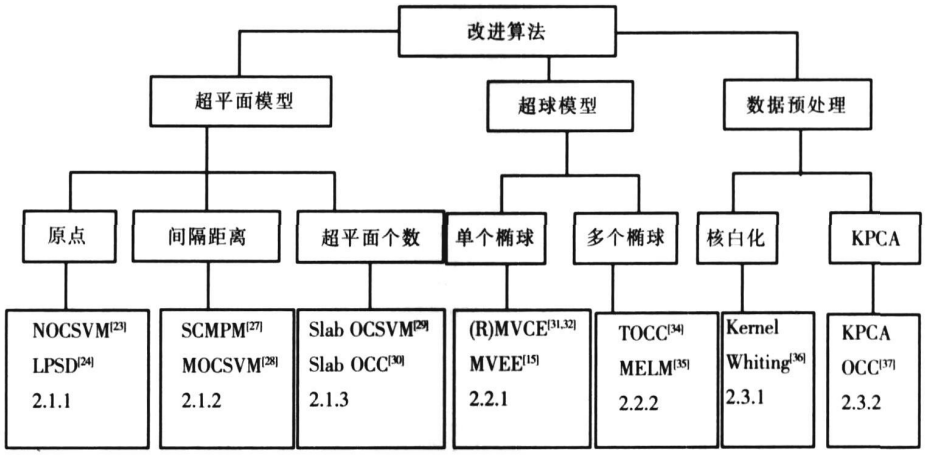

图 3 OCSVM 和 SVDD 的改进算法

Fig.3 The improved/variant algorithms of OCSVM and SVDD

2 1 基于超平面模型

2 1 1 针对原点的改进

OCSVM 采用原点作负类, 事实上已隐含假设了负类的位置[23], 这在一定程度上影响了算法的合理性, 于是出现了相应的改进算法.

(1)假定有部分负类信息. 假定有部分负类数据 (简称为 NOCSVM 算法), 尽管不具有代表性甚至仅仅是人工生成的数据, 但相对于仅考虑目标数据, 充分利用这些信息仍能一定程度上提高单类分类器的性能, 因此这里不再直接以原点作为负类[23], 而是用现有负类数据的中点取代原点. 之所以采用负类中点而不像两类 SVM 那样单独考虑负类的每一个样本, 是因为这里的负类信息并不具有代表性, 然而它的中点毕竟能够提供一部分信息来辅助寻找超平面, 引入核技巧后, 输入空间中的负类中点映射到高维空间中, 并不仅仅只是某个固定的点的映射, 相反却具有了现有全部负类数据的信息, 因此对界定目标数据边界提供了有效的帮助.

(2)线性规划算法. 受 NOCSVM 算法尽量远离负类数据中点的启发, 由 Campbell等提出的线性规划算法 LPSD[24]不再强制地将目标数据之外的任何一点作为参照来设计算法, 而是将目标数据的输出均值作为参照, 在保证大部分目标数据均落入正半空间的前提下, 通过减小全部样本数据输出结果的均值, 从而使分类超平面自动靠近样本数据. 该算法不仅显式地克服了以原点为参照的不足, 在计算速度上, 也因为线性规划而具有优势. 为了更好地覆盖目标数据所在的高密度区, 可通过嵌入局部密度来更为准确地反

映数据的分布,具体可参照文[25].

2.1.2 针对间隔距离度量的改进

OCSVM所采用的欧氏度量并未考虑数据的结构分布,导致目标数据所在的正半空间未必能准确地覆盖高密度区而很可能是次优解[26],于是出现了相应的改进算法:

(1)单类最小最大概率机(SCMPM).由 Lanckriet等提出的单类最小最大概率机(Single Class Minimax Probability Machine, SCMPM)[27]采用了概率约束来寻找一个尽量远离原点的超平面.根据 Chebychev不等式,可将目标数据的马氏距离与落入负半空间的概率密度上界建立联系,因此可通过最大化原点和超平面之间的马氏距离 $\frac{}{\sqrt{\boldsymbol{w}^{\mathrm{T}}\ \boldsymbol{w}}}$ 来优化目标数据所在的正半空间,并保证数据分布在最坏情形下仍至少有指定百分比的目标数据落入超平面内部.给定该百分比,对于任意数据分布,只要知道其相应均值和方差,SCMPM即能寻找到相应的超平面而无需知道全部样本的数值,这一思想可用于隐私保护任务.

受 SCMPM启发,Tang等用马氏距离取代了 OCSVM中欧氏距离设计了 MOCSVM,并将其引入到核学习中[28].

2.1.3 针对超平面的改进

与 OCSVM采用一个超平面来区分目标数据和异常数据不同,为描述物体的表面形状,Slab SVM[29]引入两个超平面,称之为厚板,来约束目标数据所在区域,这里的目标数据是对相应物体表面的测定.当厚板宽度为无穷时,Slab SVM退化为 OCSVM算法.

Tao等[30]将 Slab SVM进一步引申,将两个超平面之间的条带区域作为目标数据所在空间,并按照两类支持向量机的设计思想,从统计学理论出发定义了单类问题的损失函数、推广误差以及间隔,使单类分类器如同 SVM一样具有了完备的统计学习理论体系,并且提出了相应的单类算法 Slab OCC.

2.2 基于超球

与 OCSVM围绕原点、间隔以及超平面个数等多方面改进不同,超球的改进模型非常单一,即用超椭球取代超球.因为椭球模型考虑了数据在各方向的分布情形,相对于超球模型能更为紧凑地描述数据分布,这里所列出的改进算法,仅是椭球数目的不同.

2.2.1 优化单个椭球

对于给定的一组目标数据,最小包围椭球体积不会大于相应的最小包围超球体积.根据椭球体积计算公式,最小体积覆盖椭球(Minimum Volume Covering Ellipsoid, MVCE)[31]通过优化椭球体积来寻找最小超椭球以包围目标数据,其目标函数由两部分组成,其一是相当于 SVDD中对应超球半径的椭球半径,另一部分则是影响椭球体积的协方差矩阵行列式.需要指出,这里并没有如通常那样采用优化方法来求解对偶表达式,而是通过迭代来寻优,终止条件是根据马氏距离等于特征空间维数来达到最优.

上述过程借助核 PCA可得到 KMVCE,对于任意数据即圆心在任意点,可通过中心化推广到输入空间和高维特征空间[31].

核化后的高维空间若样本数小于维数,上述 KMVCE算法会导致椭球收缩到很小以致任何目标数据都会落在椭球外部.针对此不足,RMVCE算法[32]通过在对偶目标函数中加入正则化项 $r\boldsymbol{I}$ 来保证椭球在任何方向上的直径至少为 r,并通过 PAC-bayes方法[33]分析了算法在高维空间的推广误差上界,克服了 Rademacher复杂性仅能分析线性形式的不足[2].

另外,需要指出,采用类似目标函数的还有最小体积包围椭球算法 MVEE[15],不同处是后者将目标数据包围在一个单位椭球中,与 MVCE相比无需优化椭球半径,随之而来的是对偶形式的变化.而最大的不同,是两者的优化方式.MVEE仍沿用了优化对偶目标函数的方法,且因为没有推广到高维特征空间,导致算法的应用受到限制.

2.2.2 优化多个椭球

若目标数据呈多簇分布,上述所优化的单个超(椭)球不仅包围了目标数据,也包围了簇间的无关空白区域,而导致欠拟合.虽然这在一定程度上可以通过调整核参数,如减小高斯核带宽来改善,但因模型本身的不足,并不能从根本上解决这一问题.为此,Wang等[34]提出了结构化单类分类器(sTructure One-Class Classification, TOCC).TOCC通过寻找若干个最小超椭球来包围各簇目标数据,实验结果验证了其推广性

能较单个超(椭)球方法有显著提高. 若数据呈单簇分布, TOCC即退化为马氏椭球学习机 MEIM[35]. 与MVCE同时优化椭球半径和协方差矩阵不同, TOCC仅优化椭球半径, 从该意义上来看, TOCC更接近于SVDD的目标函数.

2.3 数据预处理

上述算法是直接针对模型进行改进. 由于算法的数据依赖性, 因此改善数据分布也是提高单类分类器算法性能的又一途径. 考虑到高斯核 OCSVM 和 SVDD 的等价性, 故下面仅以 SVDD 展开讨论.

2.3.1 核白化目标数据

用最小超球来包围目标数据仅适合于数据呈各向同性的情形, 对于非球形分布数据, SVDD会因为包含过多的空白区域而不够紧凑. 为此, 可采用单椭球来改进. 另外, 也可通过改变数据在核空间的分布来解决. 为使数据分布在一个超球内, 首先可通过 KPCA 寻找到核空间的各主分量, 而后将数据投影到该主分量上得到在新的坐标系下以原点为圆心的单位超球中, 上述过程称为白化. 经白化处理后的数据, 可采用SVDD算法来寻找最小超球, 所得到的边界更为紧凑[36].

仔细分析上述过程可发现, 数据白化 + SVDD 相当于在原空间中寻找最小椭球, 只是前者分为两个阶段. 如此做的最大好处是白化后的数据可采用任何单类分类器来处理.

2.3.2 KPCA 单类分类器

白化处理作为数据预处理方式需要尽量保持全部目标数据的信息, 然而若目标数据分布于子空间中时, 上述白化会因为有相当一部分特征值为零或接近于零而带来计算量的增加, 显然也会影响到后续分类器的效果. 为此, 可直接采用 KPCA 作为单类分类器, 通过计算高维空间的重建误差并使之最小, 实现分类和异常检测[37].

3 结语

OCSVM 和 SVDD 是两类以核方法为基础的单类分类器, 两者同属于统计学习理论应用范畴. 以 VC 维降低为目标, 前者通过寻找一个远离原点的超平面, 使目标数据所在的正半空间最小, 而后者通过寻找一个包含大部分目标数据的最小超球, 实现体积最小化. 两者在取高斯核时等价. 两算法坚实的理论背景和符合直觉的算法解释使之随后出现了大量的变异算法. 本文即以此为主线, 分别从模型改进和数据预处理的角度, 较为全面地回顾了相应的改进算法. 其中模型改进部分包括超平面模型各元素 (原点、间隔和平面数量)的改进以及超球模型的各类椭球改进算法. 数据处理部分, 主要采用核主成分分析来实现数据的白化预处理或通过改变数据在空间的分布直接实现异常检测.

[参考文献] (References)

[1] Sch lkopf B, Smola A. Learning With Kernels[M]. Cambridge, MA: MIT Press, 2002.

[2] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis[M]. Cambridge: Cambridge University Press, 2004.

[3] Vapnik V. Statistical Learning Theory[M]. New York: Addison-Wiley, 1998.

[4] Cristianini N, Taylor J S. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. Cambridge: Cambridge University Press, 2000.

[5] Mika S, R tsch G, Weston J, et al. Fisher discriminant analysis with kernels[C] // Neural Networks for Signal Processing IX. Piscataway, NJ: IEEE, 1999.

[6] Sch lkopf B, Smola A J, M ller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998(10): 1 299-1 319.

[7] Sch lkopf B, Williamson R C, Smola A J. Support vector method for novelty detection[C] // Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2000.

[8] Tarassenko L, Hayton P, Brady M. Novelty detection for the identification of masses in mammograms[C] // Proc 4th Int IEE Conf Artif Neural Netw. Cambridge: Oxford University Press, 1995.

[9] Lazarevic A, Ertoz L, Kumar V, et al. A comparative study of anomaly detection schemes in network intrusion detection[C] // SDM 2003. San Francisco: SIAM, 2003.

[10] Roth V. Outlier detection with one-class kernel fisher discriminants[C] // Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2005.

[11] Chen Y, Zhou X, Huang T. One-class SVM for learning in image retrieval[J]. Image Processing 2001(1): 34-37.
[12] Manevitz L, Yousef M. One-class SVMs for document classification[J]. Journal of Machine Learning Research, 2001(2): 139-154.
[13] Markou M, Singh S. Novelty detection: a review-part 1: statistical approaches[J]. Signal Processing, 2003, 83(12): 2 481-2 497.
[14] Moya M, Koch M, Hostetler L. One-class classifier networks for target recognition applications[C]//Proceedings World Congress on Neural Networks. Portland, OR: International Neural Network Society, 1993: 797-801.
[15] Juszczak P. Learning to recognise: a study on one-class classiifcation and active learning[D]. Delft: Delft University of Technology, 2006.
[16] Tax D. One-class classification: concept-learning in the absence of counter-examples[D]. Delft: Delft University of Technology, 2001.
[17] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.
[18] Schlkopf B, Platt J C, Shawe-Taylor J. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1 443-1 471.
[19] Schlkopf B, Smola A J, Williamson R C, et al. New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1 207-1 245.
[20] Platt J. Fast training of support vector machines using sequential minimal optimization[C]// Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1999.
[21] Tax D, Duin R P. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11-13): 191-1 199.
[22] Tax D, Duin R P. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.
[23] Scholkopf B, Platt J, Smola A. Kernel method for percentile feature extraction, MSR-TR-2000-22[R]. Microsoft Technical Report, 2000.
[24] Campbell C, Bennett K P. A linear programming approach to novelty detection[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001.
[25] 冯爱民, 陈斌. 基于局部密度的单类分类器 LP改进算法[J]. 南京航空航天大学学报, 2006, 38(6): 727-731.
Feng Aimin, Chen Bin. Improving LP algorithms of one-class classifier based on the local density factor[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2006, 38(6): 727-731. (in Chinese)
[26] Alberto M, JM. One-class support vector machines and density estimation: the precise relation[C]// Progress in Pattern Recognition, LNCS. Berlin: Springer, 2004.
[27] Lanckriet G R G, Ghaoui L E, Bhattacharyya C, et al. A robust minimax approach to classification[J]. Journal of Machine Learning Research, 2002(3): 555-582.
[28] Tsang I W, James T K, Li S. Learning the kernel in Mahalanobis one-class support vector machines[C]// Proceedings of the International Joint Conference on Neural Networks. Canada: Vancouver, 2006.
[29] Schlkop B. Kernel methods for implicit surface modeling[C]// Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: NIPS, 2004.
[30] Tao Q, Wu G W, Wang J. A new maximum margin algorithm for one-class problems and its boosting implementation[J]. Pattern Recognition, 2005, 38(77): 1 071-1 077.
[31] Dolia A, Harris C, Shawe-Taylor J K, et al. Kernel ellipsoidal trimming[J]. Computational Statistics and Data Analysis, 2007, 52(1): 309-324.
[32] Dolia A N, Bie T D, Harris C J, et al. The minimum volume covering ellipsoid estimation in kernel-defined feature spaces[C]// Proc of the 17th European Conference on Machine Learning. Berlin: Springer-Verlag, 2006.
[33] Langford J, Shawe-Taylor J. PAC Bayes and margins[C]// Advances in Neural Information Processing Systems. Vancouver and Whistler, British Columbia: MIT Press, 2003.
[34] Wang D, Yeung D S, Tsang E C C. Structured one-class classification[J]. IEEE Trans on Systems, Man, and Cybernetics-Part B: Cybernetics, 2006, 36(6): 1 283-1 294.
[35] Wei X K, Huang G B, Li Y H. Mahalanobis ellipsoidal learning machine for one class classification[C]// Proceedings of the 6th International Conference on Machine Learning and Cybernetics. Hong Kong: IEEE Press, 2007.
[36] Tax D, Juszczak P. Kernel whitening for one-class classification[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2003, 17(3): 333-347.
[37] Hoffmann H. Kernel PCA for novelty detection[J]. Pattern Recognition, 2007, 40: 863-874.

[责任编辑: 严海琳]