基于高维空间典型样本 Steiner 最小树覆盖模型的一类分类算法

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

摘 要：最小生成树数据描述方法在刻画高维空间样本点分布时，将所有图形的边作为新增虚拟样本以提供同类样本分布描述，这种描述存在分支多覆盖模型复杂，且局部覆盖不够合理的问题。针对该问题，依据特征空间中同类样本分布的连续性规律，提出基于高维空间典型样本 Steiner 最小树覆盖模型的一类分类算法，该算法首先对目标类训练集进行样本修剪，去除冗余信息和噪声信息，选择最具代表性的样本作为训练集，然后对保留的典型样本构建 Steiner 最小树覆盖模型。算法分析和仿真实验结果表明，相比最小生成树数据描述，文中提出的方法能在较低覆盖模型复杂度的前提下更合理的描述目标类样本空间分布，构建更合理的覆盖模型，在分类正确率和适用样本规模上都表现出一定的优越性。

关键词：一类分类器；高维空间；最小生成树；Steiner 最小树
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1003-0530(2011)06-0874-09

A One-class Classification Algorithm Based on Steiner Minimal Tree of Typical Samples Covering Model in High-dimensional Space

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

Abstract: Minimum Spanning Tree Class Descriptor (MSTCD) describes the target class with the assumption that all the edges of the graph are also basic elements of the classifier which offers additional virtual training data for better description of sample distribution in high dimensional space. However, this descriptive model has too many branches, which makes the model more complicated, and its local coverage is not so reasonable. In this case, according to the continuity law of the feature space of similar samples, a one-class classification algorithm based on Steiner minimal tree of typical samples covering model is presented in this paper. The method first prunes the training set, eliminates redundant information and noise information and selects the most representative samples as a new training set; then it builds Steiner minimal tree covering model on the retained typical samples. Theoretical analysis and simulation experimental results show that the presented method can describe the distribution of target class more reasonably, construct more reasonable covering model without increasing the model complexity. It performs better than MSTCD in accuracy of classification and applicable sample size.

Key words: One-class classifier; High-dimensional space; Minimum spanning tree (MST); Steiner minimal tree (SMT)

1 引言

传统基于划分分类的模式识别方法一般需要多个类别的训练样本,用来设计两类和多类分类器。然而在实际应用中常常存在不少一类分类问题[1],例如基于生物特征的身份识别和验证[2],机器故障检测[3],异常行为检测[4],疾病检测[5],文本分类[6]等等。在这些问题中,有时无法获取多类样本,或者获取代价极高

收稿日期：2010 年 12 月 23 日；修回日期：无
基金项目：国家自然科学基金(61071199)；河北省自然科学基金(F2010001297)；河北省自然科学基金(F2008000891)；中国博士后自然科学基金(20080440124)；第二批中国博士后基金特别资助(200902356)

(如在机器故障检测中,为了获取异常样本而故意破坏机器设备),或者获取的异常样本不可信任(如在基于人脸图像的身份识别中,任意非本人的人脸图像或者非人脸图像都属于异常样本)。和两类分类问题不同,由于仅有一类样本数据可用,一类分类器的设计目标是确定目标类样本的覆盖函数,使得目标类的样本被接受,而非目标类的样本则被拒绝。

国内外研究者针对一类分类器设计展开不少工作,根据其原理将其大致分为四类: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模型,设置密度门限,测试样本点密度小于门限时将被拒绝,例如高斯混合模型和 parzen 窗函数法[7]。在目标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特征,难以对目标类数据的稀疏区域做出正确识别。(2)神经网络模型主要包括自动编码器(Auto-Encoders)、学习矢量量化 LVQ(Learning Vector Quantization)和自组织特征映射 SOM(Self-organizing Map)等[1]。神经网络模型对一些大规模和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不少参数,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 和 k-centers。数据聚类模型有较低的运算复杂度,然而这些方法对簇类中心的选择非常敏感,且簇类数 k 值的选取仍然是开放问题。为此文献[8]提出基于单簇聚类的数据描述方法,避免了簇类数选择的问题。(4)边界描述方法就是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 SVDD[9-11] 和 OCSVM[12-13],还有一些非参数的边界描述方法,如 1NN,kNN 法。数据聚类模型和边界描述方法对目标类样本有较直观的数据分布描述能力,但这些方法对于高维空间下样本非规则复杂分布形状描述不够紧凑,存在不少覆盖冗余。为此文献[14]提出基于最小生成树数据描述 MSTCD(Minimum Spanning Tree Class Descriptor)的一类分类器,MSTCD 利用训练样本的最小生成树构建目标类的覆盖模型,能较好的对非规则复杂数据分布进行描述,在高维空间小样本问题中表现出了良好的性能,然而由于该方法将所有训练样本最小生成树的边都作为新增虚拟样本以提供同类样本分布描述,使得其存在分支多、覆盖模型复杂的问题,且最小生成树描述存在局部覆盖不够合理的问题。针对这些问题,本文提出基于高维空间典型样本 Steiner 最小树覆盖模型的一类分类算法,该算法首先对训练集进行样本修剪,选择最具代表性的样本作为训练集,然后以保留的典型样本为节点构建目标类样本的 Steiner 最小树覆盖模型。相比 MSTCD,文中提出的方法能在较低模型复杂度的前提下更合理的描述目标类数据的空间分布。

2　最小生成树数据描述 MSTCD 原理

给定一个有 n 个目标类样本的训练集 $X=\left\{x_i\in R^N\right\}_{i=1}^n$,令 $\left\{x_i,x_j\right\}\in X\subset R^N$ 表示目标类中的两个样本,根据特征空间中同类样本的连续性规律[15],同类样本之间具有相互接近的性质。如果这两个样本描述现实中相似的物体,则在特征空间中它们也应该是近邻,且在这两个样本点之间存在一个连续变换,这个连续变换上的点也属于目标类。当两个样本点在特征空间中的位置很近时,可以用这两点的线性变换来近似这样一条曲线:

$$F(x_i,x_j)=x_i+\lambda_{ij}(x_j-x_i) \tag{1}$$

为了满足同类样本的连续性假设,仅需要在目标训练集中选择 $(n-1)$ 个线性变换,则同类样本将构成一个连续性整体。假定 $G=\{V,E\}$ 表示定义在目标训练集 X 上的全连接无向图,其中,$V=X$ 表示 G 的顶点集,$E=\{e_{ij}=(x_i,x_j)\}$ 表示 G 的边的集合。边的权重定义为:$w_{ij}=\|e_{ij}\|=\|x_i-x_j\|$,即两个顶点之间的欧氏距离。考虑寻找图 G 的一个子图 g:连接所有的顶点,没有环路,并且总的权重最小,这样一个子图提供了最可能的变换集合。这等价于寻找图 G 的最小生成树 MST(Minimum Spanning Tree),即寻找 $(n-1)$ 条边,这些边形成一个具有最小权重的树。

由于训练集是有限的,假定不仅最小生成树的边属于目标类,而且边的适当邻域也属于目标类。如果一个测试对象位于最小生成树描述的适当邻域内,则被判为属于目标类。定义点 x 到边 e_{ij} 的距离为 $d(x|e_{ij})$,则根据高维几何关系,点 x 在边 e_{ij} 上的投影点为:

$$P(x,e_{ij})=x_i+\frac{(x_j-x_i)^{\mathrm{T}}(x-x_i)}{\|x_j-x_i\|^2}(x_j-x_i) \tag{2}$$

如果 $P(x,e_{ij})$ 位于 e_{ij} 的边上，则 $d(x|e_{ij})$ 等于点 x 到 $P(x,e_{ij})$ 之间的距离，否则为其与两个顶点之间的最短距离。即有：

$$\begin{aligned} &if && 0\le\frac{(x_j-x_i)^{\mathrm{T}}(x-x_i)}{\left\|x_j-x_i\right\|^2}\le 1\\ &then && d(x|e_{ij})=\left\|x-P(x,e_{ij})\right\| \\ &else && d(x|e_{ij})=min\left\{\left\|x-x_i\right\|,\left\|x-x_j\right\|\right\}\\ &end \end{aligned} \tag{3}$$

测试对象 x 到最小生成树的距离定义为点 x 到 $(n-1)$ 条边的最短距离：

$$d_{MST}(x|X)=\min_{e_{ij}\in MST} d(x|e_{ij}) \tag{4}$$

如果定义 MSTCD 的覆盖半径为 θ，则当 $d_{MST}(x|X)\le\theta$ 时，x 被判为目标类，否则判为非目标类。图 1 示意了一组 banana 型数据的 MSTCD，其中图 1(a) 为一组二维 banana 形数据，图 1(b) 为相应的最小生成树，图 1(c) 为 MSTCD 覆盖模型。从图中可以看出 MSTCD 较好的表达了数据集内在的流形结构，实现了对数据集的有效覆盖。

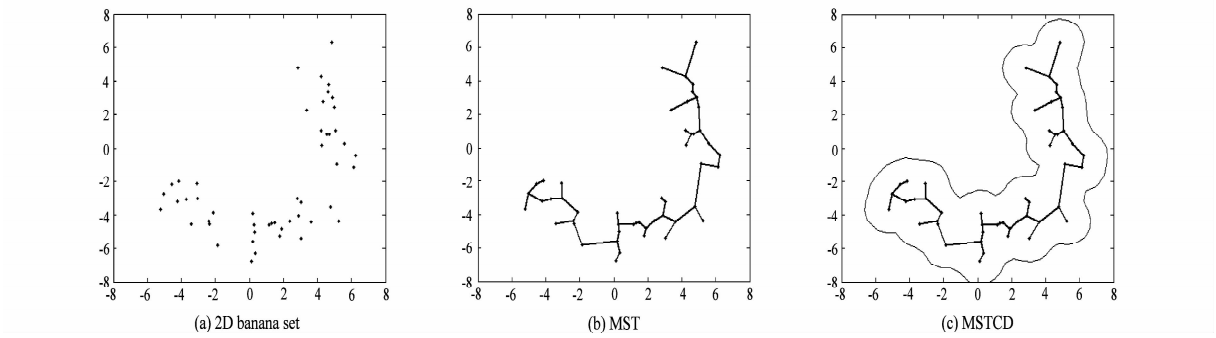


图 1 二维数据的最小生成树描述模型

Fig. 1 The MST covering model on 2D space

MSTCD 是基于边界描述的一类分类器，与其它数据覆盖模型相比，MSTCD 描述更为紧，更能反映数据的流形结构，适合于高维小样本问题。然而其也存在一些不足，由于该方法将所有训练样本最小生成树的边都作为新增虚拟样本以提供同类样本分布描述，使得其存在分支多、覆盖模型复杂的问题，且 MSTCD 存在局部覆盖不够合理的问题。

3 改进的典型样本 Steiner 最小树覆盖模型

针对 MSTCD 描述的不足，本文构造的典型样本 Steiner 最小树覆盖模型如图 2 所示，该系统模型首先对训练样本进行样本修剪，选择最具代表性的典型样本作为新的训练集，然后以新训练集构建目标类样本的 Steiner 最小树覆盖模型。

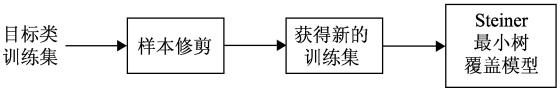


图 2 典型样本 Steiner 最小树覆盖模型原理框图

Fig. 2 Diagram of steiner MST covering model

3.1 样本修剪策略

最小生成树在描述空间样本点分布时，将样本点最小生成树的每条边作为新增虚拟样本以提供目标类样本分布描述。这所面临的问题是一些样本集中包含许多相似度很高的样本，存在大量冗余信息，这些冗余信息对目标类覆盖模型的贡献很小，但其存在却会大大增加模型复杂度，降低模型的推广能力。另一方面样本集中可能存在的噪声信息会造成模型不必要的覆盖区域，引起非目标类样本的错误接受。因此在设计覆盖模型时，有必要对训练集进行样本修剪，去除冗余信息和噪声信息，仅保留典型代表样本作为训练集，从而在较低的模型复杂度前提下提高分类器正确率。

这里采用如下的策略对目标类训练集进行样本修剪。给定一个有 n 个目标类样本的训练集 $X=\left\{x_i\in R^N\right\}_{i=1}^n$，首先构造训练集的 k 近邻有向图 G_k，图中节点代表目标类样本，如果节点 i 在节点 j 的 k 最近邻邻居，那么存在一条由节点 j 指向节点 i 的边，边的权重为样本间欧氏距离。将图 G_k 的有向边无向化，重复的边合并，求得样本间最短路径。对于任一样本点 x_i，若该点的 d_r 邻域内存在训练样本，则表明这些点密度较高，存在冗余样本；若该点与其最近邻间距离大于阈值 d_n，则认为该点属于噪声样本，将该样本从训练集中剔除。样本修剪具体算法如下：

(1) 计算每个目标类样本的 k 近邻，构建 k 近邻有

向图 G_k,将 G_k 无向化,求得样本间最短路径矩阵 $D_G(i,j)$。

(2)从有向图 G_k 中选择入度最大(入度相等则随机选择)的样本点 x_i,若该点与其最近邻距离 $d(x_i)\geq d_n$,$X=X-x_i$;否则求得到该点路径小于阈值 d_r 的点集 S_i,$S_i=S_i+x_i$,$X=X-S_i$。

(3)如果 $X\neq\varnothing$,返回(2),否则转向(4)。

(4)计算每个点集 S_i 的中心作为该点集的代表样本。

上述算法中需要确定三个参数:k、d_r 和 d_n。其中 k 值控制邻域的大小,为了获得较稳定的数据结构描述,通常取较小的 k 值,以使得 G_k 为连通图为宜。d_n 控制噪声样本的离群度,令 d_{mean} 表示 G_k 中边的平均长度,通常认为当 $d_n\geq(3\sim5)d_{mean}$ 时,样本属于噪声样本。d_r 控制样本修剪比例,其值越大,保留的典型样本越少;反之,保留的典型样本越多。三个参数中,d_r 对样本修剪结果影响最大,其通过控制典型样本个数控制覆盖模型的描述精度和推广能力。图 3 示意了上述 banana 形数据在 k 值固定为 3,不同参数 d_r 下经样本修剪的最小生成树覆盖模型,从中可以看出 d_r 越大,保留的典型样本越少,描述边界越光滑,描述精度较低,但推广能力较强;d_r 越小,保留的典型样本越多,描述边界越复杂,描述精度较高,但推广能力较弱,当 d_r 足够小时,覆盖模型即为原训练集的最小生成树数据描述。实际应用时,可通过交叉验证来实现模型描述精度和推广能力的平衡。

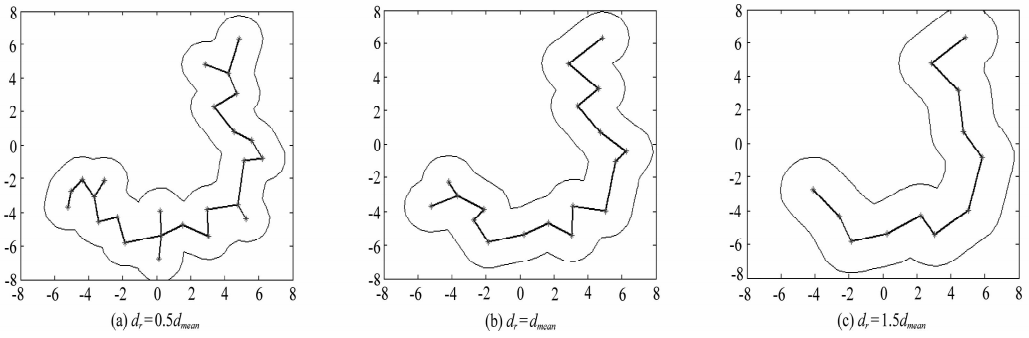


图 3　banana 形数据集上不同参数 d_r 作用下的 MSTCD

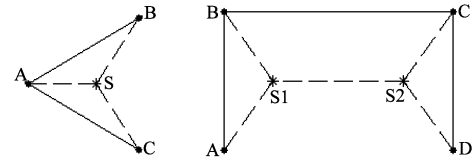
Fig. 3　The different MSTCD covering model on the banana set with parameter d_r

3.2　Steiner 最小树覆盖模型

MSTCD 用最小生成树 MST 描述目标类样本空间分布,MST 是连接样本点的最小长度的树,其能提供样本连续性描述以构建目标类覆盖模型,然而 MST 存在对样本局部分布描述不够合理的问题,影响覆盖模型的识别准确率。Steiner 最小树 SMT(Steiner Minimal Tree)是比 MST 长度更小的生成树,其通过引入 Steiner 点以使得树的总长度最小,这种长度更小的 SMT 能提供比 MST 更合理的目标类样本分布描述,从而构建更合理的覆盖模型。

3.2.1　SMT 问题

SMT 问题描述如下:给定空间中点集 $V=\{P_1,P_2,\cdots,P_n\}$,要求生成连接点集 V 所有顶点的最小树 $T(V)$。与最小生成树 MST 问题的不同之处在于,SMT 允许引入辅助点,以使得生成树的总长度最小。假设已经给定 n 个顶点,需要引进的辅助点数至多为 $n-2$,此种点成为 Steiner 点。过每一 Steiner 点至多有三条边通过,若为三条边,则它们两两交成 120°角;若为两条边,则此 Steiner 点与某一给定的点重合,且此两条边的夹角必大于或等于 120°。图 4 示意了三个顶点和四个顶点的 MST 与 SMT。图中实线代表 MST,虚线代表 SMT,"*"号代表 Steiner 点。



(a) 三个顶点的MST与SMT　(b) 四个顶点的MST与SMT

图 4　SMT 示意图

Fig. 4　Diagram of SMT

3.2.2　构建 SMT

SMT 问题在集成电路设计、交通线路规划,无线通信等方面有着广泛应用,但其求解是 NP 困难的,为此一些学者提出了启发式算法[16-18],其中一种代表性的方法就是基于 MST 的近似算法。

给定空间点集 V,这里采用一种嵌入 Steiner 点启发式算法[19]。该算法首先构建样本点的 MST,然后对 MST 的相邻边进行分析,嵌入 Steiner 点时试图满足其角度条件,即 Steiner 树中两条相邻边的夹角都大于或

等于 120°。其具体算法如下：

(1)构建点集 V 的 MST。

(2)对于 MST 的每条边(P_i,P_j),按以下步骤执行：

(a)求得与该边夹角最小的边(P_i,P_k)或(P_j,P_k),P_k 可为给定顶点或 Steiner 点。

(b)如果该夹角小于 120°,那么

ⅰ. 在点 P_k 上放置新的 Steiner 点 S_n。

ⅱ. 去除边(P_i,P_j)和(P_j,P_k),这两条边在循环(2)中将不考虑。

ⅲ. 添加新边(S_n,P_i)、(S_n,P_j)和(S_n,P_k)。

(3)局部优化求得 Steiner 点坐标。

图 5 示意了上述经样本修剪后的一组 banana 形数据的 MST 覆盖模型和 SMT 覆盖模型。从图 5(b)可以看出得到的 SMT 相比 MST 可能增添一些很小的树分支,通过移去这些边和顶点可得到一个更稀疏的覆盖模型,这可通过如下的简单修剪实现：

(1)求得给定顶点中与 Steiner 点相连的顶点集 V_S。

(2)对于点集 V_S 中每个顶点 P_i,计算该点与与其相连 Steiner 点 P_S 间距离 $d(P_i,P_S)$,若 $d(P_i,P_S)\leq T$,去除顶点 P_i 和边(P_i,P_S),与顶点 P_i 相连的边连接至对应 Steiner 点 P_S;若 $d(P_i,P_S)>T$,保持原图结构。

由此可得到修剪的 SMT 覆盖模型,如图 5(c)所示,从图中可以看出,修剪的 SMT 覆盖模型具有更低的模型复杂度,描述边界更为光滑,图结构变化很小。

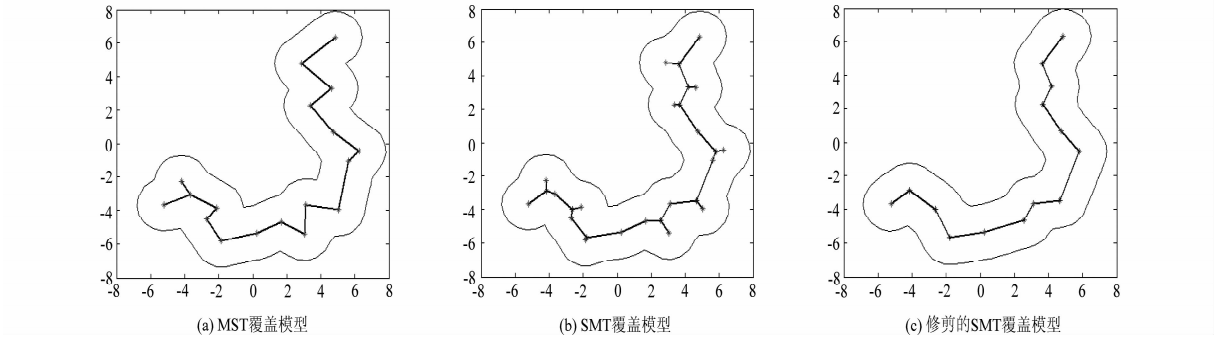


图 5 banana 形数据的 MST 与 SMT 覆盖模型示意图

Fig. 5 Comparison of different covering model on banana datasets

3.2.3 覆盖半径设置

在样本数目较多时,SMT 覆盖模型的半径通过原训练集样本的拒绝比 ε 来控制,为此计算所有训练样本到 SMT 的距离,进行从小到大排序,然后根据拒绝比 ε 选择第 $n\times(1-\varepsilon)$ 个样本到 SMT 的距离作为覆盖半径。在高维小样本情况下,样本修剪失去意义,几乎所有训练集样本都为 SMT 的节点,参照 MSTCD[14],覆盖半径可根据 SMT 中边的长度估计得到。

3.3 和 MSTCD 的比较与分析

相对于 MSTCD,本方法表现出以下优势：

(1) 分类正确率有望提高。由于对训练集进行了样本修剪,本文构造的覆盖模型有着更强的泛化能力,且通过构建 SMT 代替 MST,能构建目标类数据的更合理覆盖模型,分类正确率可能更高。

(2) 模型复杂度更低。训练集经过修剪后,用于构建 SMT 的典型样本大大减少,而覆盖模型的复杂度与图的边数相关,即与图的节点数相关,相对于 MSTCD 以所有训练集样本为节点构建 MST,本文构造的覆盖模型复杂度更低。

(3) 可适用于较大规模训练集。MSTCD 以所有训练集样本为节点构建 MST,使得其难以向大规模训练集推广,本文提出的样本修剪策略能大大减少训练集个数,因此可适用于较大规模训练集。

当然,这些优势是通过额外的训练时间获得的,样本修剪过程中采用 K-D 树方法可以 $o(n\log n)$ 的复杂度构建 k 近邻图,采用改进型 Dijkstra 算法计算训练样本间最短路径矩阵的复杂度为 $o(n\log n+E)$,其中 E 为 k 近邻图边数;假定样本修剪后保留的典型样本个数为 m,构建 SMT 在最坏情况下复杂度为$o(m^3)$[19]。而对于 MSTCD 来说,构建 MST 的复杂度仅为

$$o\left(\frac{n^2-n}{2}\log n\right)$$[14]。

4 仿真实验

为了验证本文提出算法的有效性和合理性,采用 UCI 数据库、MNIST 手写体数字库、MIT-CBCL 人脸识别数据库进行了实验,并将本文方法与 1NN、kNN、SVDD、MSTCD 四种方法进行对比,其中 kNN 通过最小化留一法错误率优化 k 值,SVDD 采用高斯核函数,核

宽度 $\sigma=5$,所有分类器的可容忍错误率设为 $\varepsilon=0.1$。

一类分类器的性能评价常采用 ROC(Receiver Operating Characteristic)曲线[20]。ROC 是一类分类器目标类接受率与非目标类接受率比值的函数,其通过对一类分类器决策变量阈值的变化提供了一类分类器的动态性能观测。AUC 是进一步衡量一类分类器性能的评价指标,其反映了一类分类器的综合性能,故本文采用 AUC 值作为分类器的评价指标。

4.1 UCI 数据集分类实验

本组实验选择了 UCI 数据库中的 iris 数据集、letter 数据集、landsat 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 10 次重复实验取平均值,实验结果见表 1,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。

从实验结果可以看出,本文提出的算法在各个数据集上都表现出良好的性能。对于低维的 iris 数据集,与 SVDD 有相当的描述性能,优于 MSTCD。对于高维数据,SVDD 难以有效的描述,本文方法性能略优于 MSTCD,这表明本文方法是有效可行的,能够用较低的模型复杂度实现对目标类数据集的较好覆盖。

表 1　UCI 数据集实验结果

Tab. 1　Experimental results of different methods trained on the UCI data sets

数据集(维数)	目标类	训练集	测试集	1NN	kNN	SVDD	MSTCD	本文方法
iris(4)	vesicular	25	25/125	0.9248	0.9780	**0.9840**	0.9788	0.9804
	virginica	25	25/125	0.9129	0.9523	0.9578	0.9537	**0.9584**
letter(16)	A	394	395/19211	0.9681	0.9972	0.9941	0.9973	**0.9973**
	B	383	383/19234	0.9697	0.9868	0.9804	0.9879	**0.9887**
	C	368	368/19264	0.9664	0.9941	0.9781	0.9943	**0.9943**
landsat(36)	1 类	536	461/1539	0.9442	0.9908	0.8562	0.9908	**0.9909**
	2 类	239	224/1776	0.7969	0.9883	0.7669	0.9888	**0.9890**
	3 类	480	397/1603	0.9286	0.9708	0.9117	0.9716	**0.9716**
sonar(60)	mines	55	56/97	0.7647	0.8014	0.5	0.8066	**0.8104**

4.2 MNIST 手写体数字识别实验

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。

1. 实验 1

本组实验分别用数字体 1、3、5、7、9 做目标类,其余数字体做非目标类,比较本文方法与 1NN、kNN、SVDD、MSTCD 四种方法的分类性能。实验中从训练集随机抽取 500 个目标类样本作为训练集,从测试集每类随机抽取 500 个样本作为测试集。实验结果经 10 次重复实验取平均值,结果表 2。

表 2　MNIST 数字体识别实验结果

Tab. 2　Experimental results of different methods trained on the MNIST data set

目标类	训练集	测试集	1NN	kNN	SVDD	MSTCD	本文方法
1	500	500/4500	0.9109	0.9966	0.9884	0.9974	**0.9976**
3	500	500/4500	0.7999	**0.9345**	0.9102	0.9259	0.9301
5	500	500/4500	0.8454	0.9328	0.8867	0.9516	**0.9570**
7	500	500/4500	0.8732	0.9520	0.9496	0.9584	**0.9606**
9	500	500/4500	0.8562	0.9388	0.9587	0.9572	**0.9590**

实验结果表明,本文提出的方法在数字体识别中表现出了优越的性能。对实验中的数据,本文方法都优于 SVDD 和 MSTCD,仅对数字体 3 性能略低于 kNN。实验中经样本修剪后的训练集样本数为 160 左右,这

表明提出的典型样本 SMT 覆盖模型比较合理,该模型能在较低模型复杂度前提下较好的描述数字体在高维空间的分布。

2. 实验 2

本组实验分别用数字体 0、2、4、6、8 做目标类,其余数字体做非目标类,比较在不同 d_r 值下经样本修剪后的典型样本集训练的 MST 覆盖和 SMT 覆盖模型的分类性能。实验时分别计算了在 20 个不同 d_r 值下的 MST 覆盖和 SMT 覆盖模型的 AUC 值,结果如图 6 所示。分析数据可得到如下结论:

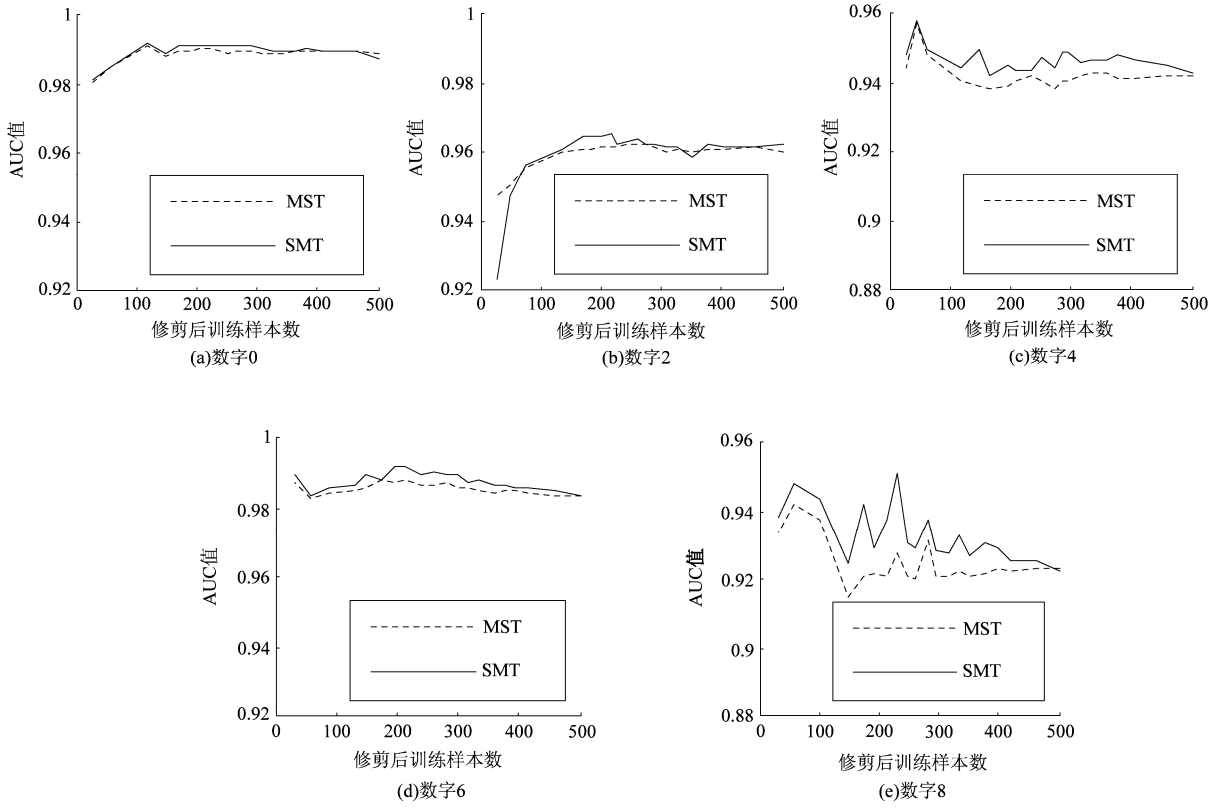


图 6 不同 d_r 下经样本修剪的 MST 覆盖和 SMT 覆盖模型性能比较

Fig. 6 Performance comparison of MST and SMT

(1)本文提出的样本修剪策略是有效的合理的。从图 6 可以看出,对于实验的每个数字体,不论 MST 还是 SMT,经样本修剪后的典型样本集训练的 MST 或 SMT 覆盖模型其性能都优于用全部样本集的训练结果,典型样本训练的 MST 或 SMT 覆盖模型可以用较少的边实现对目标类的较好覆盖,有着更强的推广能力。

(2)SMT 覆盖是比 MST 更为合理的覆盖模型。从图 6 可以看出,对于实验的数字体,不论在训练样本数目多少的情况下,SMT 覆盖模型的性能都普遍优于 MST 覆盖,这表明相比 MST,SMT 能更合理的描述目标类样本的空间分布。

4.3 MIT-CBCL 人脸识别实验

本组实验数据来源于 MIT-CBCL 人脸识别数据库的 training-synthetic 子库,该子库包括 3D 形态模型合成的标准人脸灰度图像 3 240 幅,共 10 人,每人 324 幅,分辨率为 200×200,实验中将图像双三次插值为16×16。该子库中所有人脸图像仅包含无遮挡的椭圆形颈面部区域,以姿态和光照的变化为主。(1)姿态变化:水平左向旋转 0°~32°,以 4°为增量。(2)光照变化:以头部为中心,水平右向旋转 15°~90°,以 15°为增量;竖直仰角 0°~75°,以 15°为增量。

本组实验按光照变化将数据集分为两个子集:$set1:\{0°,30°,60°\}$,$set2:\{15°,45°,75°\}$。每个子集包含每个人 162 幅图像。实验中从 $set1$ 中选择 1 个人的人脸图像作为训练类样本,其余人脸作为非目标类测试样本,$set2$ 中选取对应的人脸图像作为目标类测试样本。实验结果见表 3:

表 3　MIT-CBCL 人脸识别实验结果

Tab. 3　Experimental results of different methods trained on the MIT-CBCL face database

目标类	训练集	测试集	1NN	kNN	SVDD	MSTCD	本文方法
第 1 组人脸	162	162/1458	0.9777	0.996	0.5	0.9993	0.9994
第 2 组人脸	162	162/1458	0.998	0.9986	0.5	1	1
第 3 组人脸	162	162/1458	1	1	0.5	1	1
第 4 组人脸	162	162/1458	0.9835	0.9942	0.5	0.9997	0.9997
第 5 组人脸	162	162/1458	0.9969	0.9998	0.5	1	1

实验结果表明,本文方法对人脸识别也表现出优越的性能,与 MSTCD 同时表现出最好的性能,而 SVDD 对于人脸数据难以进行有效的描述。实验中经样本修剪后的训练集样本数为 60 左右,可见本文提出的方法是有效可行的,其能用较低的模型复杂度实现对人脸数据的较好覆盖,这在实际应用中有一定的价值。

5　结论

依据特征空间中同类样本分布的连续性规律,本文提出基于高维空间典型样本 Steiner 最小树覆盖模型的一类分类算法。该方法首先对训练集进行样本修剪,选择最具代表性的样本作为训练集,然后对保留的典型样本构建 Steiner 最小树覆盖模型。相比 MSTCD,文中提出的方法能在较低覆盖模型复杂度的前提下更合理的描述目标类样本空间分布,构建更合理的覆盖模型。最后的实验结果表明该方法有效可行,可以实现对高维空间目标类数据的有效覆盖,具有一定的应用价值。

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]. Tien Tzu Hsueh Pao/ Acta Electronica Sinica,2009,37(11): 2496-2503. (in chinese)

[2] Koppel M, Schler J. Authorship verification as a one-class classification problem [A]. International Conference on Machine Learning [C], 2004: 489-495.

[3] 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.

[4] Kassab R, Alexandre F. Incremental data-driven learning of a novelty detection model for one-class classification with application to high-dimensional noisy data [J]. 2009,(74)2: 191-234.

[5] 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.

[6] Fung G. P C, Yu J X, Lu H J, et al. Text classification without negative examples revisit [J]. IEEE Transactions on Knowledge and Data Engineering, 2006,18(1): 6-20.

[7] Hempstalk K, Frank E, Witten I H. One-class classification by combining density and class probability estimation [A]. Machine Learning and Knowledge Discovery in Databases-European Conference [C], 2008: 505-519.

[8] 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)

[9] Tax D, Duin R. Support vector data description [J]. Machine Learning, 2004, 54(1):45-56.

[10] 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.

[11] Guo S M, Chen L C, Tsai J. A boundary method for outlier detection based on support vector domain description [J]. Pattern Recognition, 2009,42(1):77-83.

[12] Hao Pei-Yi. Fuzzy one-class support vector machines [J]. Fuzzy Sets and Systems, 2008,159(18): 2317-2336.

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

[14] Juszczak P, Tax D, Pekalska E, et al. Minimum span-

ning tree based one-class classifier [J]. Neurocomputing, 2009, 72: 1859-1869

[15] 王守觉.仿生模式识别(拓扑模式识别)——一种模式识别新模型的理论与应用[J].电子学报,2002,30(10): 1417-1420.
Wang Shou Jue. Bionic (topological) pattern recognition - A new model of pattern recognition theory and its applications [J], Acta Electronica Sinica, 2002, 30(10):1417-1420. (in chinese)

[16] Fampa M, Anstreicher K M. An improved algorithm for computing Steiner minimal trees in Euclidean *d*-space [J]. Discrete Optimization, 2008,5(2): 530-540.

[17] Chlebik M, Chlebikova J. The Steiner tree problem on graphs: Inapproximability results [J]. Theoretical Computer Science, 2008,406(3): 207-214.

[18] Muller-Hannemann M, Tazari S. A near linear time approximation scheme for Steiner tree among obstacles in the plane [J]. Computational Geometry: Theory and Applications, 2010,43(4): 395-409.

[19] Dreyer D R, Overton M L. Two heuristics for the Euclidean Steiner tree problem [J]. Journal of Global Optimization, 1998,13(1):95-106.

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

作者简介

胡正平，男(汉族)，生于四川仪陇县，在站博士后，副教授，硕士生导师，燕山大学通信电子工程系副主任，1996年于燕山大学无线电专业获得学士学位，并获得推荐研究生资格，1999获得电路与系统硕士学位，2007年于哈尔滨工业大学获得信号信息处理专业博士学位，目前为中国电子学会高级会员，中国图像图形学会高级会员，目前研究方向为统计学习理论与模式识别。E-mail:hzp@ysu.edu.cn

路 亮(1987-)，男(汉族)，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：统计学习理论与一类分类器。