高维空间多分辨率最小生成树模型的自适应一类分类算法

胡正平 [1] 冯 凯 [1]

摘 要 基于最小生成树数据描述 (Minimum spanning tree class descriptor, MSTCD) 法覆盖半径固定不变, 难以形成局部结构紧致性描述. 本文将多尺度分析思想和最小生成树 (Minimum spanning tree, MST) 结构相结合, 提出最小生成树的自适应多分辨率覆盖模型. 该模型利用样本流形的自身结构特点实现对数据的多分辨分析, 任意位置的分辨率由对应的 "点" 结构和 "边" 结构共同决定, 整体覆盖模型拥有多个覆盖半径, 数据当前位置不同、分辨率不同, 实现在多分辨尺度下对数据流形的自适应紧致覆盖. 实验结果表明该方法具有一定的合理性.

关键词 一类分类, 最小生成树, 多分辨分析, 自适应覆盖

引用格式 胡正平, 冯凯. 高维空间多分辨率最小生成树模型的自适应一类分类算法. 自动化学报, 2012, **38**(5): 769−775

DOI 10.3724/SP.J.1004.2012.00769

An Adaptive One-class Classification Algorithm Based on Multi-resolution Minimum Spanning Tree Model in High-dimensional Space

HU Zheng-Ping[1] FENG Kai[1]

Abstract The coverage radii in the algorithm of MSTCD (Minimum spanning tree class descriptor) are generally fixed without diversification, which makes it difficult to construct a close coverage for the local structure. This article combines multi-resolution thought and minimum spanning tree (MST) covering model, and proposes an adaptive multi-resolution covering model based MST in high-dimensional space. In this algorithm, multi-resolution of data is determined by the distributed feature of data manifold itself. The resolution of any location is depended on the structure of sample point and edge. The proposed novel algorithm permits that the whole covering model could have different coverage radii or resolutions and the resolution has relationship with location. Experiments show that the algorithm is reasonable.

Key words One-class classification, minimum spanning tree (MST), multi-resolution analysis, adaptive coverage

Citation Hu Zheng-Ping, Feng Kai. An adaptive one-class classification algorithm based on multi-resolution minimum spanning tree model in high-dimensional space. *Acta Automatica Sinica*, 2012, **38**(5): 769−775

模式识别分类算法往往需要多类训练样本来训练出多类分类器. 然而, 有些实际生产应用问题如基于计算机的遥感事物检测[1]、疾病图像识别[2]、生物身份特征检验[3] 等却有较为特殊的分类识别要求: 仅识别出定义的目标类样本即可, 其他非目标类样本一律可做拒绝识别处理, 这就是可拒绝一类分类问题[4]. 而且在某些实践中, 获取合理的非目标类样本往往比较困难. 一方面, 非目标类样本搜集困难. 有些情况下样本数据是随机出现的, 具有一定程度的偶然性, 而且这样的样本可信度很小. 另一方面, 产生非目标类样本本身就很困难. 有些领域如军事应用产生非目标类样本的代价高昂, 仅产生目标类样本就要花费大量的人力物力. 因此仅依靠目标类样本设计可拒绝一类分类器有重要意义.

针对可拒绝一类分类问题, 国内外学者已经开展了许多有价值的研究工作. 重要进展主要有三方面: 基于样本概率密度估计的研究、基于数据聚类的研究和基于样本边界覆盖的研究. 基于样本概率密度估计的方法如高斯混合模型 (Gaussian mixture model)[5] 和 Parzen 窗函数法[6]. 核心是先估计出目标类样本的概率密度函数, 然后再依据该函数和相应的密度阈值来进行判决. 这种方法的性能严重依赖于先验知识估计的合理性和真实性, 而且需要大量的样本数据作为实验基础以获得非常准确的概率密度函数, 这在实际中有时并不满足, 因此其仅在理论上可以获得良好效果. 基于数据聚类的方法核心是以聚类代替分类. 先对目标类样本做聚类处理, 得到相应的聚类中心和阈值半径然后进行判决. 代表方法有 K-means 和 1-nn 等. 该方法对实验参数

收稿日期 2011-05-30 录用日期 2011-08-30
Manuscript received May 30, 2011; accepted August 30, 2011
国家自然科学基金 (61071199), 中国博士后自然科学基金 (20080440124), 第二批中国博士后基金 (200902356), 河北省自然科学基金 (F2008000891, F2010001297) 资助
Supported by National Natural Science Foundation of China (61071199), Postdoctoral Science Foundation of China (20080440124), the Second Batch of China Postdoctoral Science Foundation (200902356), and National Natural Science Foundation of Hebei Province (F2008000891, F2010001297)
本文责任编委 刘一军
Recommended by Associate Editor LIU Yi-Jun
1. 燕山大学信息科学与工程学院 秦皇岛 066000
1. Institute of Information Science and Engineering, Yanshan University, Qinhuangdao 066000

极其敏感, 聚类中心和半径的合理选择及自适应性改进目前仍是个难题. 不过其模型复杂度较低, 操作简单, 具有较好的推广能力. 基于样本边界覆盖的方法核心是要先找到包含所有目标类样本的流行体的边界, 只要在边界内的样本就可以认为是目标类. 主要方法有一类支持向量机 (One-class support vector machines, OCSVM)[7−9]、支持向量域数据描述 (Support vector data description, SVDD)[10−12] 等.

各方法的区别在于所创建的流行边界不同. OCSVM 是在原点和目标类样本之间找到一个超平面, 对于测试样本与目标类样本同侧的就识别, 异侧的就拒绝. 该方法可以看成是支持向量机 (Support vector machine, SVM)[13] 的一个变种, 只是将原点看成异类而已. 同样模型训练的复杂度较高, 且不适合大量样本的情况. SVDD 是找到一个包含目标类样本的最小体积的超球, 以待测样本所处超球的内外侧位置作为判决依据. 该方法同样需要像 SVM 一样的优化训练, 复杂度较高, 不过对数据的分布描述较 OCSVM 更加合理. 缺点是如果数据分布不均匀, 覆盖模型会存在冗余区域, 不能产生紧致的覆盖. 对于高维空间少量样本情况, 问题更加严重, 并由此造成很高的误判率. 以上方法均是以规则几何模型对样本数据进行的覆盖估计, 对于具有复杂几何分布的数据集, 估计的覆盖模型难以符合数据的先验分布规律, 较难对数据进行紧致性覆盖, 分类器的性能将急剧下降.

针对呈复杂几何分布的数据样本集, 一些研究者提出基于最小生成树数据描述 (Minimum spanning tree class descriptor, MSTCD)[14] 的分类算法. 该方法以数据的最小生成树 (Minimum spanning tree, MST)[14] 作为样本流形的拟合估计主体, 更适合数据无规则分布的一般情况, 在实际应用中效果更好. 但是 MSTCD 算法对 MST 全树使用了相同的补充覆盖描述 (覆盖半径), 在目标样本数据密集区和稀疏区采用同样的分辨精度, 这种忽略数据本征流形结构、无差别的做法虽然操作简单, 但容易形成误识. 因此, 基于 MSTCD 的自适应覆盖思想, 自适应模型覆盖方法[15] 在一定程度上实现了模型的局部区分性分析. 但其分辨能力只能做到 “树枝” 级别, 仅对样本点的空间位置分布做到多分辨率分析, 没有精细尺度的 “点” 分析. 本文提出的基于高维空间最小生成树自适应覆盖模型的多分辨率可拒绝一类分类算法, 在 MSTCD 模型基础上引入了多分辨率的思想, 在数据流形的不同区域使用不同的分辨率 (覆盖范围), 根据流形结构的分布特点对数据进行有差别的分辨分析. 数据密集区覆盖范围较大, 数据稀疏区覆盖范围较小, 减少覆盖模型的冗余区域, 以使估计的覆盖模型与数据的真实流形分布结构更加贴合.

1 MSTCD 覆盖模型简述

针对一类可拒绝分类问题, 受到仿生模式识别[16] 思想的启发, 研究同类 (目标类) 样本的关系是根本出发点. 目前学界普遍认为同类样本的流形分布具有连续性的特点: 同类样本分布在同一个光滑的流形上而不存在孤立的样本点, 且越是相似的样本, 分布位置越接近 (连续性的接近). 给定一个有 n 个目标类样本的训练集 $X = \{\boldsymbol{x}_i \in \mathbf{R}\}_{i=1}^{n}$, 当目标类中任意两个样本点 $\{\boldsymbol{x}_i, \boldsymbol{x}_j\} \in X$ 在本征流形上接近到一定程度后, 用一条直线将二者相连, 可以认为该直线上的点及其适度邻域内的点都属于目标类. 本文称该直线上的点为虚拟目标类样本点. 当然, 也可以用曲线来连接 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$, 但是基于最近距离原则, 直线连接的可信度更高. 同理, 直线越短可信度越高.

推而广之, 用目标训练集中最短的 $n-1$ 个线性连接 (直线段) 的组合来构成对整体数据实际模型的连续性拟合估计, 进而把它看成数据本征流形的简化描述模型. 这些线性连接就是 MSTCD 覆盖模型的骨干.

对任意线段 $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ 上的虚拟目标类样本点 $\boldsymbol{x}_\lambda$:

$$\boldsymbol{x}_\lambda = \boldsymbol{x}_i + \lambda_{ij}(\boldsymbol{x}_j - \boldsymbol{x}_i) = (1-\lambda_{ij})\boldsymbol{x}_i + \lambda_{ij}\boldsymbol{x}_j \quad (1)$$

其中, $0 \le \lambda_{ij} \le 1$. 寻找最短的 $n-1$ 个线性连接的工作可采用图论中的寻找最小生成树算法来完成. 首先, 构造目标训练集 X 上的全连接无向图 $G = \{V, E\}$, V 是图的顶点集, E 是图的边集, 即 X 中任意两点连线的集合. 以 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 为端点的边记为 e_{ij}, 每条边设定欧氏测度权值 $w_{ij} = \| \boldsymbol{x}_j - \boldsymbol{x}_i \|$, 然后, 寻找图 G 的最小生成树 MST: 所有的顶点连接、没有环路, 并且总的权重最小的子图. 这样就找到了目标训练集 X 的基本骨干覆盖模型. 最后, 需要找到上述 MST 模型的补充覆盖区域, 即所有样本点 (包括虚拟样本点) 的适当邻域组成的区域, 只要在补充覆盖范围内的数据点可做目标类处理. 这里任意邻域看成以对应样本点为球心的超球, 超球半径亦即 MST 的覆盖半径设为阈值 θ.

任意测试点 $\boldsymbol{x}$ 在 e_{ij} 上的投影点为

$$\boldsymbol{x}_p = \boldsymbol{x}_i + \frac{(\boldsymbol{x}_j - \boldsymbol{x}_i)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}_i)(\boldsymbol{x}_j - \boldsymbol{x}_i)}{\| \boldsymbol{x}_j - \boldsymbol{x}_i \|^2} = \boldsymbol{x}_i + \lambda_{ij}(\boldsymbol{x}_j - \boldsymbol{x}_i) \quad (2)$$

当 $\boldsymbol{x}_p$ 在边 e_{ij} 上时, 即 $0 \le \lambda_{ij} \le 1$ 时, $\boldsymbol{x}$ 到边 e_{ij} 的最小距离为

$$d(\boldsymbol{x}|e_{ij}) = \| \boldsymbol{x} - \boldsymbol{x}_p \| \quad (3)$$

当 $\boldsymbol{x}_p$ 不在边 e_{ij} 上时, 即 $\lambda_{ij} < 0$ 或 $\lambda_{ij} > 1$ 时, 最小距离为

$$d(\boldsymbol{x}|e_{ij}) = \min\{\| \boldsymbol{x} - \boldsymbol{x}_i \|, \| \boldsymbol{x} - \boldsymbol{x}_j \|\} \quad (4)$$

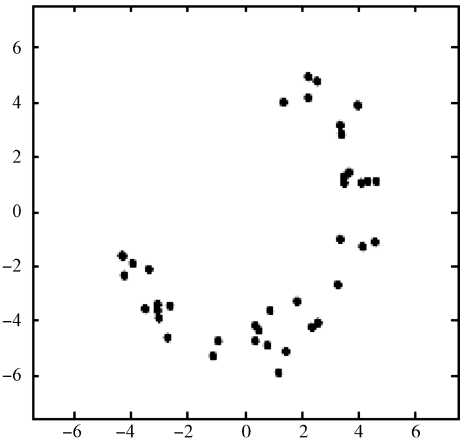
确定了测试点 $\boldsymbol{x}$ 到所有边的最小距离后, 定义 $\boldsymbol{x}$ 到目标类 X 的 MST 最小距离:

$$d_{\mathrm{MST}}(\boldsymbol{x}|X) = \min d(\boldsymbol{x}|e_{ij}) \quad (5)$$

根据上文 θ 的定义, 当 $d_{\mathrm{MST}}(\boldsymbol{x}|X) \le \theta$ 时, $\boldsymbol{x}$ 被判为目标类, 否则判为非目标类. 这就是 MSTCD 的识别判决基本原理.

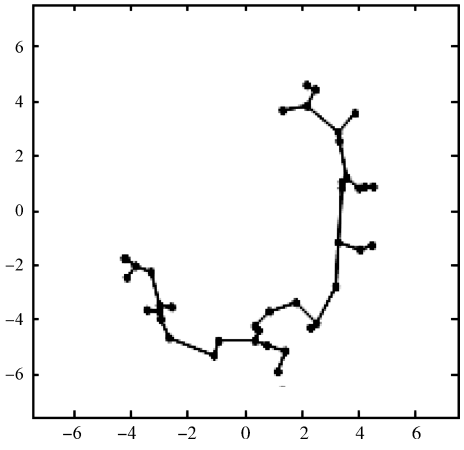
下面以一组分布呈 Banana 形状的数据为例说明 MSTCD 流形结构. 图 1 (a) 显示原始数据点分布情况. 对这样无规则分布的数据来说, 采用规则几何形状的模型 (如 SVDD) 或粗略的边界划分模式 (如 OCSVM) 描述, 得到的拟合模型必然会存在大量的冗余区域. 图 1 (b) 显示对应生成的 MST 基本模型. 这种树模型对一般流形结构拟合很有效. 图 1 (c) 显示完整的 MST 覆盖模型.

以上的 MSTCD 能够较为真实地拟合一般数据的内在流形结构, 无论其分布是否均匀, 外观是否规则, 都可以对数据形成较好的紧致估计覆盖, 这是 MSTCD 独具的优点. 然而, MSTCD 在各流形区域中都采用相同的覆盖半径, 在数据的密集区域与稀疏区域并没有区别对待. 应用固定的分辨率 (覆盖半径) 描述可信度不同的区域必然导致误识和覆盖冗余. MSTCD 能够精确地描绘出任意数据流形的自适应结构特点, 但是固定的分辨半径使覆盖模型变得粗糙化、模糊化, MSTCD 自身的优点并没有充分表现出来.



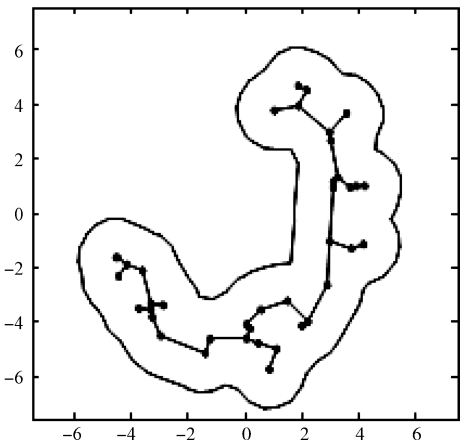
(a) 原始数据

(a) Original data



(b) MST 基本模型

(b) MST basic model



(c) MSTCD 完整覆盖模型

(c) Full coverage model of MSTCD

图 1　MSTCD 覆盖模型示意图

Fig. 1　MSTCD coverage model diagram

2　自适应多分辨率 MSTCD 覆盖模型

针对上文提到的 MSTCD 覆盖模型的不足, 本文将多分辨率[15] 的思想引入 MSTCD 覆盖模型中, 采用自适应变化的覆盖分辨率区别对待数据流形结构中不同数据密度区域. 在数据分布密集区令覆盖半径变大、分析尺度变粗; 在数据分布稀疏区令覆盖半径变小、分析尺度变细. 针对生成的 MST 覆盖模型, 既考虑全局分布: 较长边的覆盖半径普遍小于较短边的覆盖半径, 即长边的分辨率小于短边的分辨率; 又考虑局部分布: 同一条边上的不同样本点的分辨率也不同, 其覆盖半径与样本点的自身位置处点分布密度成正比例. 在全树范围内采用不同的分辨率, 对任意样本点同时考虑其所在边的长度比例因素和该点周围样本点的分布密度因素, 综合权衡后得到与其局部结构相适应的分辨半径. 这样, 覆盖模型中每个点都有自适应的独立的覆盖半径, 整体的

分辨能力就变得精细化、多样化.

具体而言, MST 中任意点 $\boldsymbol{x}_i$ 的覆盖半径 (分辨率) 定义为

$$R(\boldsymbol{x}_i) = r_0(\lambda_e \times e\_\text{density}(\boldsymbol{x}_i) + \lambda_n \times n\_\text{density}(\boldsymbol{x}_i)) \tag{6}$$

式中, r_0 表示基准半径. 一般情况下和 MSTCD 覆盖模型选择的半径相同: 将产生的 MST 的各个边长按长度排序, 去掉前 10 % 的最长边并把剩下的最长边的长度设定为 r_0. 另外, 式中 λ_e 和 λ_n 为人工调节参数. 随样本集的不同分布特点而变化. 二者调节的是参数 $e\_\text{density}(\boldsymbol{x}_i)$ 和 $n\_\text{density}(\boldsymbol{x}_i)$ 发挥影响力的大小. 根据经验, 一般情况下要求 $0 \le \lambda_e + \lambda_n \le 3$. 参数 $e\_\text{density}(\boldsymbol{x}_i)$ 反映的是 $\boldsymbol{x}_i$ 所在边的长度比例因素. 该值仅与边长有关, 且 $0 < e\_\text{density}(\boldsymbol{x}_i) \le 1$. 参数 $n\_\text{density}(\boldsymbol{x}_i)$ 反映的是 $\boldsymbol{x}_i$ 周围样本点的分布密度关系. 同理, $0 < e\_\text{density}(\boldsymbol{x}_i) \le 1$. 自适应多分辨率 MSTCD 覆盖模型的具体结构如图 2 所示.

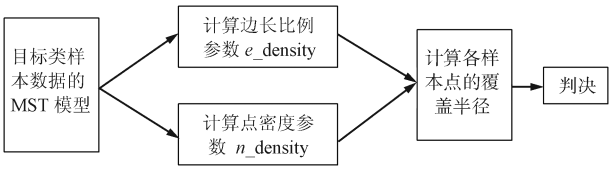


图 2 自适应多分辨率 MSTCD 覆盖模型结构

Fig. 2 Structure of adaptive multi-resolution MSTCD coverage model

2.1 数据样本点密度参数计算

MST 某一点 $\boldsymbol{x}_i$ 处的样本点密度 $n\_\text{density}(\boldsymbol{x}_i)$ 与 $\boldsymbol{x}_i$ 的近邻样本点的分布紧密状况有密切的关系, 如果 $\boldsymbol{x}_i$ 与近邻点的距离较小, 当然获得点密度就较大. 依据这个思想, 可以先自适应地选择出近邻点, 然后, 根据近邻距离算出所需参数 $n\_\text{density}(\boldsymbol{x}_i)$.

这里采用一种自适应的近邻选择方法: 对任意点 $\boldsymbol{x}_i$, 先计算它到所有其余 $n-1$ 个点的距离之和 $\text{total}(\boldsymbol{x}_i)$, 然后求出 $\text{total}(\boldsymbol{x}_i)$ 的平均数 (平均距离), 与 $\boldsymbol{x}_i$ 的距离小于等于平均数的点就认为是近邻. 近邻是个相对的、局部性的概念. 平均距离反映的是仅与点 $\boldsymbol{x}_i$ 有关的全部距离参量的总趋势, 能够体现出相对性的含义. 另外, 这里把平均距离作为默认的局部范围阈值, 每个点的平均距离不同, 默认的局部范围也就不同, 这是符合数据分布规律的. 而且平均距离的计算不会引入人为因素, 完全自适应决定. 常用的 k 近邻图虽然能反映相对性 (比较对象确实是仅与点 $\boldsymbol{x}_i$ 有关的全部距离参量), 但把各点的局部性却按统一标准来处理 (每个点的近邻都是 k 个). 由此看来, 这种方法较好.

找到近邻后算出近邻距离之和 $d\_\text{neighbor}(\boldsymbol{x}_i)$. 对所有点完成近邻操作后, 会得到集合 $\text{total} = \{\text{total}(\boldsymbol{x}_i) | i = 1, \cdots, n\}$, 找到其中最大的两个数记为: total_max1 和 total_max2. 则 $n\_\text{density}(\boldsymbol{x}_i)$ 计算公式如下:

$$n\_\text{density}(\boldsymbol{x}_i) = \exp\left(\frac{-d\_\text{neighbor}(\boldsymbol{x}_i)}{\text{total\_max1} + \text{total\_max2} - \text{total}(\boldsymbol{x}_i)}\right) \tag{7}$$

该式的意义在于当近邻距离 $d\_\text{neighbor}(\boldsymbol{x}_i)$ 越小, 且总距离 $\text{total}(\boldsymbol{x}_i)$ 越小时, 点 $\boldsymbol{x}_i$ 的点密度 $n\_\text{density}(\boldsymbol{x}_i)$ 越大. 对于虚拟目标类样本点 $\boldsymbol{x}_\lambda$, 其点密度虽不能直接计算. 但由式 (1) 可知, 对每一个 $\boldsymbol{x}_\lambda$, 仅存在唯一的 λ_{ij} 与之对应, 且所在线段的两端点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的点密度可求. 则对虚拟样本点 $\boldsymbol{x}_\lambda$ 的点密度定义为

$$n\_\text{density}(\boldsymbol{x}_\lambda) = (1-\lambda_{ij}) \times n\_\text{density}(\boldsymbol{x}_i) + \lambda_{ij} \times n\_\text{density}(\boldsymbol{x}_j) \tag{8}$$

这样, MST 上的任意点 $\boldsymbol{x}_i$ 都存在一一对应的点密度 $n\_\text{density}(\boldsymbol{x}_i)$, 每个点拥有独立自适应的覆盖半径 (分辨率), 全树就拥有多差别化的多分辨率分析能力.

2.2 MST 边长比例参数计算

当然, 仅根据样本点的分布密度来构造多分辨率覆盖半径是不全面的, 对于 MST 各边长的比例关系也要考虑在内. 短边的可信度较高, 分辨率要高一些, 长边的可信度较低, 分辨率要低一些. 判断边长长短与否要根据 MST 的自身结构特点来自适应决定.

对于 MST 中所有的边, 按照边的长度进行升序排列. 去掉前 10 % 的最长边, 对剩下的所有边取均值记为 e_mean. 本文把 e_mean 作为判断标准, 长度小于 e_mean 的做短边处理, 长度大于 e_mean 的做长边处理. 令 $\boldsymbol{x}_i$ 为 MST 上的任意一点, 所在边的长度记为 $e\_\text{length}(\boldsymbol{x}_i)$. 若 $e\_\text{length}(\boldsymbol{x}_i)$ 是短边, 那么该点所在边长度比例参数 $e\_\text{length}(\boldsymbol{x}_i) = 1$, 若 $e\_\text{length}(\boldsymbol{x}_i)$ 是长边, 那么该点所在边长度比例参数记为

$$e\_\text{density}(\boldsymbol{x}_i) = \exp\left(\frac{e\_\text{mean} - e\_\text{length}(\boldsymbol{x}_i)}{e\_\text{length}(\boldsymbol{x}_i)}\right) \tag{9}$$

式 (9) 的意义在于当 $e\_\text{length}(\boldsymbol{x}_i)$ 大于 e_mean 时, 超出的程度越少, 所获得的 $e\_\text{density}(\boldsymbol{x}_i)$ 越大, 所得分辨率 (覆盖半径) 越大.

2.3 计算各样本点的分辨率 (覆盖半径)

经过上述两个步骤, 必要的参数已经获得. 利用式 (6) 可以得到多分辨率覆盖半径 $R(\boldsymbol{x}_i)$, 并且可以看出 $R(\boldsymbol{x}_i)$ 随点 $\boldsymbol{x}_i$ 位置不同而变化, 充分体现了多分辨率分析的思想. 随着点 $\boldsymbol{x}_i$ 对 MST 的遍历, 使原来的 MST 模型实现动态覆盖效果: 在数据密集区覆盖范围大, 在数据稀疏区覆盖范围小. 参数 λ_e 和 λ_n 的作用在于调节参数 $n\_\text{density}(\boldsymbol{x}_i)$ 和 $e\_\text{density}(\boldsymbol{x}_i)$ 的比重大小. 不同的数据集有自己的结构特点, 所以在具体测试应用中 λ_e 和 λ_n 也不可能相同. 依据经验来说, 二者尽量不要相等, 且不要出现负数, 精确到小数点后第三位时会有较高精度.

在全树范围内, 测试点 $\boldsymbol{x}$ 必然存在一个到 MST 的最小距离投影点 $\boldsymbol{x}_p$, 结合式 (1) 的表述形式, 当 $0 < \lambda_{ij} < 1$ 时, 最小距离投影点 $\boldsymbol{x}_p$ 就是前文的虚拟样本点 $\boldsymbol{x}_\lambda$. 此时的分辨率半径为

$$R(\boldsymbol{x}_p) = R(\boldsymbol{x}_\lambda) = r_0(\lambda_e \times e\_\text{density}(\boldsymbol{x}_\lambda) + \lambda_n \times n\_\text{density}(\boldsymbol{x}_\lambda)) \tag{10}$$

测试点 $\boldsymbol{x}$ 到 MST 的最小距离为

$$d_{\text{MST}}(\boldsymbol{x}|X) = \| \boldsymbol{x} - \boldsymbol{x}_p \| = \| \boldsymbol{x} - \boldsymbol{x}_\lambda \| \tag{11}$$

当 $\lambda_{ij} \le 0$ 时, 最小距离投影点 $\boldsymbol{x}_p$ 就是 $\boldsymbol{x}_i$. 此时的分辨率半径为

$$R(\boldsymbol{x}_p) = R(\boldsymbol{x}_\lambda) = r_0(\lambda_e \times e\_\text{density}(\boldsymbol{x}_i) + \lambda_n \times n\_\text{density}(\boldsymbol{x}_i)) \tag{12}$$

测试点 $\boldsymbol{x}$ 到 MST 的最小距离为

$$d_{\text{MST}}(\boldsymbol{x}|X) = \| \boldsymbol{x} - \boldsymbol{x}_p \| = \| \boldsymbol{x} - \boldsymbol{x}_i \| \tag{13}$$

当 $\lambda_{ij} \ge 1$ 时, 最小距离投影点 $\boldsymbol{x}_p$ 就是 $\boldsymbol{x}_j$. 此时的分辨率半径为

$$R(\boldsymbol{x}_p) = R(\boldsymbol{x}_j) = r_0(\lambda_e \times e\_\text{density}(\boldsymbol{x}_j) + \lambda_n n\_\text{density}(\boldsymbol{x}_j)) \tag{14}$$

测试点 $\boldsymbol{x}$ 到 MST 的最小距离为

$$d_{\text{MST}}(\boldsymbol{x}|X) = \| \boldsymbol{x} - \boldsymbol{x}_p \| = \| \boldsymbol{x} - \boldsymbol{x}_j \| \tag{15}$$

这样, 根据 $R(\boldsymbol{x}_p)$ 与 $d_{\text{MST}}(\boldsymbol{x}|X)$ 的大小即可判断识别.

3 实验仿真

为验证本文思路的正确性, 进行了三组实验, 实验采用的样本数据分别来源于 UCI 数据库、MNIST 手写体数字库、MIT-CBCL 人脸识别数据库. 采用的对比实验为经典的 SVDD 方法, 采用高斯核函数, 核宽度为 5, 错误容忍率为 0.1; 经典的 MSTCD 方法, 其中覆盖半径与本文的基准半径 r_0 取法相同, 皆是去掉前 10 % 的最长边后剩下的最长边的长度; 自适应模型覆盖方法.

实验结果的评价指标为目标类正确识别率和目标类正确拒识率. 目标类正确识别率表示测试样本中目标类正确识别数与测试样本中目标类识别数的商. 目标类正确拒识率表示测试样本中非目标类正确识别数与测试样本中非目标类识别数的商. 以上两个指标互相影响、牵制, 不可能存在单纯地提高一个目标函数而不影响另一指标大小的方法. 本文的实验结果为均衡考虑二者后的结果, 实际上存在单一目标函数的较好结果, 但另一结果的表现会下降.

3.1 UCI 数据库

该组实验的参考数据集为 UCI 数据库的 Iris 数据集、LandSat 数据集和 Letter 数据集. 每个数据集随机进行 20 次试验, 取平均值作为最终实验结果. 对 Iris 数据集, 样本维数为 4, 训练集: 1、2 类为目标类, 每类随机提取 30 个样本; 测试集: 3 类为非目标类, 提取 50 个样本并加入剩余目标类样本. 对 Letter 数据集, 样本维数为 16, 训练集: A、B 类为目标类, 每类随机提取 200 个样本; 测试集: C 类为非目标类, 提取 400 个样本并加入剩余目标类 600 个样本. 对 LandSat 数据集, 样本维数为 36, 训练集: 1、3 类为目标类, 每类随机提取 200 个样本; 测试集: 2、4、5 类为非目标类, 提取 672 个样本并加入剩余目标类 858 个样本. 实验情况如表 1 所示.

根据结果可以看出, 在数据维数较低情况下, 本文方法效果十分明显. 随着训练样本数的增加, 正识率和拒识率相应提高, 而随着数据维数的增加, 正识率和拒识率相应减小. 另一方面, 针对表中个别数据指标, 本文方法并不是最优, 这与数据自身的潜在流形分布特点有关. 综合两项指标可见本文的多分辨率做法是比较合理的.

3.2 MNIST 手写体数字识别实验

MNIST 手写体数字数据库是由 0 ∼ 9 共 10 类数字的手写体样本组成的, 本身已经分为训练集和测试集两部分. 每一个样本分辨率归一化为 28 像素 × 28 像素. 实验选择不同的数字组合来验证本文方法的有效性: 第 1 组 5、6 为目标类, 9 为非目标类; 第 2 组 0、9 为目标类, 6 为非目标类; 第 3 组 1、2 为目标类, 4、7 为非目标类. 每个目标类训练集随机抽取 500 个样本, 测试集每类样本 200 个. 实验情况如表 2 所示.

MST 模型对数据主体流形拟合得较好, 在此基础上, 引入的多分辨率方法在细节上形成了符合数据分布规律的较好覆盖. 相对于同类的自适应覆盖模型法, 可以做到单一评价指标提高, 二者在总体性

能上几乎相当, 这与所使用的数据集的流行分布有密切关系, 相对而言, MNIST 手写体数字数据库的数据分布较为均匀规范, 所选数据分布密度没有明显变化, 因此本文方法的优越性并未完全显现.

3.3 MIT-CBCL 人脸识别实验

实验数据取自 MIT-CBCL 人脸识别数据库中的 Training-synthetic 子库. 该库数据是标准人脸的 3D 模型合成的灰度图像, 包含 10 个人, 每个人有 324 幅图像, 像素的分辨率为 200 像素 × 200 像素. 每幅图像记录了一个人的姿态和光照的不同角度变化数据. 本次实验以姿态为标准把数据分为训练集 $\{0°, -8°, -16°, -24°\}$ 和测试集 $\{-4°, -12°, -20°, -28°\}$. 每个子集有 144 幅人脸, 在训练集合中随机选取 5 个人的图像做目标类训练样本, 而所有测试集的样本均作测试使用, 实验结果如表 3 所示.

MST 模型适合对人脸模型拟合, 而且使用分辨率不断调整策略对于分类性能影响较为深刻. 本文方法可以使正识率和拒识率同时得以提高, 一方面说明将多分辨率思想贯彻得越深, 实验效果越好, 另一方面说明人脸数据分布是疏密有间的, 本文数据描述方法是符合人脸数据分布规律的.

4 结论

在 MSTCD 算法的基础之上, 本文提出自适应多分辨率 MST 覆盖一类分类算法. 以 MST 为基本覆盖模型, 把 MST 各边上的点看作新增的虚拟样本点, 利用现有数据分布流行结构实现对测试样本的多分辨分析, 即在数据分布密集区使用较细尺度描述数据, 在数据分布稀疏区使用较粗尺度描述数据. 对任一样本点来说, 其所在流形区域的数据描述分辨率既与周围实际样本点的分布疏密情况有关, 又与其所在边长的比例长度因素 (虚拟样本点分布疏密情况) 有关. 不同的样本区域有不同的描述尺度, 各样本点 (包括虚拟样本点) 处的分辨率相互独立, 因此形成多尺度、多精度的数据描述形式. 原始算法中, 不同流形区域会使用相同的数据分辨尺度, 而且这种尺度固定不变, 不能反映数据本身的流形结构. 本文提出的数据描述模型改善了原始算法粗糙固定的数据分辨率, 形成灵活精确的多重数据分辨率, 并可以很好地体现数据样本自身的流形结构特点和规律, 这种数据描述模型减少了原始算法中冗余的覆盖区域, 一定程度上提高了拒识率.

表 1 UCI 数据库实验结果

Table 1 Experimental results of UCI database

实验方法	Irish (正识率 : 拒识率) (%)	Letter (正识率 : 拒识率) (%)	Landsat (正识率 : 拒识率) (%)
SVDD	90.00 : 78.00	82.00 : 72.50	70.63 : 77.68
MSTCD	92.50 : 80.00	98.83 : 97.25	81.24 : 87.94
自适应覆盖模型法	92.50 : 82.00	98.17 : 99.25	85.08 : 89.43
本文方法	94.06 : 85.38	98.31 : 97.47	85.26 : 90.89

表 2 MNIST 数据库实验结果

Table 2 Experimental results of MNIST database

实验方法	第 1 组 (正识率 : 拒识率) (%)	第 2 组 (正识率 : 拒识率) (%)	第 3 组 (正识率 : 拒识率) (%)
SVDD	75.75 : 81.50	92.50 : 95.00	68.50 : 92.25
MSTCD	82.25 : 83.50	94.50 : 95.50	81.50 : 91.25
自适应覆盖模型法	82.00 : 84.00	96.50 : 96.00	83.25 : 92.00
本文方法	83.35 : 84.59	96.69 : 95.85	83.43 : 91.15

表 3 MIT-CBCL 数据库实验结果

Table 3 Experimental results of MIT-CBCL database

实验指标	SVDD (%)	MSTCD (%)	自适应覆盖模型法 (%)	本文方法 (%)
正识率	91.11	92.78	92.64	94.08
拒识率	96.11	97.22	97.78	98.72

本文算法忠实体现并遵循数据自身流形结构的分布规律, 其性能严重依赖于现有数据样本点的分布流形状况. 这样, 训练样本的合理选取变得至关重要. 如果所选训练样本能较好地拟合出数据本征流形结构, 效果会较好, 相反效果就会变差. 如果在样本数据中混入噪声数据, 算法性能在一定程度上会有所下降. 在以后的工作中希望可以更好地解决对样本数据选取不敏感的算法, 加强数据描述模型的鲁棒性, 并且能够抵御一定程度的噪声干扰.

References

1 Li W K, Guo Q H, Elkan C. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 2011, **49**(2): 717−725

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

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

4 Xu J, Chen Q C, Wang X L, Wei Z Y. One-class classification models for financial industry information recommendation. In: Proceedings of the International Conference on Machine Learning and Cybernetics. Qingdao, China: IEEE, 2010. 3329−3334

5 Gosztolya G, Banhalmi A, Toth L. Using one-class classification techniques in the anti-phoneme problem. In: Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis. Povoa de Varzim, Portugal: Springer, 2009. 433−440

6 Oliveira H, Caeiro J J, Correia P L. Improved road crack detection based on one-class Parzen density estimation and entropy reduction. In: Proceedings of the 17th IEEE International Conference on Image Processing. Hong Kong, China: IEEE, 2010. 2201−2204

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

8 Tian Jiang, Gu Hong. Outlier one class support vector machines. *Journal of Electronics and Information Technology*, 2010, **32**(6): 1284−1288
(田江, 顾宏. 孤立点一类支持向量机算法研究. 电子与信息学报, 2010, **32**(6): 1284−1288)

9 Gu H, Zhao G Z, Qiu J. One-class support vector machine with relative comparisons. *Tsinghua Science and Technology*, 2010, **15**(2): 190−197

10 Tax D, Duin R. Support vector data description. *Machine Learning*, 2004, **54**(1): 45−56

11 Sakla W, Chan A, Ji J, Sakla A. An SVDD-based algorithm for target detection in hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 2011, **8**(2): 384−388

12 Huang G X, Chen H F, Yin F. Improved support vector data description. In: Proceedings of the International Conference on Machine Learning and Cybernetics. Qingdao, China: IEEE, 2010. 1459−1463

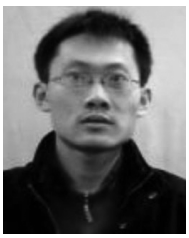
13 Zhang X L, Ren F. Improving svm learning accuracy with adaboost. In: Proceedings of the 4th International Conference on Natural Computation. Jinan, China: IEEE, 2010. 221−225

14 Juszczak P, Tax D, Pekalska E, Duin R. Minimum spanning tree based one-class classifier. *Neurocomputing*, 2009, **72**(7−9): 1859−1869

15 Hu Zheng-Ping, Xu Cheng-Qian, Jia Qian-Wen. A classification algorithm with reject option based on adaptive minimum spanning tree covering model in high-dimensional space. *Journal of Electronics and Information Technology*, 2010, **32**(12): 2896−2900
(胡正平, 许成谦, 贾千文. 基于高维空间最小生成树自适应覆盖模型的可拒绝分类算法. 电子与信息学报, 2010, **32**(12): 2896−2900)

16 Wang Shou-Jue. Bionic (topological) pattern recognition — a new model of pattern recognition theory and its applications. *Acta Electronica Sinica*, 2002, **30**(10): 1417−1420
(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. 电子学报, 2002, **30**(10): 1417−1420)

胡正平 燕山大学博士, 教授. 主要研究方向为统计学习理论, 模式识别, 医学图像处理. 本文通信作者.
E-mail: hzp@ysu.edu.cn
(**HU Zheng-Ping** Ph. D., professor at Yanshan University. His research interest covers statistical learning theory, pattern recognition, and medical image processing. Corresponding author of this paper.)



冯 凯 燕山大学信息科学与工程学院硕士研究生. 2009 年获得燕山大学信息科学与工程学院学士学位. 主要研究方向为一类分类算法.
E-mail: fklegend@gmail.com
(**FENG Kai** Master student at the Institute of Information Science and Engineering, Yanshan University. He received his bachelor degree from Yanshan University in 2009. His main research interest is one-class classification.)