基于 L1 范数稀疏距离测度学习的单类分类算法

胡正平,路 亮,许成谦
(燕山大学信息科学与工程学院,河北秦皇岛 066004)

摘 要: 已有单类分类算法通常采用欧氏测度描述样本间相似关系,然而欧氏测度有时难以较好地反映一些数据集样本的内在分布结构,为此提出一种用于改善单类分类器描述性能的高维空间单类数据距离测度学习算法,与已有距离测度学习算法相比,该算法只需提供目标类数据,通过引入样本先验分布正则化项和 L1 范数惩罚的距离测度稀疏性约束,能有效解决高维空间小样本情况下的单类数据距离测度学习问题,并通过采用分块协调下降算法高效的解决距离测度学习的优化问题.学习得到的距离测度能容易地嵌入到单类分类器中,仿真实验结果表明采用学习得到的距离测度能有效改善单类分类器的描述性能,特别能够改善覆盖分类的描述能力,从而使得单类分类器具有更强的推广能力.

关键词: 模式识别; 稀疏距离测度学习; L1 范数; 单类分类器
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2012) 01-0134-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.01.022

L1 Norm Sparse Distance Metric Learning for One-class Classifier

HU Zheng-ping, LU Liang, XU Cheng-qian
(*School of information Science and engineering & Yanshan University. Qinhuangdao, Hebei* 066004, *China*)

Abstract: Most one-class classification algorithms measure similarity based on Euclidean distance between samples. Unfortunately, the Euclidean distance couldn't reveal the internal distribution of some datasets, and so reduced the descriptive ability of these methods. A distance metric learning algorithm was proposed to improve the performance of one-class classifiers in this paper. Compared with existing distance metric learning algorithm, the algorithm only needed to provide target class data, it could effectively solve distance metric learning problem for one-class samples in high-dimensional space by imposing sample distribution prior and sparsity prior with l1-norm constraint on the distance metric, and the formulation could be efficiently optimized in a block coordination descent algorithm. The learned metric can be easily embedded into one-class classifiers, the simulation experimental results show that the learned metric can effectively improve the description performance of one-class classifiers, in particular the description of covering classification model and obtain better generalization ability of one-class classifiers.

Key words: Pattern recognition; Sparse distance metric learning; L1-norm; One-class classifier

1 引言

在故障诊断、身份识别、异常检测、疾病监测等应用中普遍存在单类分类问题,与二分类和多分类问题不同,单类分类器仅有目标类样本可用[1~3].单类分类器的设计目标是确定目标类样本的覆盖函数,使得目标类样本因覆盖被接受,而非目标类样本未被覆盖而被拒绝.

针对单类分类问题,国内外已有工作按其原理可分为密度函数法、神经网络模型、数据聚类模型和边界描述方法:(1)密度函数法就是通过参数化或非参数化方法估计训练样本的密度分布模型,通过设置密度门限,当测试样本点密度小于给定门限时将被拒绝,例如高斯混合模型和 Parzen 窗函数法.在目标类样本集维数较低且样本数较多时密度函数法比较有效,但在高维有限样本情况下密度估计的方法不能真实反映模式的统计分布特征,难以对目标类数据的稀疏区域做出正确判决;

收稿日期:2010-08-20;修回日期:2011-08-31
基金项目:国家自然科学基金(No.61071199);河北省自然科学基金(No.F2008000891, No.F2010001297);中国博士后自然科学基金(No.20080440124);第二批中国博士后特别资助基金(No.200902356)

(2)神经网络模型主要包括自动编码器(Auto-Encoders)、学习矢量量化(Learning Vector Quantization,LVQ)和自组织特征映射(Self-Organizing Map,SOM)等.神经网络模型对非线性问题有较好的分类效果,其缺点在于网络训练需预先确定较多参数,如网络隐层数和每层神经元数目;(3)数据聚类模型认为目标类样本满足某种聚类假设,对数据进行聚类,以测试样本到最近簇类中心的距离判定是否为目标类,如 k-means、k-centers 和单类聚类数据描述模型[4];(4)边界描述方法通过对目标类数据的学习,形成一个围绕目标类的边界,如超平面、超球等,并且最小化目标类数据支撑域的体积,以达到错误接受率最小的目的,代表方法是支持向量数据描述(Suppor Vector Data Description,SVDD)[5]和单类支持向量机(One-Class Support Vector Machines,OCSVM)[6],还有一些非参数的边界描述方法,如最近邻(1-Nearest Neighbour,1-NN)、k 近邻(k-Nearest Neighbour ,k-NN)法和最小生成树覆盖模型(Minimum Spannning Tree Class Descriptor,MSTCD)[7].数据聚类模型和边界描述方法能提供对目标类样本较直观的数据分布描述,然而数据聚类模型和边界描述方法的性能都依赖于样本间的距离测度,通常这些算法采用经典欧氏测度来描述样本间相似关系,面临的问题是欧氏测度忽略了样本分布的统计特性,采用欧氏测度有时无法合理揭示数据的内在分布结构,从而影响这些方法对数据的描述能力.如果能从目标类样本中学习到一个适合描述样本间相似关系的距离测度,则将会保证这些方法对数据的描述性能,提高其推广能力,这就是本文工作的出发点.

针对距离测度学习的问题,已经有不少学者做了相关研究[8~14].基本思想是通过引入有限的同类样本相似约束和不同类样本不相似约束学习一个距离测度,学习的距离测度被用于改善数据聚类或分类.如文献[8]中提出的用于改善 kNN 分类性能的距离测度学习算法,文献[9,10]中提出的用于改善数据聚类或分类的距离测度学习算法.核矩阵距离测度学习则是一种更加灵活的距离测度学习算法,这种距离测度学习等价于学习样本空间的非线性变换,如文献[13,14]中提出用于半监督聚类的核矩阵距离测度学习算法.这些算法都从不同类别的最佳划分出发,使得同类样本间距离较近,不同类样本间距离较远,通过最小化相应的损失函数使得不同类别之间存在较大的间隔.然而对于仅有单类数据的距离测度学习的相关研究存在不足,文献[15]中提出一种流形嵌入的支持向量数据描述,用测地距离代替原空间欧氏距离,对于高维空间的流形数据描述取得不错的效果,该方法的缺点是测试样本到目标类样本覆盖模型的距离计算比较复杂.鉴于此,本文提出 L1 范数约束的单类数据距离测度学习算法,该算法通过引入样本先验分布正则化项和距离测度稀疏性约束,能有效解决高维空间小样本情况下单类数据的距离测度学习问题,学习的距离测度能容易地嵌入到单类分类器中,进而改善分类器的性能.

2　一类数据距离测度学习问题

距离测度是描述样本间相似关系的一种度量,对于不少数据集,直接使用某种固定相似性度量(如欧氏距离)有时难以反映样本间相似关系,如何保持二者的一致性是距离学习的关键问题.单类数据距离测度学习就是利用已有目标类样本间相似程度的先验知识来学习距离测度,使之能较好的描述数据样本间相似关系.

2.1　目标类样本间相似约束

给定一个有 n 个目标类样本的训练集 $\boldsymbol{X}=\{\boldsymbol{x}_i\in R^N\}_{i=1}^n$,单类数据距离学习目标是学习一个马氏距离测度 $\boldsymbol{M}$:

$$d_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{M}}^2=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}_i-\boldsymbol{x}_j)\quad(1)$$

为保证学习距离的有效性,式中 $\boldsymbol{M}\in R^{N\times N}$是一个半正定矩阵,当 $\boldsymbol{M}$ 为单位阵时即为欧氏距离.假定存在一个线性变换 $\boldsymbol{G}$ 使得 $\boldsymbol{M}=\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$,则该问题等价于学习一个线性变换后的欧氏距离,即有:

$$\begin{aligned}d_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j)&=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_{\boldsymbol{M}}^2\\&=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}_i-\boldsymbol{x}_j)\\&=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{x}_i-\boldsymbol{x}_j)\\&=\|\boldsymbol{G}(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2\end{aligned}\quad(2)$$

目标样本间相似约束根据先验知识确定,当没有先验知识可用时,可计算每个目标样本欧氏距离的 k 近邻与其组成的点对作为相似约束先验知识,定义关联矩阵 $\boldsymbol{K}$ 表示目标类样本间相似约束:

$$K_{ij}=\begin{cases}1, & \boldsymbol{x}_i\text{ 与 }\boldsymbol{x}_j\text{ 相似}\\0, & \boldsymbol{x}_i\text{ 与 }\boldsymbol{x}_j\text{ 不相似}\end{cases}\quad(3)$$

学习的距离测度就是最小化样本间的相似约束,即最小化下式的损失函数:

$$\begin{aligned}loss&=\frac{1}{2}\sum_{i,j=1}^n\|\boldsymbol{G}(\boldsymbol{x}_i-\boldsymbol{x}_j)\|_2^2K_{ij}\\&=\sum_{i,j=1}^n(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x}_j)K_{ij}\\&=\sum_{i,j=1}^n(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}_j)\boldsymbol{K}_{ij}\\&=\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{X}\boldsymbol{D})-\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{X}\boldsymbol{K})\\&=\mathrm{tr}(\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M})-\mathrm{tr}(\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M})\\&=\mathrm{tr}(\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{K})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M})\\&=\mathrm{tr}(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M})\end{aligned}\quad(4)$$

式中 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$ 表示样本集,$\boldsymbol{D}$ 为对角阵,其

对角阵元素等于矩阵 $\boldsymbol{K}$ 相应列元素的和，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{K}$ 表示拉普拉斯矩阵.

2.2 数据分布先验知识约束

在不引入数据分布先验知识约束情况下，最小化式(4)将收敛于零解.这里引入一个基于信息论的先验分布正则化项[9]，给定一个初始矩阵 $\boldsymbol{M}_0$ 确定关于目标类数据分布的先验知识，学习目标是最小化由凸函数 $g(\boldsymbol{X})=-\log\det(\boldsymbol{X})$ 确定的所求马氏距离测度 $\boldsymbol{M}$ 与先验分布 $\boldsymbol{M}_0$ 之间的 Bregman 偏移

$$D_g(\boldsymbol{M},\boldsymbol{M}_0)=\mathrm{tr}(\boldsymbol{M}\boldsymbol{M}_0^{-1})-\log\det(\boldsymbol{M}\boldsymbol{M}_0^{-1})-N \tag{5}$$

式中 N 是数据维数.最小化式(5)即使得学习的距离测度 $\boldsymbol{M}$ 与数据先验分布 $\boldsymbol{M}_0$ 尽可能的接近.该优化是基于信息论的，因为最小化式(5)也即最小化以 $\boldsymbol{M}$ 和 $\boldsymbol{M}_0$ 为协方差矩阵的两个多变量高斯函数的相关熵.当样本分布为正态分布时，$\boldsymbol{M}_0$ 可取为样本的协方差矩阵，当无法确定其分布时可取为确定欧氏距离的单位阵，因为单位阵给出了数据分布无偏的先验分布信息.

2.3 距离测度稀疏性约束

已有研究表明，样本空间特别是高维样本空间，样本不同特征间具有稀疏相关性的特点，样本的精度矩阵(协方差矩阵的逆)中非对角线元素往往仅有少量的非零元素，稀疏距离测度能真实反映样本空间的特征相关性.为了获得距离测度的稀疏解，可以最小化距离测度 $\boldsymbol{M}$ 的 L_0 范数，然而 L_0 范数的求解是 NP 困难的.在样本空间足够稀疏时，可以等价的用 L_1 范数替代，即最小化距离测度 $\boldsymbol{M}$ 中元素绝对值的和[16].另一方面，样本的精度矩阵中对角线元素往往有比非对角线元素较大的值，不具有稀疏性.因此这里对学习的距离测度 $\boldsymbol{M}$ 施加非对角线元素稀疏性约束，即最小化式(6)来寻求距离测度 $\boldsymbol{M}$ 的稀疏解.

$$\|\boldsymbol{M}\|_{1,\mathrm{off}}=\sum_{i\neq j}|M_{i,j}| \tag{6}$$

上述的数据分布先验知识约束和距离测度的稀疏性约束相互补充，稀疏性约束就是寻求一个稀疏的距离测度来控制测度复杂性，而数据先验分布束使得学习的距离测度尽可能与样本先验测度相近.

综合上述先验分布正则化项和距离测度稀疏性约束，最终距离测度学习的目标函数为

$$\begin{aligned}\min\quad&\mathrm{tr}(\mu\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M})+\mathrm{tr}(\boldsymbol{M}_0^{-1}\boldsymbol{M})\\&-\log\det(\boldsymbol{M})+\lambda\|\boldsymbol{M}\|_{1,\mathrm{off}}\\=\;&\mathrm{tr}((\mu\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}+\boldsymbol{M}_0^{-1})\cdot\boldsymbol{M})\\&-\log\det(\boldsymbol{M})+\lambda\|\boldsymbol{M}\|_{1,\mathrm{off}}\\\text{subject to:}\;&\det(\boldsymbol{M})\geqslant 0\end{aligned} \tag{7}$$

式中略去了关于 $\boldsymbol{M}_0$ 的常数项，参数 μ 和 λ 分别是目标样本相似性和距离测度稀疏性与先验分布 $\boldsymbol{M}_0$ 正则化项的平衡参数.目标函数中 $-\log\det(\boldsymbol{M})$ 和 $\|\boldsymbol{M}\|_{1,\mathrm{off}}$ 是关于 $\boldsymbol{M}$ 的凸函数，另一方面 $\mathrm{tr}((\mu\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}+\boldsymbol{M}_0^{-1})\cdot\boldsymbol{M})$ 为线性项，所以总的目标函数为凸函数，存在一个全局最优解.

2.4 距离测度学习算法

式(7)的优化可以通过正定优化算法求解，然而正定优化的求解比较复杂[17].这里采用一种效率较高的分块协调下降算法.与正定优化不同，该算法通过优化矩阵 $\boldsymbol{M}$ 的逆来求解而不是直接优化矩阵 $\boldsymbol{M}$.令 $\boldsymbol{W}$ 表示矩阵 $\boldsymbol{M}$ 逆的估计，$\boldsymbol{S}=\mu\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}+\boldsymbol{M}_0^{-1}$，该算法一次只优化矩阵 $\boldsymbol{W}$ 的一行和一列，直到解收敛.具体的，将 $\boldsymbol{W}$ 和 $\boldsymbol{S}$ 分块为：

$$\boldsymbol{W}=\begin{bmatrix}\boldsymbol{w}_{11}&\boldsymbol{w}_{12}\\\boldsymbol{w}_{12}^{\mathrm{T}}&\boldsymbol{w}_{22}\end{bmatrix},\boldsymbol{S}=\begin{bmatrix}\boldsymbol{S}_{11}&\boldsymbol{S}_{12}\\\boldsymbol{S}_{12}^{\mathrm{T}}&\boldsymbol{S}_{22}\end{bmatrix} \tag{8}$$

式中 $\boldsymbol{w}_{11}\in R^{(N-1)\times(N-1)}$，$\boldsymbol{w}_{12}\in R^{(N-1)}$.文献[17]表明 $\boldsymbol{w}_{12}$ 的解满足：

$$\boldsymbol{w}_{12}=\arg\min\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{w}_{11}^{-1}\boldsymbol{y}:\|\boldsymbol{y}-\boldsymbol{S}_{12}\|_{\infty}\leqslant\lambda\} \tag{9}$$

这是一个简单界二次规划问题，可通过其对偶问题求解：

$$\min_{\alpha}\left\{\frac{1}{2}\|\boldsymbol{w}_{11}^{1/2}\alpha-\boldsymbol{w}_{11}^{-1/2}\boldsymbol{S}_{12}\|^2+\lambda\|\boldsymbol{\alpha}\|_1\right\} \tag{10}$$

如果式(10)解为 $\boldsymbol{\alpha}$，则式(9)解为 $\boldsymbol{w}_{12}=\boldsymbol{w}_{11}\boldsymbol{\alpha}$.每次优化最后一列后，将下一目标列交换至最后一列优化，更新 $\boldsymbol{W}$，迭代直到解收敛.研究表明如果用一个正定阵初始化 $\boldsymbol{W}$，则优化过程中 $\boldsymbol{W}$ 保持正定且不可逆，且这个优化过程是收敛的[17].求得 $\boldsymbol{W}$ 后可通过 $\boldsymbol{W}$ 的逆求得 $\boldsymbol{M}$.距离测度的分块协调下降学习算法流程图如图 1 所示.

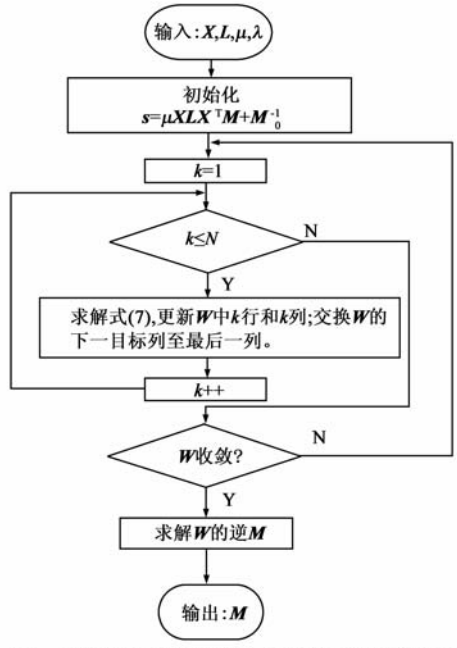


图1 距离测度的分块协调下降学习算法流程图

3 嵌入距离测度的单类分类器

学习的距离测度能容易地嵌入到基于欧氏距离的单类分类器中,如 k-means、1-NN、k-NN,SVDD 等.在这些算法中,用学习的距离测度代替欧氏距离,即计算$(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}_i-\boldsymbol{x}_j)$.然而对于大型数据集,这种计算是比较耗时的.如前所述,如果存在一个线性变换 $\boldsymbol{G}$ 使得 $\boldsymbol{M}=\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$,则样本间马氏距离测度等价于相应线性变换空间的欧氏距离测度,即有:

$$d_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}_i-\boldsymbol{x}_j)_2^2=\|\boldsymbol{G}(\boldsymbol{x}_i-\boldsymbol{x}_j)\| \tag{11}$$

这里采用一种 $\boldsymbol{M}$ 的特殊分解方式:$\boldsymbol{M}=\boldsymbol{M}^{1/2}\boldsymbol{M}^{1/2}$,对称阵 $\boldsymbol{M}^{1/2}$定义为 $\boldsymbol{M}^{1/2}=\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}$,$\boldsymbol{\Gamma}$ 是正定阵 $\boldsymbol{M}$ 的特征向量,$\boldsymbol{\Lambda}$ 是对应的特征值.对数据集运用线性变换 $\boldsymbol{M}^{1/2}$ 后,则在变换空间的欧氏测度等价于原特征空间学习得到的距离测度.如此,只需通过一个线性变换,学习的距离测度就可以进行维数约简[18,19],然后可嵌入各种基于欧氏距离的一类分类器中.

4 实验仿真与分析

为了验证本文提出距离测度学习算法的有效性,本文进行了三组实验,第一组采用两种随机产生的满足高斯分布的三类数据样本(其中一类作为目标类样本),后两组分别采用 UCI 数据集和 MNIST 手写体数据集.实验中距离测度学习的参数 μ 和 λ 经两次交叉验证选择,先验分布矩阵 $\boldsymbol{M}_0$ 取单位阵.单类分类器选用了 1-NN、k-NN、k-means 和 SVDD,其中 kNN 中 k 值的选取经最小化留一法错误率优化,k-means 算法中 $k=5$,SVDD 采用高斯核,核带宽 $\sigma=8$,所有分类器的容错率设为 0.1.

单类分类器的性能评价常采用 ROC(Receiver Operating Characteristic)曲线,ROC 是单类分类器目标类接受率与非目标类接受率比值的函数,其通过对单类分类器决策变量阈值的变化提供了单类分类器的动态性能观测.AUC(Area Under ROC Curve)是进一步衡量单类分类器性能的评价指标,其反映单类分类器的综合性能,故本文采用 AUC 作为分类器的评价指标.

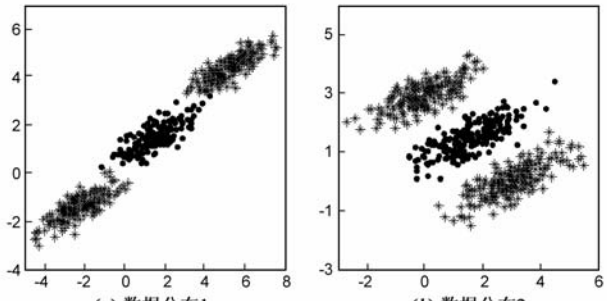
4.1 高斯分布样本点分类实验

本组实验比较如图 2 所示两种数据分布下采用欧氏距离、马氏距离和学习距离测度的单类分类器性能,各数据簇按高斯分布抽样,每类包含 200 个样本.

实验时选取中间的数据簇样本作为目标类,其余两类作为非目标类.实验重复 20 次,每次实验中样本数据按照图中相应高斯分布重新抽样生成.实验结果如表 1 所示,其中测试集表示为:目标类测试样本数/非目标类测试样本数.

表 1　两种高斯分布数据样本点的实验结果

数据集	训练集	测试集	距离测度	1-NN	k-NN	k-means	SVDD
数据分布 1	200	200/400	Euclidean	0.922	0.9934	0.9954	0.9977
			Mahalanobis	0.9276	0.9931	0.9967	0.9978
			LM	0.9299	0.9934	0.9971	0.9985
数据分布 2	200	200/400	Euclidean	0.9663	0.9992	0.9991	0.9565
			Mahalanobis	0.9591	0.9997	0.9998	0.9998
			LM	0.964	0.9996	0.9998	0.9988



(a) 数据分布1　　(b) 数据分布2

图2　两种高斯分布数据

实验结果表明,采用学习的距离测度单类分类器表现出更好的描述性能,仅对数据分布 2,1-NN 法的性能略有下降,其原因是数据分布 2 中目标类样本中心与非目标类样本距离较近,使得边界样本的错误率增加,这也使得采用学习距离测度 SVDD 性能较采用欧氏距离的 SVDD 获得较大提高.同时马氏距离测度也对一类分类器的性能有所改进.图 3 示意了对两种数据分布分别采用欧氏距离测度的 SVDD 描述边界和学习距离测度对应线性变换空间欧氏距离测度的 SVDD 描述边界.从图中可看出,直接在样本空间采用欧氏距离的 SVDD 对数据分布描述不够紧,存在误判率较高的问题,而进行距离学习后,对应线性变换空间的 SVDD 对数据的描述更为紧,更能实现对目标数据集的有效覆盖.

4.2 UCI 数据集实验

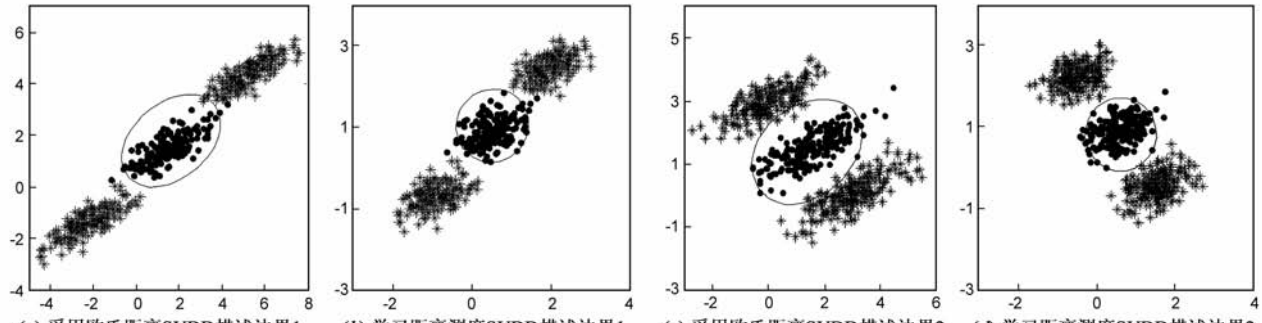
本组实验选择了 UCI 数据库中的 iris 数据集、wine 数据集、ionosphere 数据集以及 sonar 数据集作为研究对象.实验中每次随机选择一半目标类样本作为训练集,其余所有样本作为测试集.实验结果经 20 次重复实验取平均值,实验结果见表 2,其中测试集表示为:目标类测试样本数/非目标类测试样本数.UCI 数据集上距离

学习的优化时间见表 2 最后一列.

从表 2 实验结果看出,采用学习的距离测度后各种单类分类器都表现出更好的描述性能,可见将数据投影到合适的空间,建立与数据分布合适的距离测度有利于建立性能更佳的分类器.对于 wine 数据集单类分类器的描述性能改进最大.同时对于 iris、wine、ionosphere 三个数据集采用马氏距离也取得较好的效果,然而对于高维的 sonar 数据集,采用马氏距离无法描述样本间的相似关系,而采用学习的距离测度对于 1-NN、k-means 和 SVDD 的性能仍获得较大的提高.从表 2 中目标类距离学习的优化时间看出特征维数为 60 维的 sonar 数据集的距离学习过程仅需 50s 左右.

表 2 UCI 数据集的实验结果

数据集(维数)	目标类	训练集	测试集	距离测度	1-NN	k-NN	k-means	SVDD	优化时间(s)
iris(4)	vesicular	25	25/100	Euclidean	0.9152	0.974	0.9708	0.9665	1.4
				Mahalanobis	0.8968	0.9759	0.9860	0.9810	
				LM	0.9183	0.977	0.9807	0.9769	
	virginica	25	25/100	Euclidean	0.9332	0.9556	0.9535	0.9621	1.4
				Mahalanobis	0.9268	0.9608	0.9567	0.9713	
				LM	0.9575	0.9607	0.9605	0.9723	
wine(13)	class1	29	30/119	Euclidean	0.7503	0.9017	0.8984	0.8753	4.2
				Mahalanobis	0.9643	0.962	0.9612	0.9724	
				LM	0.9735	0.9907	0.9871	0.9912	
	class2	35	36/107	Euclidean	0.6717	0.8187	0.7684	0.8109	4.3
				Mahalanobis	0.8922	0.953	0.964	0.96	
				LM	0.8738	0.9424	0.9388	0.9217	
ionosphere(34)	good	112	113/126	Euclidean	0.8461	0.9545	0.9563	0.8134	12.6
				Mahalanobis	0.9405	0.96	0.9602	0.9566	
				LM	0.9218	0.962	0.9618	0.9617	
sonar(60)	mines	55	56/97	Euclidean	0.7642	0.8216	0.7019	0.5162	50.6
				Mahalanobis	0.3	0.3	0.3	0.3	
				LM	0.7828	0.8165	0.7341	0.6968	



(*a*) 采用欧氏距离SVDD描述边界1 (*b*) 学习距离测度SVDD描述边界1 (*c*) 采用欧氏距离SVDD描述边界2 (*d*) 学习距离测度SVDD描述边界2

图3 采用欧氏距离测度和学习距离测度的SVDD描述比较

4.3 MNIST 手写数字体集实验

本组实验的数据来源于 MNIST 手写体数字数据库,该数据库包括 0 ~ 9 共 10 类数字手写体样本.训练集有 6 万个样本,测试集有 1 万个样本.每一个样本都归一化到 28 × 28 大小.实验中分别选用数字体 1、3、5、7、9 做目标类,其余数字体做非目标类.目标类随机抽取 200 个样本进行训练.非目标类测试集每类随机抽取 200 个样本,实验结果经 10 次重复实验取平均值,结果见表 3,表格中最后一列为目标类距离学习的优化时间.

实验结果表明,采用学习的距离测度后单类分类器对于数字体样本分布表现出更好的描述性能,特别是对于 SVDD 的性能提高最为明显,采用学习距离测度后的 SVDD 对于实验的数字体都表现出最好的性能,此外马氏距离无法描述样本间相似关系,这表明本文引入的高维空间下样本特征间稀疏相关性约束下的距离测度学习是有效的合理的.此外,从表中最后一列可看出本文算法对高维数字体集的距离学习时间约 16min,这表明本文学习算法的高效性.

5 结论

本文提出一种高维空间下针对单类数据的稀疏距离测度学习算法,并将学习的距离测度应用于单类分

类器中.该算法充分考虑了高维空间中样本特征间稀疏相关性的特点,引入样本先验分布正则化项和距离测度 L1 范数稀疏性约束构造距离学习的目标函数,构造的目标函数具有全局最优解,可通过分块协调下降算法高效优化求解.学习的距离测度能容易的嵌入各种基于欧氏距离的一类分类器中.在人工和实际数据集上的实验结果表明相比欧氏距离采用学习的距离测度能有效改善单类分类器的描述性能,特别是能够较好地改善 SVDD 的描述能力,从而使得单类分类器具有更强的推广能力.

表 3　MNIST 手写数字数据集的实验结果

目标类	训练集	测试集	距离测度	1 – NN	k – NN	k – means	SVDD	优化时间(s)
数字体 1	200	200/1800	Euclidean	0.8805	0.9978	0.9967	0.9924	953.5
			Mahalanobis	0.2	0.2	0.2	0.2	
			LM	0.9572	0.997	0.997	0.997	
数字体 3	200	200/1800	Euclidean	0.7815	0.9246	0.9343	0.9054	959.7
			Mahalanobis	0.2	0.2	0.2	0.2	
			LM	0.8611	0.9272	0.9323	0.9363	
数字体 5	200	200/1800	Euclidean	0.8296	0.9376	0.9206	0.8114	955.3
			Mahalanobis	0.2	0.2	0.2	0.2	
			LM	0.8813	0.9343	0.942	0.9443	
数字体 7	200	200/1800	Euclidean	0.8596	0.9577	0.9553	0.9429	953.7
			Mahalanobis	0.2	0.2	0.2	0.2	
			LM	0.9142	0.9591	0.9584	0.9596	
数字体 9	200	200/1800	Euclidean	0.848	0.9543	0.9461	0.9403	956.8
			Mahalanobis	0.2	0.2	0.2	0.2	
			LM	0.9066	0.9604	0.9621	0.9641	

参考文献

[1] 潘志松,陈斌,缪志敏,等.One-Class 分类器研究[J].电子学报,2009,37(11):2496 – 2503.
Pan Zhi-Song, Chen Bin, Miao Zhi-Min, et al. Overview of study on one-class classifiers[J]. Acta Electronica Sinica, 2009,37(11):2496 – 2503.(in Chinese)

[2] Mahadevan S, Shah S L. Fault detection and diagnosis in process data using one-class support vector machines[J]. Journal of Process Control,2009,19(10):1627 – 1639.

[3] Mena L, Jesus A G. Symbolic one-class learning from imbalanced datasets: application in medical diagnosis[J]. International Journal on Artificial Intelligence Tools,2009,18(2):273 – 309.

[4] 陈斌,冯爱民,陈松灿,等.基于单簇类聚类的数据描述[J].计算机学报,2007,30(8):1325 – 1332.
Chen Bin, Feng Ai-Min, Chen Song-Chan , et al. One-cluster clustering based data description[J]. Chinese Journal of Computers,2007,30(8):1325 – 1332.(in Chinese)

[5] Lee K, Kim D W, Lee K H, et al. Density-induced support vector data description[J]. IEEE Transactions on Neural Networks,2007,18(1):284 – 289.

[6] Choi Y S. Least squares one-class support vector machine[J]. Pattern Recognition Letters,2009,30(13):1236 – 1240.

[7] Piotr J, Tax D M J, Elzbieta P, et al. Minimum spanning tree based one-class classifier[J]. Neurocomputing, 2009, 72(7 – 9):1859 – 1869.

[8] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. The Journal of Machine Learning Research,2009,10(1):207 – 244.

[9] Davis J V, Kulis B, Jain P, et al. Information-theoretic metric learning[A]. Proceedings of the 24th International Conference on Machine Learning[C]. Corvalis, United states: ACM International Conference Proceeding Series,2007(227).209 – 216.

[10] Xiang Shiming, Nie Feiping, Zhang Changshui. Learning a Mahalanobis distance metric for data clustering and classification[J]. Pattern Recognition,2008,41(12):3600 – 3612.

[11] Kulis B, Jain P, Grauman K. Fast similarity search for learned metrics[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(12):2143 – 2157.

[12] Qi Guo-Jun, Tang Jinhui, Zha Zheng-Jun, et al. An efficient sparse metric learning in high dimensional space via l1-penalized log-determinant regularization[A]. Proceedings of the 26th International Conference On Machine Learning[C]. Montreal, Canada: ACM International Conference Proceeding Series,2009(382).841 – 848.

[13] Yeung D Y, Chang H. A kernel approach for semi-supervised metric learning[J]. IEEE Transactions on Neural Networks, 2007,18(1):141 – 149.

[14] Mahdieh S B, Saeed B S. Kernel-based metric learning for semi-supervised clustering[J]. Neurocomputing,2010,73(7-9): 1352 – 1361.

[15] 陈斌,李斌,潘志松,等.流形嵌入的支持向量数据描述[J].模式识别与人工智能,2009,22(4):548 – 553.

Chen Bin, Li Bin, Pan Zhi-Song, et al. Support vector data description with manifold embedding[J]. Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence, 2009, 22(4): 548 – 553. (in Chinese)

[16] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210 – 227.

[17] Banerjee O, Ghaoui L E, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data[J]. Journal of Machine Learning Research, 2008, 9(7): 485 – 516.

[18] 刘中华,周静波,陈●,金忠.距离保持投影非显影降维技术的可视化与分类[J].电子学报,2009,37(8):1821 – 1824.
Liu Zhi-hua, Zhou Jing-bo, Chen Yi, Jin Zhong. Non-linear dimensionality reduction techniques of distance-preseving projection for visalization and classification[J]. Acta Electronica Sinica, 2009, 37(8): 1821 – 1824. (in Chinese)

[19] 邵超,黄厚宽.P-ISOMAP:一种新的对邻域大小不甚敏感的数据可视化算法[J].电子学报,2006,34(8):1497 – 1501.
Shao Chao, Huang Hou-kuang, et al. P-ISOMAP: a new isomap based data visualization algorithm with less sensitivity to the neighborhood size[J]. Acta Electronica Sinica, 2006, 34(8): 1497 – 1501. (in Chinese)

作者简介

胡正平 男(汉族),1970 年 8 月生于四川仪陇县,博士,教授.目前研究方向为统计学习理论与模式识别,医学图像处理.
E-mail: hzp@ysu.edu.cn

路 亮 男(汉族),1987 年生于山西,燕山大学通信与信息系统专业硕士研究生.主要研究方向:统计学习单类分类与空间距离度量学习.