计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
45(11): 1858-1864, 2008

增强型单类支持向量机

冯爱民[1] 薛 晖[1] 刘学军[1] 陈松灿[1] 杨 明[2]
[1](南京航空航天大学信息科学与技术学院 南京 210016)
[2](南京师范大学计算机科学系 南京 210097)
(amfeng@nuaa.edu.cn)

Enhanced One-Class SVM

Feng Aimin[1], Xue Hui[1], Liu Xuejun[1], Chen Songcan[1], and Yang Ming[2]
[1](*College of Information Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 210016)
[2](*Department of Computer Science, Nanjing Normal University, Nanjing* 210097)

Abstract One-class-classifier (OCC) aims to distinguish a target class from outliers. Existing OCC algorithms based on hyperplane, such as one-class SVM (OCSVM) and Mahalanobis one-class SVM (MOCSVM), solve this problem by finding a hyperplane with the maximum distance to the origin. However, since they either neglect the structure of the given data or just takes the structure into account in a relatively coarse granularity, only the suboptimal hyperplane may be abtained. In order to mitigate this problem, a novel OCC named enhanced one-class SVM (EnOCSVM) is proposed. First obtaining the distribution of the target data by the unsupervised methods such as agglomerative hierarchical clustering, and then embedding the cluster information into the original OCSVM framework, EnOCSVM can optimize the tightness of target data and maximizes the margin from the origin simultaneously. In this way, EnOCSVM not only takes much more priori knowledge into account than the above algorithms, but also provides a general method to extend the present SVM algorithm to consider intrinsic structure of the data. Moreover, the optimization of the EnOCSVM can be solved using the standard SVM implementation similar to OCSVM, and all the advantages of SVM are preserved. Experiment results on benchmark data sets show that EnOCSVM really has better generalization than OCSVM and MOCSVM significantly.

Key words one-class-classifier; hyperplane; structure information; support vector machine; multi-clustered

摘 要 现有基于超平面的单类分类器,包括 one-class SVM(OCSVM)和马氏 one-class SVM(MOCSVM),由于未考虑数据的结构信息或粒度较粗,寻找的超平面很可能是次优解.为此,增强型单类支持向量机(enhanced OCSVM, EnOCSVM)通过在现有 SVM 算法中加入数据先验信息以克服其不足.首先,EnOCSVM 通过聚类得到数据的内在分布簇,而后将各簇结构信息嵌入到 OCSVM 框架中,最大化间隔的同时,优化输出空间中各簇数据的紧性.由于保留了 SVM 框架不变,EnOCSVM 仍具备原算法的全部优点,并因结合了数据的簇结构信息而具有更好的推广性.标准数据集上的实验表明,EnOCSVM 的推广性能较 OCSVM 和 MOCSVM 均有显著提高.

关键词 单类分类器;超平面;结构信息;支持向量机;簇分布

中图法分类号 TP391.4

收稿日期:2008-04-30;**修回日期**:2008-06-26

基金项目:国家自然科学基金项目(60603029,60703016);江苏省博士后科研资助计划基金项目(05225)

单类问题[1]是指训练样本中只有一类目标数据,其他非目标数据因为采集代价过高(如故障诊断,不可能为了获取大量的非目标数据而人为破坏机器)或者数据类型过多而使现有数据无代表性(如人脸检测,任何非人脸都可以作为非目标数据),不得不放弃获取,造成样本缺失.目标数据又称为正常类,而非目标数据亦称为异常类或负类.针对单类问题所设计的单类分类器,可用于文本检测[2]、图像抽取[3]、异常检测[4-6]等,典型的研究方法有基于密度和基于边界的方法[7-8]等.

基于密度的方法通过参数化或非参数化方法来估计样本数据的概率密度[9],而后设置阈值以判定测试数据是否属于正常数据.现实中有限的目标数据反映的常常是数据所处的区域而非密度信息,因此采用密度估计很可能把正常数据的稀疏区域作为低密度区而误判,并且由于维数灾难[10],该方法更适合低维数据.

基于边界的方法是通过几何形状如超平面[7,11-12]、超(椭)球[8,13-15]等,将目标数据中的高密度区映射到一个正半空间或者封闭的超(椭)球里,从而在尽可能包含大部分目标数据的前提下,最小化上述几何形状的体积,以达到错误率最小的目的.由于该方法的目标是寻找数据的支持域而非密度,从而适合处理高维、有噪和有限样本的单类问题,成为单类分类器研究的热点.

单类支持向量机(one-class SVM,OCSVM)[7]作为超平面模型的代表,由于采用的欧氏度量无法考虑数据结构分布,导致所获得的超平面很可能是次优解[12].单类最小化最大概率机 one-class MPM[11]虽然利用了样本分布的先验信息,但因为所采用的概率模型必须通过二次锥规划求解,且因无法采用对偶理论而丧失了解的稀疏性.Tsang 等人[12]直接将马氏距离引入到 OCSVM(MOCSVM),避免二次锥规划求解的同时,一定程度上克服了原算法采用欧氏距离的不足.

支持向量数据描述(support vector data description,SVDD)[8,16]作为超球模型典型算法,当目标数据在各方向呈现不同的分布趋势或数据并非球状分布时,SVDD 所找到的超球会因包含过多的非目标区域而不够紧凑[17],Wei 等人[13]利用超椭球代替了 SVDD 中的超球以考虑数据的结构信息,类似的椭球模型还有最小体积包含椭球(minimum volume enclosing ellipsoid,MVEE)[14]以及核最小体积覆盖椭球(kernel minimum volume covering ellipsoid,KMVCE)[15],它们均是通过优化椭球体积来寻找最小超椭球.

上述基于边界的算法由于没有考虑数据分布(如 OCSVM 或 SVDD)或结构信息粒度过粗(如 MOCSVM 或 MVEE,KMVCE),因而更适合解决数据呈单簇分布的情形.然而现实中越来越多的单类应用如网络入侵检测、手写体识别、人脸检测等,目标数据均呈多簇分布,如图 1 即为 UCI 标准数据集 Sonar 部分数据经 KPCA[18]降维取最大三维主分量后的可视化分布[19],图 1(a)(b)分别为正负类相应的簇分布,分别对应着实验部分 Sonar1 和 Sonar2.尽管低维空间中的分布并不能完全表示高维空间的情形,但从实验部分的结果可充分说明图 1 可视化的代表性.因此,设计可用于多(单)簇数据

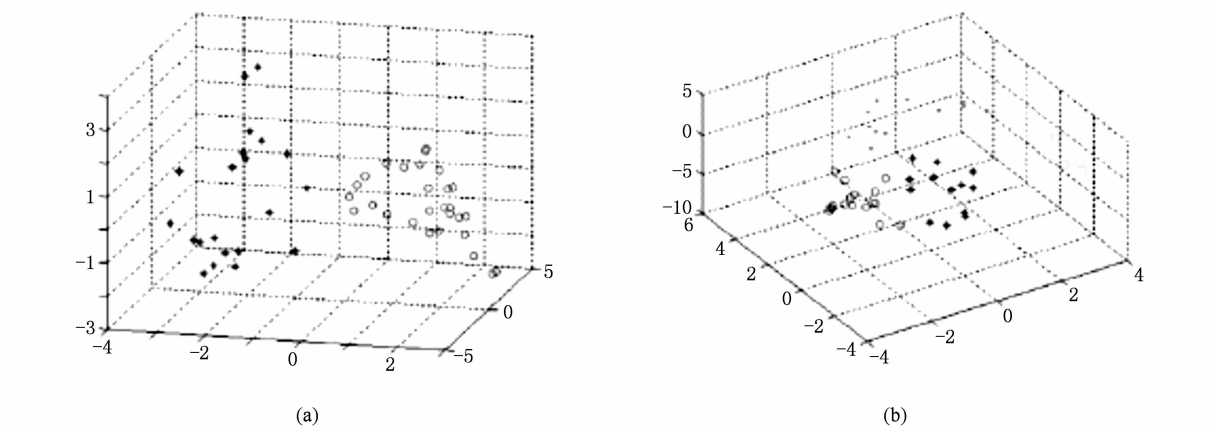


Fig. 1 Visualization of multi-clustered data structure of Sonar data set by projecting the data images in the RBF kernel space onto the three most principal kernel components. (a) Data structure in positive class and (b) Data structure of negative class.

图 1 Sonar 数据集采用径向基核三维可视化簇分布[19]. (a) 正类结构;(b) 负类结构

分布的单类分类器已成为应用驱动的必然.

结构化单类分类器(structure one-class classification, TOCC)[20]作为首次提出的针对数据呈多簇分布的单类分类器,首先通过聚类找到数据的簇分布,而后借助超椭球模型即若干个马氏距离SVDD来寻找各簇数据的超椭球,椭球数目等于聚类个数,其推广性能较其他算法有显著提高[20]. 然而,TOCC为得到好的推广性能设计了多个椭球并依次采用二次锥规划优化求解,且受样本数目和聚类算法的影响,所寻找的椭球有相当的不确定性,该不确定性会在一定程度上影响算法的性能.

本文所提出的增强型单类支持向量机(enhanced one-class SVM, EnOCSVM)旨将超平面模型应用于多簇数据,为此算法也分为两个阶段:首先如TOCC那样通过聚类得到数据的内在分布簇以获得各簇结构信息;而后在分类阶段,EnOCSVM并未像TOCC那样完全根据聚类的结果来设计各椭球,而是在现有OCSVM框架下构建出新的目标函数,最大化间隔的同时,通过嵌入各簇结构信息组成的总体协方差矩阵以反映数据的先验信息,不仅保留了原算法的诸多优点如全局最优性、对偶稀疏性、可二次规划求解以及易核化等,并且由于结合了数据的结构信息,带有更多的先验知识,因此,与基于超平面的同类算法相比,理论上具有更好的推广性.

1 单类支持向量机

给定数据集 $X=\{x_1,x_2,\cdots,x_n\}$, n 为样本个数,单类支持向量机OCSVM通过最大化原点和目标数据之间的最小欧氏距离 $\frac{\rho}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}$ 来寻找最优超平面,其中 $\boldsymbol{w}$ 是超平面的法向量, ρ 是超平面截距. 目标函数中引入松弛因子 $\boldsymbol{\xi}=[\xi_1,\cdots,\xi_n]^{\mathrm{T}}$ 使算法具有一定的鲁棒性,具体刻画如下:

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

其中, $\upsilon\in[0,1]$ 是所谓的百分比估计,和支持向量的数目有密切联系,即 υ 是边界支持向量的上界,是全部支持向量的下界,称为 υ 属性[7].

引入向量 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_v]^{\mathrm{T}}$, $\boldsymbol{1}_n=[1,\cdots,1]^{\mathrm{T}}$,式(1)的对偶形式以矩阵表示为

$$\begin{aligned}&\min_{\alpha}=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}X\boldsymbol{\alpha}\\&\text{s.t.}\quad\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}_n=1,\boldsymbol{0}\leqslant\boldsymbol{\alpha}\leqslant\frac{1}{\upsilon n}\boldsymbol{1}_n.\end{aligned}\tag{2}$$

这是一个二次规划(quadratic programming, QP)问题,可采用经典的QP软件包或者序列最小优化(sequential minimal optimization, SMO)算法[21]来优化.

对于新来的测试样本 z,判别函数如下:

$$f(x)=\operatorname{sgn}[\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}z-\rho]=\begin{cases}1,\ \text{target.}\\-1,\ \text{outlier.}\end{cases}\tag{3}$$

2 增强型单类支持向量机

上述OCSVM在计算间隔时并未考虑数据的结构信息,因而所寻找的超平面很可能是次优解. 针对此不足且考虑到算法要同时应用于多(单)簇数据分布,增强型单类支持向量机EnOCSVM亦分为两个阶段:首先借助聚类算法获取数据的内在分布簇;而后将各簇结构信息嵌入到OCSVM框架下构建出包含数据先验信息的目标函数,详细过程如下.

2.1 获取簇信息

为使分布簇能用协方差矩阵表示数据分布信息,要求所采用的聚类算法所得到的子簇需是球形紧凑聚类,凝聚型层次聚类Ward算法因为满足此要求亦在本文采用. 通过重复计算当前各聚类中心点之间的距离且合并最近一对聚类这一过程,所有数据点最终可成为一类. 在整个聚类过程中,随着聚类数目的减少,每次所合并的两个聚类中心之间的距离(称为合并距离)逐渐增大,将上述聚类数目和合并距离之间的变化汇成曲线,其中曲率变化最大的点(拐点)所对应的聚类数目即为聚类个数. 为找到该拐点,本文采用文献[22]所使用的贪婪法来遍历任一点作为拐点时所拟合曲线的误差,其中,误差最小即为拐点.

由于凝聚型层次聚类的树型结构,因此可在任意深度确定各聚类所包含的样本点,于是,根据拐点所确定的聚类个数将各样本点划分到各自独立的子类,自此,聚类阶段完成.

2.2 线性EnOCSVM

上述聚类过程将目标数据分成若干个簇,记为 $S_1,\cdots,S_i,\cdots,S_M$,并分别以协方差矩阵 $\boldsymbol{\Sigma}_{S_1},\cdots,\boldsymbol{\Sigma}_{S_i},\cdots,\boldsymbol{\Sigma}_{S_M}$ 表示各簇结构信息. 为将此聚类结果加入到基于超平面的分类器设计中,现将上述各协方差整体求和后组成新的总体协方差矩阵 $\boldsymbol{\Sigma}$,且 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{S_1}+\cdots+\boldsymbol{\Sigma}_{S_i}+\cdots+\boldsymbol{\Sigma}_{S_M}$. 考虑到OCSVM所优化的目标函数可看作类间(目标数据和原点之间)距离,而协方差矩阵所揭示的结构信息恰好指示目标数据

的类内紧性,因此受 LDA 类间尽量分开且类内尽量紧凑思想的启发[10],EnOCSVM 在现有 OCSVM 框架下嵌入分布簇结构信息 $\boldsymbol{\Sigma}$,相应目标函数如下:

$$\begin{aligned}&\min_{\boldsymbol{w},\boldsymbol{\xi},\rho}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{\lambda}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w}-\rho+\frac{1}{\nu n}\sum_{i=1}^{n}\xi_i\\&\text{s.t.}\quad \boldsymbol{w}^{\mathrm{T}}x_i\geqslant\rho-\xi_i,\ \xi_i\geqslant 0,i=1,\cdots,n,\end{aligned}\tag{4}$$

其中参数 λ 用于类内紧性和类间的平衡,且 $\lambda\geqslant 0$. 从式(4)可看出,λ 较大则更强调目标数据的结构分布信息,而 λ 较小则更强调目标数据与原点之间的可分性,$\lambda=0$ 退化为 OCSVM 算法,即不考虑数据的类内结构信息而只关注数据间的可分性. 由此可见,EnOCSVM 是原算法的推广,它通过在原算法的基础上加入结构信息充分考虑数据的先验知识,而加入先验正是提高分类器精度的有效途径.

EnOCSVM 实现了最大化类间间隔$\frac{\rho}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}$的同时最小化类内散度 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w}$,当进一步整理式(4)并调整参数 λ 的位置,可得

$$\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_r\boldsymbol{w}-\rho+\frac{1}{\nu n}\sum_{i=1}^{n}\xi_i,\tag{5}$$

其中,$\boldsymbol{\Sigma}_r=\frac{1}{\lambda}\boldsymbol{I}+\boldsymbol{\Sigma}$,表示总体协方差矩阵 $\boldsymbol{\Sigma}$ 估计值的不确定性,而$\frac{\rho}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_r\boldsymbol{w}}$表示所寻找的超平面与原点之间的马氏距离,即整理后的 EnOCSVM 目标函数可看作加入簇结构信息的马氏距离 MOCSVM,或者说 MOCSVM 是 EnOCSVM 将全部数据样本聚为一类的特殊情形,一致起见,下文称 MOCSVM 为全局 EnOCSVM,简称 gEnOCSVM.

将式(4)变换得到相应的对偶形式:

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{I}+\lambda\boldsymbol{\Sigma})^{-1}\boldsymbol{X}\boldsymbol{\alpha}\\&\text{s.t.}\quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}_n=1,\boldsymbol{0}\leqslant\boldsymbol{\alpha}\leqslant\frac{1}{\nu n}\boldsymbol{1}_n.\end{aligned}\tag{6}$$

和 OCSVM 的对偶形式(2)比较不难看出,式(6)所在的对偶空间已不再是样本内积 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 作用的输入空间,而是 $\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{I}+\lambda\boldsymbol{\Sigma})^{-1}\boldsymbol{X}$ 所在的输出空间,它相当于把样本通过$(\boldsymbol{I}+\lambda\boldsymbol{\Sigma})^{-\frac{1}{2}}$映射到输出空间中寻找最优超平面,该超平面相对于原空间来说是非线性的,因此不难理解,理论上 EnOCSVM 优于 OCSVM. 尽管该结论对于 gEnOCSVM 也同样成立,但由于其并未如 EnOCSVM 那样考虑数据分布簇而丢失了一部分先验信息,因而从理论上分析其性能不会超过后者,这点可从后续实验部分进一步说明. 值得指出,由于协方差矩阵可事先计算及存储,因此上述对偶形式在计算复杂度上并未带来额外计算量,并且由于同样是 QP 问题,因此优化可直接采用 OCSVM 的方法.

对于待测样本 z,可通过如下判别函数决定是否属于异常类:

$$f(x)=\mathrm{sgn}[\boldsymbol{\alpha}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{I}+\lambda\boldsymbol{\Sigma})^{-1}z-\rho].\tag{7}$$

2.3 核 EnOCSVM

这里仅指分类阶段的核化,简单起见,省略了核化空间的目标函数,而直接采用线性空间中的对偶式(6)引入核技巧,为此,需求所有簇的总体协方差矩阵 $\boldsymbol{\Sigma}$,首先某簇结构信息表示为

$$\boldsymbol{\Sigma}_{S_i}=\frac{1}{|S_i|}\widetilde{\boldsymbol{X}}_{S_i}\widetilde{\boldsymbol{X}}_{S_i}^{\mathrm{T}}-\frac{1}{|S_i|^2}\widetilde{\boldsymbol{X}}_{S_i}\boldsymbol{1}_{S_i}\boldsymbol{1}_{S_i}^{\mathrm{T}}\widetilde{\boldsymbol{X}}_{S_i}^{\mathrm{T}}=\widetilde{\boldsymbol{X}}_{S_i}\boldsymbol{H}_{S_i}\widetilde{\boldsymbol{X}}_{S_i}^{\mathrm{T}},\tag{8}$$

其中,$\boldsymbol{H}_{s_i}=\left(\frac{1}{|S_i|}\boldsymbol{I}_{S_i}-\boldsymbol{E}_{S_i}\boldsymbol{E}_{S_i}^{\mathrm{T}}\right)$,$|S_i|$ 表示某簇的样本个数,$i=1,\cdots,M$,$\widetilde{\boldsymbol{X}}_{S_i}=[x_1,\cdots,x_{|S_i|}]$表示该簇的样本,$\boldsymbol{I}_{S_i}$ 表示 $|S_i|$ 维数的 1 向量,$\boldsymbol{E}_{S_i}=\frac{1}{|S_i|}\boldsymbol{I}_{S_i}$. 因此,总体协方差矩阵 $\boldsymbol{\Sigma}$ 可整理为:

$$\boldsymbol{\Sigma}=\widetilde{\boldsymbol{X}}\boldsymbol{H}\widetilde{\boldsymbol{X}}^{\mathrm{T}},\tag{9}$$

其中 $\widetilde{\boldsymbol{X}}=[\widetilde{\boldsymbol{X}}_{S_1},\cdots,\widetilde{\boldsymbol{X}}_{S_M}]$,表示按簇分布重新排序的样本数据,$\boldsymbol{H}$ 表示各簇的结构分布信息. 为求式(6)中的$(\boldsymbol{I}+\lambda\boldsymbol{\Sigma})^{-1}$,此处需借助 Woodbury 公式[23]:

$$(\boldsymbol{A}+\boldsymbol{UBV})^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{UB}(\boldsymbol{B}+\boldsymbol{BVA}^{-1}\boldsymbol{UB})^{-1}\boldsymbol{BVA}^{-1},$$

得

$$(\boldsymbol{I}+\lambda\boldsymbol{\Sigma})^{-1}=\boldsymbol{I}-\lambda\widetilde{\boldsymbol{X}}\boldsymbol{H}(\boldsymbol{H}+\lambda\boldsymbol{H}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}\boldsymbol{H})^{-1}\boldsymbol{H}\widetilde{\boldsymbol{X}}^{\mathrm{T}}.\tag{10}$$

将上述结果代入矩阵表示的对偶式(6),并将样本数据之间的内积用核函数来替代,化简如下:

$$\min\ \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{K}-\lambda\widetilde{\boldsymbol{K}}\boldsymbol{H}(\boldsymbol{H}+\lambda\boldsymbol{H}\hat{\boldsymbol{K}}\boldsymbol{H})^{-1}\boldsymbol{H}\widetilde{\boldsymbol{K}}^{\mathrm{T}})\boldsymbol{\alpha},\tag{11}$$

其中 $\boldsymbol{K}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 表示输入样本数据的核矩阵,$\widetilde{\boldsymbol{K}}=\boldsymbol{X}^{\mathrm{T}}\widetilde{\boldsymbol{X}}$ 表示样本数据和聚类后重新排序数据的核矩阵,$\widetilde{\boldsymbol{K}}^{\mathrm{T}}$ 是其相应转置,$\hat{\boldsymbol{K}}=\widetilde{\boldsymbol{X}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}$ 是聚类后重新排序数据之间的核矩阵,原有约束条件不变. 同样,该核化目标函数仍属二次规划,优化方法可参照 QP 求解.

对于新来的样本矩阵 $\boldsymbol{Z}$,将式(10)带入式(7),可得到高维特征空间的判别函数矩阵:

$$\boldsymbol{f}=\boldsymbol{sgn}[\boldsymbol{\alpha}^{\mathrm{T}}(\widehat{\boldsymbol{K}}-\lambda\widetilde{\boldsymbol{K}}\boldsymbol{H}(\boldsymbol{H}+\lambda\boldsymbol{H}\hat{\boldsymbol{K}}\boldsymbol{H})^{-1}\boldsymbol{H}\widehat{\boldsymbol{K}}-\boldsymbol{\rho}],\tag{12}$$

这里 $\hat{\boldsymbol{K}}$ 表示训练样本和测试样本的核矩阵，$\bar{\boldsymbol{K}}$ 表示分簇排序后的训练样本和测试数据之间的核矩阵，而超平面的偏移向量 $\boldsymbol{\rho}=\rho\boldsymbol{1}^{\mathrm{T}}$，向量 $\boldsymbol{1}$ 的意义同前，$\boldsymbol{f}$ 和 $\boldsymbol{sgn}$ 分别对应式(7)中 $f(x)$ 和 sgn(·)的向量形式.

3 实 验

由于单类少有专用数据集，因此实验中采用了 UCI 的 5 个两类数据集分别产生 10 个单类数据，分别称为 Data Set 1, Data Set 2，其中 Data Set 1 是指将数据集中较多的一类作为目标类，而数据较少的一类作为异常类，Data Set 2 则相反. 这里比较了 EnOCSVM 与 OCSVM 以及 gEnOCSVM 算法，后者是省略了聚类过程的 EnOCSVM 算法，即将数据看成一个簇或者说省略聚类过程而直接考虑数据的整体结构信息，粒度与 MOCSVM 相当. 实验中采用了高斯核函数

$$K(x,y)=\mathrm{e}^{-\|x-y\|^2/\sigma^2}.$$

核带宽 σ 和 λ 采用网格式搜索，5-fold 交叉验证，随机选择 80% 的目标数据作为训练样本，剩余 20% 的目标数据和异常类分别作为正类和异常类的测试数据，重复 10 轮.

单类问题根据测试样本的真实类别和分类结果构成了表 1 的混淆矩阵：

Tabel 1 Confusion Matrix of One-Class Classification

表 1 单分类的混淆矩阵

Classified Label	True Label	
	Target	Outlier
Target	True Positive(TP)	False Negative(FN)
Outlier	False Positive(FP)	True Negative(TN)

从表 1 可看出单类有两类错误，即目标类被错分为异常类 FN 和异常类被错分为目标类 FP，实验中即采用了 FP/FN 评价指标，显然，这两个结果越小越好. 表 2 列出了上述 5 个数据集的 10 组结果，其中黑体表示每组 3 个算法中 $\frac{(FP+FN)}{2}$ 的最小结果，称为 BalanceLoss(BL)[24]. 数据集括号内 $Tr:TeT:TeO$ 依次表示该数据集的训练样本、目标类和异常类测试样本. 3 个算法的 v 均取 0.1，即在训练样本上，最多允许 10% 的数据落在边界或分类面之外而成为支持向量，根据统计学习理论[25]，该值不可能为零，也就是说，单类必然存在训练误差. 表中 FN 括号中列出了相应的训练误差. 优化方法采用单类 SMO 算法[7].

Table 2 *FP/FN/BL* Results on UCI Benchmark Data Sets with OCSVM and (g)EnOCSVM

表 2 OCSVM 和(g)EnOCSVM 算法在 UCI 数据集上的 *FP/FN/BL*

Data Set	OCSVM			gEnOCSVM			EnOCSVM		
($Tr:TeT:TeO$)	FP	$FN(Tr)$	BL	FP	$FN(Tr)$	BL	FP	$FN(Tr)$	BL
Breast1(367:91:241)	0.0261	0.0593(0.0649)	0.0427	**0.0274**	**0.0473**(0.0629)	**0.0373**	0.0166	0.0582(0.0676)	0.0374
Breast2 (193:48:458)	0.0328	0.1021(0.1093)	0.0674	0.0367	0.0708(0.0710)	0.0538	**0.0413**	**0.0375**(0.0907)	**0.0394**
Heart1(123:41:139)	0.6165	0.2125(0.1576)	0.4145	0.5165	0.2531(0.1530)	0.3848	**0.5360**	**0.1594**(0.1788)	**0.3477**
Heart2(112:27:164)	0.5579	0.3519(0.2205)	0.4549	0.6805	0.1815(0.1054)	0.4310	**0.6280**	**0.0778**(0.1170)	**0.3529**
Import1(71:17:71)	0.1775	0.3353(0.1901)	0.2564	**0.1817**	**0.2294**(0.1493)	**0.2056**	0.2085	0.2529(0.2169)	0.2307
Import2(57:14:88)	0.4273	0.1714(0.1368)	0.2994	0.3864	0.1143(0.1070)	0.2503	**0.2864**	**0.2071**(0.2702)	**0.2468**
Sonar1(89:22:97)	0.1155	0.6591(0.6169)	0.3873	0.1351	0.5409(0.4461)	0.3380	**0.1784**	**0.3773**(0.4461)	**0.2778**
Sonar2(78:19:111)	0.3847	0.3789(0.2474)	0.3818	0.3784	0.3474(0.1936)	0.3629	**0.2856**	**0.3105**(0.4487)	**0.2981**
Wine1(86:21:71)	0.4127	0.1238(0.1628)	0.2682	0.4197	0.0619(0.1116)	0.2408	**0.3282**	**0.1238**(0.1756)	**0.2260**
Wine2(57:14:107)	0.2523	0.3000(0.1842)	0.2762	0.2766	0.2143(0.1491)	0.2455	**0.2972**	**0.1143**(0.1649)	**0.2057**

分析表 2 可看出：

1) 考虑数据类内结构信息的 gEnOCSVM 和 EnOCSVM 所对应的 BL 均一致优于仅考虑数据类间信息的 OCSVM，尤其对于 Heart 和 Sonar 数据集，EnOCSVM 的 BL 较 OCSVM 降低了 7%～11%，gEnOCSVM 在 Import1 上也降低了 5%，验证了考虑数据结构信息的有效性.

2) 进一步比较结构信息的粒度，EnOCSVM 除在 Breast1 上和 gEnOCSVM 结果相当，在 Import1 上 BL 差于 gEnOCSVM 外，其他 8 组结果均优于 gEnOCSVM，最高达 6%～8%(见 Heart2 和 Sonar 数据集)，说明对于大部分数据，内部的确呈簇分布，EnOCSVM 因为合理考虑了该簇信息因此性能更优；相反，若数据内部没有明显的分簇，EnOCSVM 性能只与 gEnOCSVM 相当甚至出现降低.

3) 分析上述结果不难发现，考虑结构化信息的算法相对于 OCSVM 性能提高的主要原因是 FN 降低：从表 2 看，除 gEnOCSVM 在 Heart1 上和 EnOCSVM

在 Import 2 上 FN 增大之外,其他 9 组数据集上两算法的 FN 均一致减小,这是因为考虑了数据的结构信息可以更准确地把握数据的分布趋势,因而可使更多的正类落在目标区域内. 当然,两者降幅不同,如 Heart2 和 Sonar1,EnOCSVM 的 FN 降幅高达 28%,而 gEnOCSVM 只降低了 11%~17%,总的来说,多数情况下 EnOCSVM 降幅要高于 gEnOCSVM.

4) 对于 FP,由于在训练过程中没有该类数据,因此相对于 OCSVM,两算法的变化不同,但主要表现为 FP 增大,其中 gEnOCSVM 有 7 组 FP 增大,EnOCSVM 也有 5 组增大. 这是因为训练过程中由于考虑了数据的分布趋势而使数据域增大,因此可能将一部分异常数据所在的空间包含进来,并且 EnOCSVM 因为粒度更细而导致增幅更大,这也从另一侧面揭示了 EnOCSVM 性能提升的主要原因是 FN 降低.

5) 与其他两算法相比,EnOCSVM 在目标数据上的测试误差除 Import1 外均小于训练误差,这说明从更细粒度上考虑数据结构信息的 EnOCSVM 相对于其他两算法具有更好的推广性,它能在训练误差并不占优的情况下得到更小的 FN,而推广性能才是机器学习的本质.

4 结 论

尽可能利用先验知识,是提高分类器推广性能的关键所在,本文提出的增强型单类支持向量机 EnOCSVM,便是在现有 OCSVM 框架下嵌入先验信息的单类算法,为了能够更有效地处理单类多簇数据,整个算法分为两步:首先通过聚类得到数据内在分布簇,并构建反映各簇结构信息的总体协方差矩阵;而后将此先验信息作为类内紧性嵌入到经典的 OCSVM 框架中,最大化间隔的同时,最小化输出空间中数据各簇之间的类内散度. 作为 OCSVM 的推广,EnOCSVM 不仅继承了原算法全部的优点,克服了 OCSVM 并未考虑数据结构信息的不足,并且从更精细的粒度上刻画了数据的分布,因此在实际数据集上表现出更好的推广性.

由于考虑数据类内结构信息所引入的参数 λ,目前只能靠经验学习来获得,未来工作中,将关注该参数的选择,以便对于不同的数据分布具有一定的指导意义.

参 考 文 献

[1] Tax D. One-class classification: Concept-learning in the absence of counter-examples [D]. Delft: Delft University of Technology, 2001

[2] Manevitz L, Yousef M. One-class SVMs for document classification [J]. Journal of Machine Learning Research, 2002, 2: 139-154

[3] Chen Y, Zhou X, Huang T. One-class SVM for learning in image retrieval [J]. Image Processing, 2001, 1: 34-37

[4] Campbell C, Bennett K P. A linear programming approach to novelty detection [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press. 2001: 395-401

[5] Hoffmann H. Kernel PCA for novelty detection [J]. Pattern Recognition, 2007, 40(3): 863-874

[6] Pan Zhisong, Ni Guiqiang, Tan Lin, *et al*. One-class classification and immune framework in abnormal detection [J]. Journal of Nanjing University of Science and Technology: Natural Science, 2006, 30(1): 48-52 (in Chinese)
(潘志松, 倪桂强, 谭琳, 等. 异常检测中单类分类算法和免疫框架设计[J]. 南京理工大学学报:自然科学版, 2006, 30(1): 48-52)

[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] Bishop C. Novelty detection and neural network validation [C] //IEE Proc of Vision, Image and Signal Processing. 1994: 217-222

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

[11] Lanckriet G R G, Ghaoui L E, Jordan M. Robust novelty detection with single-class MPM [C] //Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002: 905-912

[12] Tsang I W, James T K, Li S. Learning the kernel in Mahalanobis one-class support vector machines [C] //Proc of the Int Joint Conf on Neural Networks (IJCNN'06). Los Alamitos: IEEE Computer Seeiety, 2006: 1169-1175

[13] Wei X K, Huang G B, Li Y H. Mahalanobis ellipsoidal learning machine for one class classificationin [C] //Proc of the 6th Int Conf on Machine Learning and Cybernetics. Los Alamitos: IEEE Computer Seeiety, 2007: 3528-3533

[14] Juszczak P. Learning to recognise: A study on one-class classiifcation and active learning [D]. Delft: Delft University of Technology, 2006

[15] Dolia A, Harris C, Shawe-Taylor J. Kernel ellipsoidal trimming [J]. Computational Statistics and Data Analysis, 2007, 52(1): 309-324

[16] Tax D, Duin R P. Support vector data description [J]. Machine Learning, 2004, 54(1): 45-66

[17] Tax D, Juszczak P. Kernel whitening for one-class classification [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2003, 17(3): 333-347

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

[19] Yeung D S, Wang D, Wing W Y N. Structured large margin machines: Sensitive to data distributions [J]. Machine Learning, 2007, 68(2): 171-200

[20] 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): 1283-1294

[21] Platt J. Fast training of support vector machines using sequential minimal optimization [C] //Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1999: 185-208

[22] Salvador S, Chan P. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms [C] //Proc of the 16th IEEE Int Conf on Tools with AI. Washington: IEEE Computer Society, 2004: 576-584

[23] Zhang Xianda. Matrix Analysis and Applications [M]. Beijing: Tsinghua Qinghua University Press, 2004 (in Chinese)
(张贤达. 矩阵分析与应用[M]. 北京: 清华大学出版社, 2004)

[24] James T K, Tsang I W, Zurada J M. A class of single-class minimax probability machines for novelty detection [J]. IEEE Trans on Neural Networks, 2007, 18(3): 778-785

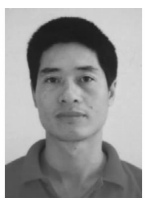
[25] Vapnik V. Statistical Learning Theory [M]. New York: Addison-Wiley, 1998

Feng Aimin, born in 1971. Ph. D. candidate and lecturer. Her main research interests include artificial intelligence, pattern recognition, and machine learning.
冯爱民,1971 年生,博士研究生,讲师,主要研究方向为人工智能、模式识别和机器学习.

Xue Hui, born in 1979. Ph. D. candidate. Her current research interests include pattern recognition and machine learning.
薛 晖,1979 年生,博士研究生,主要研究方向为模式识别和机器学习.

Liu Xuejun, born in 1976. Ph. D. and associate professor. Her main research interests focus on bioinformatics and machine learning.
刘学军,1976 年生,博士,副教授,主要研究方向为生物信息学、机器学习.

Chen Songcan, born in 1962. Professor and Ph. D. supervisor. Senior member of China Computer Federation. His main research interests include pattern recognition, bioinformatics engineering, image processing, and neural networks.
陈松灿,1962 年生,教授,博士生导师,中国计算机学会高级会员,主要研究方向为模式识别、生物信息工程、图像处理和神经网络.



Yang Ming, born in 1964. Ph. D. and professor. His main research interests include data mining and knowledge discovery, machine learning, rough sets theory and its applications.
杨 明,1964 年生,博士,教授,主要研究方向为数据挖掘、知识发现、机器学习、粗糙集及其应用.

Research Background:

Different from binary (or multi) class classification problems where one tries to distinguish between two (or more) classes of objects, one-class classification tries to describe one class of objects, and distinguish it from all other possible objects. These classification tasks can be found in many real-world scenarios like machine faulty diagnosis, network intrusion detection and document classification, *etc*. To solve this problem, an extreme approach is to estimate the probability density function of the given data. New objects that fall under some density threshold are considered outliers. Inspired by Vapnik's principle that never solves a problem that is more general than the one we actually need to solve, the domain-based approach has become the main method recently. As the state-of-the-art SVM introduced in this field, one-class SVM (OCSVM) uses a hyperplane to separate the given data from the origin with maximal margin. Its modified version, MOCSVM uses Mahalanobis metric instead of the original Euclidean distance for improving. But since the above algorithms both neglect the priori knowledge of the given data which is possibly composed of homogeneous group, the suboptimal hyperplane may be obtained. Partially supported by the National Natural Science Foundation of China(grant No. 60603029 and 60703016), the proposed enhanced one-class SVM tries to alleviate this problem by two steps. Firstly, it obtains the multi-clustered distribution of the given data by some kind of clustering methods, and then embeds this information into the original OCSVM framework. Since this novel method pays much more attention to the priori information than the above algorithms, its generalization power should be better than both of them. In addition, this new algorithm can also be used as a theoretical framework to extend the present SVM including the binary and multi class algorithms.