文章编号: 1000-6788(2012)08-1784-09 中图分类号: TP181 文献标志码: A

LSVDD: 基于局部支持向量数据描述的稀有类分析算法

熊海涛 [1], 吴俊杰 [1], 刘 鲁 [1], 李 明 [2]

(1. 北京航空航天大学 经济管理学院, 北京 100191; 2. 中国石油大学 工商管理学院, 北京 102249)

摘 要 在单类支持向量数据描述算法的基础上, 提出了一种基于局部支持向量数据描述的稀有类分析算法: LSVDD, 能够处理存在类重叠的类不平衡问题. 该算法利用支持向量数据描述算法对各类样本分别进行单类学习, 从而获得单类模型; 然后对单类模型的概念重叠区域使用属性选择进一步进行局部单类学习, 最后得到综合分类模型. 在仿真数据集和 UCI 数据集上的实验结果表明, LSVDD 能够有效和稳定地提高稀有类分析精度.

关键词 数据挖掘; 稀有类分析; 支持向量数据描述; 属性选择

LSVDD: Rare class analysis based on local support vector data description

XIONG Hai-tao[1], WU Jun-jie[1], LIU Lu[1], LI Ming[2]

(1. School of Economics and Management, Beihang University, Beijing 100191, China;
2. School of Business Administration, China University of Petroleum, Beijing 102249, China)

Abstract As a hot topic in data mining society, rare class analysis (RCA) has been widely used in various application domains including financial fraud detection, network intrusion detection, facility failure diagnosis, etc. However, it is not until recently that researchers have realized the impact of complex data structures to the RCA problem. We propose a local support vector data description algorithm LSVDD for RCA based on SVDD, which has the ability to handle class imbalance problem with the presence of class overlaps. Specifically, LSVDD firstly uses SVDD to get one-class classification model for each class and finds the concept overlapping regions between different classes. Then, the regions are locally trained using SVDD again after attribute selections. Finally, the models for non-overlapping and overlapping regions are combined to form a complete RCA model. Experimental results on artificial and real-world UCI data sets demonstrate that LSVDD can improve the performances of RCA stably and effectively.

Keywords data mining; rare class analysis; support vector data description; attribute selection

1 引言

稀有事件永远是人们关注的焦点, 其特点在于正常情形下出现的概率很小, 但一旦发生将产生巨大的影响 [1]. 典型的稀有事件包括金融欺诈、网络入侵、通信设备故障等 [2−3]. 数据挖掘中的稀有类分析技术是进行稀有事件预测的强有力工具 [2]. 研究表明, 在稀有类分析中, 稀有类的样本数目与普通类相差悬殊, 稀有类的预测难度远大于普通类. 大量算法因此被提出以解决这个问题, 如重抽样、成本敏感学习、调整归纳偏置等 [4−5]. 然而, 这些方法的出发点通常是平衡各类样本, 没有考虑数据固有结构的影响, 忽略了稀有类分析与数据固有结构之间的紧密联系, 因此也无法真正解决问题.

单类学习由于能刻画数据空间分布, 近年来得到了机器学习领域的广泛重视. 基于统计学习理论的支持向量机 (support vector machines, SVMs)[6] 是一种比较成熟的机器学习方法, 它被很多学者应用于单类学习中. 文献 [7] 在 SVMs 理论基础上提出了单类支持向量机算法, 称为支持向量数据描述算法 (support vector

收稿日期: 2010-06-04
资助项目: 国家自然科学基金 (70901002, 90924020); 高等学校博士学科点专项科研基金 (200800060005, 20091102120014)
作者简介: 熊海涛 (1983–), 男, 博士研究生, 研究方向: 机器学习, 数据挖掘, 知识管理; 吴俊杰 (1979–), 男, 副教授, 研究方向: 机器学习, 数据挖掘, 商务智能; 刘鲁 (1947–), 女, 教授, 博士生导师, 研究方向: 管理信息系统, 电子商务, 知识管理; 李明 (1981–), 男, 博士, 研究方向: 管理信息系统, 知识管理.

data description, SVDD). 该算法通过计算包含训练样本的高维空间最小超球体的边界来对数据进行描述. 通过利用其他类信息, SVDD 也被扩展来解决多分类问题, 并与其他方法相结合提高分类效果 [8−9]. 文献 [8] 将 SVDD 算法应用于多分类问题, 给出了一种多类支持向量数据描述算法的组合方法, 我们称之为组合支持向量数据描述算法 (combined-SVDD, CSVDD). 该方法分别针对每类数据进行学习, 然后通过标准化组合方法对多个单类模型进行组合, 得到最终分类模型. 文献 [9] 则在 SVDD 中引入聚类算法, 先对数据进行聚类, 使得数据分为更小的部分, 然后再进行分类. 总体而言, 由于单类模型可以分别界定各类数据的边界, 从而在一定程度上揭示数据在分布空间的固有结构, 因此对于稀有类分析是有益的.

然而上述基于 SVDD 的分类算法只是简单地利用了各个类别的信息来进行学习, 没有很好地利用各个类别获得的单类模型, 不能考虑各类数据存在类重叠的情况. 而类重叠问题正是稀有类分析乃至机器学习领域的瓶颈问题之一 [10]. 文献 [11] 通过实验也发现, 稀有类分析效果不仅与类不平衡有关, 更与类重叠的程度有关. 事实上, 大量真实数据中存在着类重叠情况 [12]. 如 UCI 数据库中的 Ionosphere 数据集, 如果采用 SVDD 算法计算各类样本边界, 那么落入类重叠区域的样本达到 192 个, 占所有样本的 54.7%. 又如 UCI 的 Breast-w 数据集, 其最大和最小类的样本在 9 个属性中有 3 个完全重叠, 重叠情况也非常严重. 因此在稀有类分析中有必要对类重叠问题进行处理. 本文在 SVDD 的基础上, 提出一种基于局部支持向量数据描述的稀有类分析算法 (local support vector data description, LSVDD). 算法针对每个类别样本进行单类学习获得单类模型, 然后利用单类模型确定类重叠区域, 并针对类重叠区域进行局部单类学习, 优化综合分类模型, 从而获得对数据边界描述更加精确的模型, 提高稀有类分析精度.

2　数据固有结构对稀有类分析算法的影响

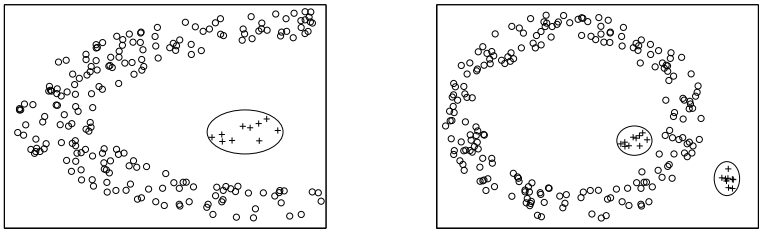
传统的分类算法, 如 C4.5、SVMs 等, 在简单数据上有较好的效果, 但在不均衡数据上却有着比较明显的缺陷. 传统观点认为这是各类样本的不平衡性对分类算法产生了影响 [11], 但这并不是导致稀有类分析如此困难的根本原因. 对于某些分类器如 SVMs, 在普通样本和稀有样本显著线性可分的情况下, 样本的不平衡性并不会对分类器学习带来实质性的影响, 尽管分类器的线性分界面可能会存在一定程度的偏移 [13]. 事实上, 稀有类分析中一个不容忽视的问题是数据固有结构的影响. 这里数据的 “固有结构” 指的是数据在属性空间的分布情况, 即数据的概念模型. 复杂的数据固有结构, 或者称为数据中的复杂概念 (complex concept), 已成为稀有类分析中隐含的且亟待解决的难点问题.

图 1 展示了三类比较典型的数据固有结构, 包括复杂类间结构、复杂类内结构和类重叠结构. (a) 中稀有类和普通类之间并非线性可分, 因此不管样本数量有无得到平衡, 采用简单的线性分类器建模是无效的; (b) 中稀有类样本被普通类样本分成了两部分, 这使得分类器的分界面学习变得更加困难, 但采用局部聚类仍能有效解决问题; (c) 中是最为困难的情况, 即两类之间出现了类重叠区域, 这导致稀有类和普通类之间根本无法正确划分, 而属性的缺失或者冗余属性的存在是造成上述类重叠的重要原因. 因此, 必须综合考虑样本的不平衡性以及数据结构的复杂性, 才能真正解决实践领域的稀有类分析问题.

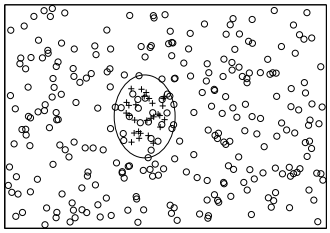
类重叠问题 (class overlapping problem) 可分为概念重叠 (concept overlapping) 和样本重叠 (sample overlapping) 两个层次 [14−16]. 概念重叠是指不同类别的区域 (一个区域刻画了一个类的概念) 在分布空间中重叠的情况, 属于宏观层面的概念. 样本重叠是指不同类别的样本点在分布空间中位置较为接近甚至重叠的情况, 属于微观层面的概念. 当两类数据较为接近时概念重叠就容易发生, 而样本重叠不一定会产生; 而当样本重叠产生则意味着概念重叠已经发生. 实践表明, 分类器划分错误往往集中在数据的边界区域, 而这正是概念重叠经常存在的区域. 因此, 对概念重叠区域进行单独学习能够有效避免其与概念非重叠区域的相互干扰, 有助于提高两类区域的分类精度. 值得一提的是, 除非出现严重的数据重复采集问题, 严格意义上的样本重叠是很少发生的. 真实数据通常表现为广泛存在的概念重叠, 且局部可能包含少量样本重叠. 有鉴于此, 本文致力于解决稀有类分析中包含少量样本重叠的概念重叠问题. 本文提出的 LSVDD 算法能够发现样本数据概念重叠区域, 并通过进一步针对概念重叠区域的局部学习, 最终提高稀有类分析的精度.

3　支持向量数据描述的原理及算法

SVMs 分类方法希望找到一条理想的分界面使得分类间隔达到最大, 而忽略了对数据本身的描述, 在处理复杂数据结构如类间重叠时往往存在问题. 近年来, 很多学者将 SVMs 应用于单类学习中, SVDD 算法是



(a) 复杂类间结构 (b) 复杂类内结构



(c) 类重叠结构

图 1 数据固有结构示例

其中一个代表. SVDD 针对单类进行学习, 寻找一个高维空间的超球体来覆盖尽可能多的数据在该属性空间的映像, 从而获得数据边界特征. 给定一个包含 n 个数据对象的集合 X=$\{x_i, i = 1, 2, \cdots, n\}$, SVDD 通过非线性映射函数 Φ 将输入空间映射到高维空间, 寻找一个半径为 R、球心为 a 的超球体来覆盖尽可能多的 x_i. SVDD 建立如下的优化问题 [7]:

$$\begin{aligned} &\min\ R^2 \\ &\text{s.t.}\ \ ||\Phi(x_i) - a||^2 \le R^2,\ \ i = 1, 2, \cdots, n \end{aligned} \tag{1}$$

在式 (1) 中引入松弛变量向量 $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$, 使得超球体能够将一部分样本作为噪音排除在外部, 优化问题变换为:

$$\begin{aligned} &\min_{R,\xi}\ q(R, \xi) = R^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.}\quad ||\Phi(x_i) - a||^2 \le R^2 + \xi_i, \qquad \xi_i \ge 0,\ i = 1, 2, \cdots, n \end{aligned} \tag{2}$$

其中, C 为噪音惩罚系数. 引入拉格朗日函数可得:

$$L(R, a, \xi, \alpha, \beta) = R^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(R^2 + \xi_i - ||\Phi(x_i) - a||^2) - \sum_{i=1}^{n}\beta_i\xi_i \tag{3}$$

令 $C = \frac{1}{n\nu}$, 式 (3) 可变换为:

$$L(R, a, \xi, \alpha, \beta) = R^2 + \frac{1}{n\nu}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(R^2 + \xi_i - ||\Phi(x_i) - a||^2) - \sum_{i=1}^{n}\beta_i\xi_i \tag{4}$$

其中, α_i 和 β_i 为拉格朗日乘子, ν 为对目标类别样本的拒绝度, $0 < \nu \le 1$, $n\nu$ 则为外点数量的上限和支持向量的下限. 令 L 分别对 R, a 和 ξ 求偏导, 并令其为 0, 可得:

$$\sum_{i=1}^{n}\alpha_i = 1, \quad a = \sum_{i=1}^{n}\alpha_i\Phi(x_i), \quad \alpha_i = \frac{1}{n\nu} - \beta_i \tag{5}$$

将式 (5) 代入式 (4), 并将内积 $\Phi(x_i)\Phi(x_j)$ 用 Mercer 核函数 $K(x_i,\ x_j)$ 代替, 可得原最优问题的 Wolfe 对偶问题为:

$$\begin{aligned} &\max{}_{\alpha}\, L = \sum_{i}^{n}\alpha_i K(x_i,\ x_i) - \sum_{i,j=1}^{n}\alpha_i\alpha_j K(x_i,\ x_j) \\ &\text{s.t.}\quad \sum_{i=1}^{n}\alpha_i = 1,\ \ 0 \le \alpha_i \le \frac{1}{n\nu} \end{aligned} \tag{6}$$

根据 KKT 条件, 样本数据可按照 α_i 和 β_i 的取值不同分为三类: 内点, 位于超球体内部, 其 $||\Phi(x_i)-a||^2<R$, 即 $\alpha_i=0$, $\beta_i=\frac{1}{n\nu}$; 支持向量, 位于超球体边界, 其 $||\Phi(x_i)-a||^2=R$, 即 $0<\alpha_i<\frac{1}{n\nu}$, $\beta_i>0$; 外点, 位于超球体外部, 其 $||\Phi(x_i)-a||^2>R$, 即 $\alpha_i=\frac{1}{n\nu}$, $\beta_i=0$. 为了验证样本数据的类型, 可通过判断测试样本是否在超球体内来定义分类函数, 决策函数如下:

$$f(x)=\mathrm{sgn}[R^2-||\Phi(x_i)-a||^2] \tag{7}$$

由式 (5) 中的 $a=\sum_{i=1}^{n}\alpha_i\Phi(x_i)$ 和 Mercer 核函数定义 $K(x_i,\ x_j)=\Phi(x_i)\Phi(x_j)$ 可得:

$$||\Phi(x_i)-a||^2=\sum_{i,j=1}^{n}\alpha_i\alpha_jK(x_i,\ x_j)-2\sum_{i}^{n}\alpha_iK(x_i,\ x)+K(x,\ x) \tag{8}$$

将式 (8) 代入式 (7), 决策函数可转化为:

$$f(x)=\mathrm{sgn}\left[R^2-\sum_{i,j=1}^{n}\alpha_i\alpha_jK(x_i,\ x_j)+2\sum_{i}^{n}\alpha_iK(x_i,\ x)-K(x,\ x)\right] \tag{9}$$

由此可得支持向量的决策函数值为 0, 内点的决策函数值大于 0, 外点的决策函数值小于 0. 高斯核函数由于其能够形成任意形状的闭合的凸包, 经常被选择作为 SVDD 中的核函数.

4　基于局部支持向量数据描述的稀有类分析算法

本文在 SVDD 的基础上提出了一种基于局部支持向量数据描述的稀有类分析算法 LSVDD. 给定训练样本, LSVDD (核函数选择高斯核函数 $K(x,y)$=$\exp(-||x-y||^2/\sigma^2)$) 首先利用 SVDD 对样本中每个类别分别进行单类学习, 然后基于得到的多个单类模型的决策函数进行计算, 寻找概念重叠区域. 如果一个样本在至多一个单类模型上的决策函数值大于 0, 那么它位于概念非重叠区域, 取标准化后的决策函数最大值即可获得样本的类别. 但如果一个样本在至少两个单类模型上的决策函数值都大于 0, 则认为其落入了概念重叠区域, 需要进一步进行概念重叠区域学习. 重叠区域产生的原因往往是属性缺失或冗余属性的存在. 属性缺失通常起因于数据采集阶段的信息缺失, 在数据分析阶段很难有针对性的解决方法. 因此, 对于属性缺失, 需要重新收集缺失的属性信息, 使得重叠的数据得以区分; 对于冗余属性, 需要考察数据所包含的各属性, 删去与类别无关或关系不大的属性, 而保留与类别关联紧密的属性. 本文基于已有公共数据集对算法进行比较, 假设采用的公共数据集不存在属性缺失问题, 而只针对属性冗余进行了处理. LSVDD 针对概念重叠区域样本, 通过属性选择方法筛除冗余属性, 减少冗余属性对样本重叠带来的影响, 然后再对其进行 SVDD 学习获得概念重叠区域的单类模型. 最后结合概念非重叠区域和概念重叠区域的单类模型得到综合分类模型以对稀有类进行预测. LSVDD 算法分为三个阶段, 其伪代码如图 2 所示.

第一阶段, 对于样本中的每个类别 $i(i=1,2,\cdots,c)$, 分别通过 SVDD 进行学习, 获得每个类别的单类模型 M_i, 并计算所有样本在各决策函数上的值, 供第二阶段的概念重叠区域确定使用.

第二阶段, 利用第一阶段的决策函数值获得概念重叠区域样本. 如果一个样本在至多一个单类模型上的决策函数值大于 0, 那么它位于概念非重叠区域; 如果一个样本在至少两个单类模型上的决策函数值都大于 0, 那么它位于概念重叠区域. 然后通过并集得到整体概念重叠区域, 并对概念重叠区域中的所有样本数据使用属性选择方法筛除冗余属性. 本文选取较为简单的 *F-socre* 属性选择方法选取 *F-socre* 值大于平均 *F-socre* 值的属性. *F-socre* 的计算方法如下 [17]:

$$F\text{-}socre(p)=\frac{\sum_{i=1}^{c}(\bar{x}_p^{(i)}-\bar{x}_p)^2}{\sum_{i=1}^{c}\frac{1}{n_i-1}\sum_{k=1}^{n_i}(x_{k,p}^{(i)}-\bar{x}_p^{(i)})^2} \tag{10}$$

其中, p 代表属性, c 代表概念重叠区域类别的数目, n_i 代表类别为 i 的概念重叠区域样本的数目, $x_{k,p}^{(i)}$ 代表类别为 i 的概念重叠区域样本 x_k 的属性 p 的值, $\bar{x}_p$ 代表所有概念重叠区域样本的属性 p 的均值, $\bar{x}_p^{(i)}$ 代表类别为 i 的概念重叠区域样本的属性 p 的均值. 最后对经过属性选择处理后的概念重叠区域再次运用 SVDD 对每个类别 i 进行局部学习, 获得概念重叠区域的单类模型 Mo_i.

第三阶段, 利用单类模型 M 和概念重叠区域单类模型 Mo 对测试样本的类别进行预测. 对于每一个测试样本通过第一阶段学习得到的单类模型 M 进行计算, 如果其在至多一个单类模型上的决策函数值大于 0, 那么它位于概念非重叠区域, 取标准化后的决策函数最大值即可获得样本的类别; 如果其在至少两个单类模型上的决策函数值都大于 0, 那么它位于概念重叠区域, 通过第二阶段学习得到的概念重叠区域单类模型 Mo 进行计算, 取标准化后的概念重叠区域的决策函数最大值即可获得样本的类别.

LSVDD (Local Support Vector Data Description)

输入: 训练集 D, 类别数 c, 测试集 T, 惩罚系数 ν, 高斯核函数参数 σ, SVDD 算法

输出: 分类结果 L

```
begin
Phase 1: SVDD 学习, 边界获取
    for class i = 1 to c //c 是训练样本中的类的数目
        M_i = train(SVDD(ν,σ),D(i));
    end for
Phase 2: 概念重叠区域学习
    for class i = 1 to c
        for class j = i to c
            Do_ij = findOverlapRegion(D, M_i, M_j);//概念重叠区域样本确定
        end for
    end for
    Do = ⋃_{1≤i<j≤c} Do_ij;//概念重叠区域样本合并
    As= attributeSelection(Do); //属性选择
    for class i = 1 to c
        Mo_i = train(SVDD(ν,σ),Do(i),As); //概念重叠区域学习, Do(i) 为 Do 中 i 类样本
    end for
Phase 3: 预测
    L= predictLabel(T, M, Mo); //利用综合模型预测
end
```

图 2 LSVDD 算法伪代码

LSVDD 旨在通过单类模型寻找概念重叠区域, 针对概念重叠区域进行局部学习, 提高概念重叠区域边界的精度, 从而获得对数据边界描述更加精确的模型. 因此 LSVDD 适用于迭代策略: 如果单类模型 M 的概念重叠区域经过属性选择后, 再次进行局部单类学习获得的单类模型 Mo 仍然存在较多的概念重叠区域, 则可对概念重叠区域样本 Do 迭代使用 LSVDD 算法. 此时, 可设置迭代次数 k 或者概念区域重叠比率阈值 r 作为迭代停止的条件, 当超过迭代次数 k 或者概念重叠区域比例低于阈值 r 时停止迭代. 最后对于迭代结束后概念重叠区域获得的单类模型仍然存在重叠区域时, 则与 CSVDD 算法的处理方法相同, 通过计算经过标准化后的概念重叠区域的各单类模型的决策函数值中的最大者对样本进行划分. 本文实验部分实现的 LSVDD 算法不对概念重叠区域进行迭代处理.

5 实验结果和分析

实验分为两个部分, 分别在仿真数据和 UCI 数据集上进行, 所有算法均使用 Tax 提供的在 MATLAB 平台的实现 Data description 工具箱 [18]. 实验通过 10 折交叉验证方法验证分类算法的分类效果.

5.1 仿真数据实验

此实验目的在于直观展示 LSVDD 算法的特点和优势, 证明 LSVDD 算法与 SVDD 算法相比能够有效地解决概念重叠区域的分类问题, 从而提高最终的分类效果. 实验需要用到的参数包括拒绝度 ν 和所选用的高斯核函数 $K(x,y)=\exp(-||x-y||^2/\sigma^2)$ 参数 σ, 取值区间分别为 [0.01,0.1] 和 [1,10], 步长分别为 0.01 和 1. 根据不同参数值的实验结果选取了 Data description 工具箱的默认值 0.05 和 5 作为参数的值. 使用的第一个仿真数据由两个 banana 形的数据集组成, 类别分别为 1 和 2. 它们是存在概念重叠的二维点集, 共有 100 个数据, 50 个为类 1, 50 个为类 2. 数据集分布如图 3 所示.

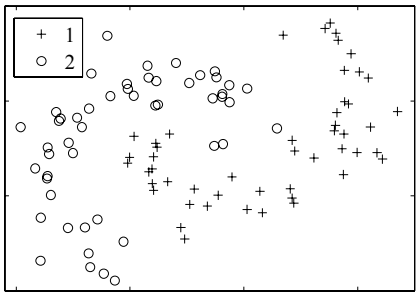
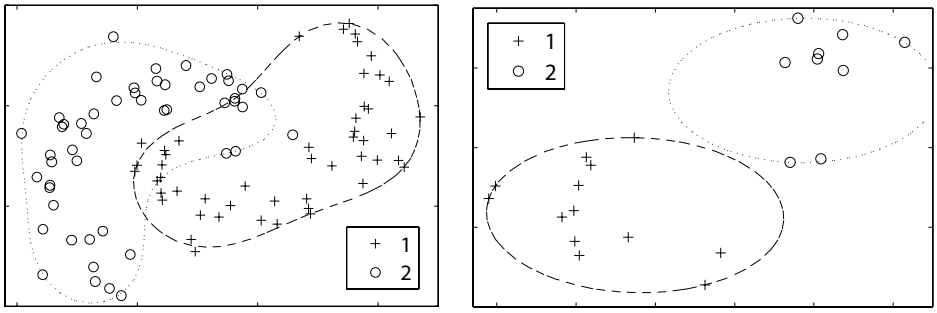


图 3 概念重叠仿真数据分布图

对概念重叠仿真数据分别使用 SVDD 和 LSVDD 算法. 图 4(a) 是对每个类分别进行 SVDD 学习后获得的单类模型边界; 图 4(b) 是 LSVDD 识别出概念重叠区域后, 对概念重叠区域各类样本进一步进行 SVDD 单类学习后获得的概念重叠区域单类模型边界. 从图 4(a) 中可以看出, SVDD 进行训练后得到的数据区域往往存在重叠情况. 因此无论选择哪个类作为目标类, 通过 SVDD 获得的边界来进行划分, 都会由于边界不精确而出现误差; 而 LSVDD 有效地找到了概念重叠区域, 并针对概念重叠区域进行局部学习, 获得了概念重叠区域的数据边界, 综合初始边界得到了更加精确的数据边界描述.



(a) SVDD 获得的数据边界　　(b) LSVDD 获得的概念重叠区域数据边界

图 4　SVDD 与 LSVDD 数据边界比较

第二个仿真数据是样本重叠仿真数据. 它是由一定区域内服从不同密度的均匀分布的数据集组成, 类别分别为 1 和 2, 共有 1100 个数据. 其中, 100 个为类 1, 1000 个为类 2. 数据集分布如图 5 所示, 中间为样本重叠区域, 两侧为各类的样本非重叠区域. 类 1 在样本重叠区域上半部分的密度与左侧的密度相同, 在样本重叠区域下半部分的密度是上半部分的 1/3. 类 2 在样本重叠区域下半部分的密度与右侧的密度相同, 在样本重叠区域上半部分的密度是下半部分的 1/3. 实验采用召回率 (recall), 精度 (precision), F 值 (F-measure) 和 AUC 作为算法绩效评价指标, 将类 1 和类 2 分别作为目标类进行评价. 其中, F 值 =(2× 召回率 × 精度)/(召回率 + 精度); AUC 是 ROC (receiver operating characteristic) 曲线下的面积 [2], 其值一般大于等于 0.5. 对于 AUC 小于 0.5 的, 通过将 ROC 曲线转换后再计算 AUC. AUC 越大, 分类器对正类的分类精度越高, 整体的分类性能也越好 [19]. 由于 SVDD 获得的决策函数值越大则为正类的概率越大, 故 AUC 存在小于 0.5 的情况, 此时表明分类效果比随机猜测的效果差 [20].

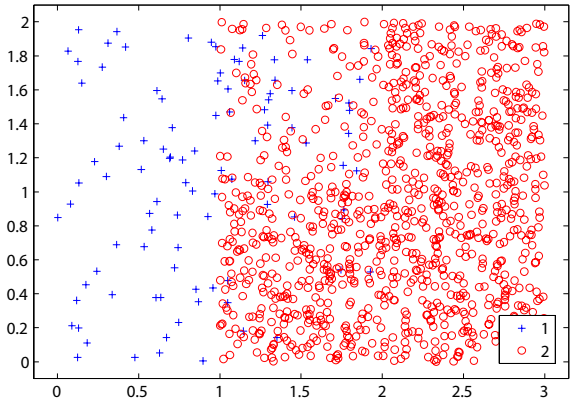


图 5　样本重叠仿真数据分布图

表 1 给出了 SVDD 和 LSVDD 算法在样本重叠仿真数据集上的评价指标值, 其中 “类别” 编号后的括号内为目标类名称及训练样本个数, 粗体表示的是对应评价指标值中最优者. 表中的 R, P, F 和 A 分别代表召回率, 精度, F 值和 AUC. 从表 1 的实验结果看, 在处理样本重叠仿真数据时, 无论是对普通类 2, 还是稀有类 1, LSVDD 都具有比 SVDD 更好的分类效果.

表 1　SVDD 与 LSVDD 在样本重叠仿真数据集上的结果比较

类别	SVDD				LSVDD			
	R	P	F	A	R	P	F	A
1 (1: 100)	0.870	0.641	0.678	0.855	**0.920**	**0.658**	**0.763**	**0.942**
2 (2: 1000)	0.578	0.872	0.697	0.593	**0.699**	**0.961**	**0.775**	**0.949**

5.2　UCI 数据实验

本实验的目的是为了验证 LSVDD 算法在 UCI 实际数据集上的分类效果, 同样采用召回率, 精度, F 值和 AUC 作为算法绩效评价指标, 将每个类作为目标类进行评价. 选取 UCI 数据库 [21] 中 6 个数据集, 其特

征如表 2 所示. 其中, Breast-w 数据集删除了存在属性缺失的 9 个样本, Glass 数据集删除了样本数少于 20 的类别. 待比较的算法选择了 SVDD、CSVDD 和 LSVDD, 核函数均为高斯核函数. 参数 ν 和 σ 取值区间分别为 [0.01,0.1] 和 [1,10], 步长分别为 0.01 和 1. 根据不同参数值的实验结果选取了 Data description 工具箱的默认值 0.05 和 5 做为作为参数 ν 和 σ 的值.

表 2 实验所用 UCI 数据集特征

数据集	样本数	类别数	属性数	最大类样本数	最小类样本数
Breast-w	683	2	9	444	239
Ionosphere	351	2	34	225	126
Sonar	208	2	60	111	97
Balance	625	3	4	288	49
Glass	175	3	9	76	29
Vehicle	946	4	18	226	240

表 3 给出了各分类算法在二分类数据集上的评价指标值, 其中数据集中 "类别" 编号后的括号内为目标类名称及训练样本个数, 粗体表示的是对应评价指标值中最优者. 从表 3 的实验结果看, 在处理二分类问题时, LSVDD 在绝大多数情况下具有比 SVDD 和 CSVDD 更好的分类效果, 特别是精度、F 值和 AUC 评价指标. 仅在 Ionosphere 数据集的类 2 上稍劣于 SVDD 和 CSVDD. 针对稀有类可以发现 LSVDD 与 SVDD 和 CSVDD 相比, 在处理二分类问题时有较好和稳定的稀有类预测效果.

表 3 算法在二分类数据集上的结果比较

数据集	SVDD				CSVDD				LSVDD			
	R	P	F	A	R	P	F	A	R	P	F	A
Breast-w												
1 (benign:444)	0.704	**0.882**	0.780	0.703	0.861	0.839	0.843	0.697	**0.865**	0.858	**0.861**	**0.887**
2 (malignant:239)	0.406	**0.959**	0.564	0.936	0.500	0.892	0.811	0.896	**0.933**	0.875	**0.907**	**0.951**
Ionosphere												
1 (b:126)	0.694	0.278	0.396	0.315	**1.000**	0.359	0.523	0.877	0.937	**0.412**	**0.569**	**0.908**
2 (g:225)	**0.795**	0.816	**0.807**	0.791	0.304	0.933	0.529	**0.828**	0.412	**0.941**	0.656	0.808
Sonar												
1 (mine:111)	**0.691**	0.470	0.553	0.501	0.607	0.544	0.561	0.578	0.609	**0.562**	**0.566**	**0.589**
2 (rock:97)	**0.832**	0.471	0.599	0.558	0.535	0.576	0.546	0.580	0.747	**0.594**	**0.623**	**0.606**

表 4 算法在多分类数据集上的结果比较表

数据集	SVDD				CSVDD				LSVDD			
	R	P	F	A	R	P	F	A	R	P	F	A
Balance												
1 (L:288)	0.726	0.367	0.424	0.773	**1.000**	0.420	0.513	0.879	0.802	**0.548**	**0.544**	**0.905**
2 (B:49)	0.557	0.715	0.571	0.544	0.479	0.816	0.582	0.710	**0.591**	**0.923**	**0.608**	**0.783**
3 (R:288)	**0.584**	0.739	0.610	0.778	0.544	0.761	0.639	0.913	0.542	**0.785**	**0.677**	**0.915**
Glass												
1 (float:70)	0.843	0.564	0.671	0.768	0.586	0.647	0.625	0.751	**0.886**	**0.692**	**0.759**	**0.831**
2 (non-float:76)	**0.882**	0.487	0.626	0.464	0.652	0.659	0.653	0.519	0.695	**0.665**	**0.682**	**0.725**
3 (headlamps:29)	0.733	0.341	0.429	0.477	**0.900**	0.698	0.744	0.881	**0.900**	**0.739**	**0.776**	**0.905**
Vehicle												
1 (bus:240)	0.368	0.256	0.304	0.659	**0.472**	**0.334**	**0.416**	**0.801**	**0.472**	**0.334**	**0.416**	**0.801**
2 (opel:240)	0.498	0.449	0.470	0.401	**0.516**	**0.654**	**0.592**	**0.630**	**0.516**	**0.654**	**0.592**	**0.630**
3 (saab:240)	**0.785**	0.256	0.411	0.438	0.564	**0.383**	**0.513**	**0.652**	0.564	**0.383**	**0.513**	**0.652**
4 (van:226)	0.300	0.233	0.277	0.488	**0.316**	**0.509**	**0.381**	**0.905**	**0.316**	**0.509**	**0.381**	**0.905**

表 4 则给出了各分类算法在多分类数据集上的评价指标值. 从表 4 的实验结果看, 在处理多分类问题

时, CSVDD 和 LSVDD 都优于或近似于 SVDD 的分类效果, 说明有效利用多个类的信息有助于提高分类效果. 同时 LSVDD 在数据集 Balance 和 Glass 上的精度、F 值和 AUC 值绝大部分都大于 CSVDD 的值, 说明 LSVDD 对概念重叠区域的处理是有效的. 在数据集 Vehicle 上, LSVDD 具有与 CSVDD 相似的分类效果, 这是由于此时概念重叠区域较少. 针对数据集中的稀有类可以发现, LSVDD 的评价指标值基本都优于或等于 CSVDD, 说明 LSVDD 有效提高了稀有类预测效果, 在处理多分类问题时有较好和稳定的稀有类预测效果.

实践中的稀有类分析问题往往存在信息不完全情况, 如在入侵检测中, 可能只知道某一类入侵模式, 而把其他入侵模式与正常模式混淆. 为了检验 LSVDD 在信息不完全数据上的分类效果, 本文利用多分类数据集模拟信息不完全情况, 即每次只将多分类数据集中的一类作为已知目标类, 其他类都归入非目标类. 表 5 给出了各分类算法在信息不完全多分类数据集上的各个评价指标值. 从表 5 的实验结果看, 在信息不完全情况下, 只有在 Vehicle 数据集的类 1 上, CSVDD 和 LSVDD 的评价指标值远小于 SVDD. 其他情况下, CSVDD 和 LSVDD 都优于或近似于 SVDD 的分类效果, 说明在信息不完全条件下有效利用非目标类的信息有助于分类效果的提高. 进一步, 由于非目标类由多类样本混合而成, 其数据边界不够精确, 更加容易产生重叠区域, 因此 LSVDD 的各评价指标在绝大部分情况下都优于 CSVDD. 这说明 LSVDD 能够有效地对概念重叠区域进行处理, 提高信息不完全情况下的整体分类效果. 针对数据集中的稀有类可以发现, LSVDD 只在 Glass 数据集的类 3 上的 AUC 值稍劣于 CSVDD, 说明 LSVDD 在信息不完全情况下也能有效提高稀有类预测效果.

表 5　算法在信息不完全多分类数据集上的结果比较

数据集	SVDD				CSVDD				LSVDD			
	R	P	F	A	R	P	F	A	R	P	F	A
Balance												
1 (L:288)	0.726	0.367	0.424	0.773	**0.822**	0.415	0.458	0.834	0.787	**0.481**	**0.516**	**0.890**
2 (B:49)	**0.557**	0.715	0.571	0.544	0.456	0.799	0.517	0.536	0.533	**0.813**	**0.592**	**0.546**
3 (R:288)	**0.584**	0.739	0.610	0.778	0.525	0.706	0.596	0.794	0.502	**0.771**	**0.638**	**0.803**
Glass												
1 (float:70)	**0.843**	0.564	0.671	0.768	0.577	0.610	0.589	0.755	0.700	**0.684**	**0.695**	**0.769**
2 (non-float:76)	**0.882**	0.487	0.626	0.464	0.630	0.652	0.643	0.485	0.681	**0.657**	**0.669**	**0.608**
3 (headlamps:29)	0.733	0.341	0.429	0.477	0.837	0.501	0.642	**0.878**	**0.895**	**0.594**	**0.686**	0.837
Vehicle												
1 (bus:240)	**0.368**	**0.256**	**0.304**	**0.659**	0.269	0.227	0.238	0.391	0.335	0.232	0.273	0.442
2 (opel:240)	0.498	0.449	0.470	0.401	**0.507**	0.584	0.482	0.470	0.505	**0.632**	**0.577**	**0.508**
3 (saab:240)	0.785	0.256	0.411	0.438	0.526	0.312	0.423	0.478	**0.541**	**0.378**	**0.494**	**0.517**
4 (van:226)	0.300	0.233	0.277	0.488	0.291	0.415	0.246	0.586	**0.304**	**0.456**	**0.319**	**0.661**

综上所述, LSVDD 通过数据边界描述可以发现数据固有结构, 并寻找概念重叠区域, 针对概念重叠区域进行局部单类学习, 优化概念重叠区域数据边界, 因此与 SVDD 和 CSVDD 相比具有更好的稀有类分析效果. 即使在信息不完全情况下, LSVDD 也能够获得较高的稀有类分析精度, 这主要是因为信息不完全情况下非目标类的数据边界更加模糊, 因此更加容易与目标类的边界产生更多的概念重叠. 总的来说, 本文提出的 LSVDD 算法在解决稀有类分析问题上是稳定和有效的.

6　结论

本文提出了一种基于局部支持向量数据描述的稀有类预测算法 LSVDD. LSVDD 首先利用 SVDD 算法进行单类学习, 并找出概念重叠区域, 然后针对概念重叠区域采用属性选择方法筛除冗余属性, 并进行进一步的局部单类学习, 最终获得综合分类模型来对稀有类进行预测. LSVDD 将传统的 SVDD 算法进行了扩展, 并解决了 CSVDD 在概念重叠区域的分类效果不佳的问题. 实验结果表明, LSVDD 算法有效改善了数据边界的识别, 提高了稀有类分析精度, 并且在信息不完全情况下也有较好效果. 未来研究将对样本重叠中的属性缺失问题进行进一步研究, 并优化 LSVDD 的迭代实现, 进一步提高分类界面的准确性.

参考文献

[1] Weiss G M. Mining with rarity: A unifying framework[J]. SIGKDD Explorations, 2004, 6(1): 7–19.

[2] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining[M]. New York: Addison Wesley, 2005: 95–98.

[3] Wu J, Xiong H, Chen J. COG: Local decomposition for rare class analysis[J]. Data Mining and Knowledge Discovery, 2010, 20(2): 191–220.

[4] He H, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263–1284.

[5] 刘叶青, 刘三阳, 谷明涛. 多项式光滑的半监督支持向量分类机 [J]. 系统工程理论与实践, 2009, 29(7): 113–118.
Liu Y Q, Liu S Y, Gu M T. Improved learning algorithm with transductive support vector machines[J]. Systems Engineering — Theory & Practice, 2009, 29(7): 113–118.

[6] Cortes C, Vapnik V N. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273–297.

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

[8] Tax D, Duin R. Growing a multi-class classifier with a reject option[J]. Pattern Recognition Letters, 2008, 29(10): 1565–1570.

[9] 徐晶, 石端银, 张亚江, 等. 基于聚类和 SVDD 的单类入侵检测模型 [J]. 控制与决策, 2010, 25(3): 441–444.
Xu J, Shi D Y, Zhang Y J, et al. Model of IDS based on SVDD and cluster algorithm[J]. Control and Decision, 2010, 25(3): 441–444.

[10] Prati R C, Batista G. Class imbalances versus class overlapping: An analysis of a learning system behavior[C] // Proceedings of the Mexican International Conference on Artificial Intelligence, Berlin: Springer, 2004: 312–321.

[11] Weiss G M, Provost F. Learning when training data are costly: The effect of class distribution on tree induction[J]. Journal of Artificial Intelligence Research, 2003, 19: 315–354.

[12] Visa S, Ralescu A. The effect of imbalanced data class distribution on fuzzy classifiers-experimental study[C]// Proceedings of the IEEE Conference on Fuzzy Systems, Washington: IEEE Press, 2005: 749–754.

[13] Wu J, Xiong H, Wu P, et al. Local decomposition for rare class analysis[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press, 2007: 814–823.

[14] Gangemi A, Pisanelli D M, Steve G. An overview of the ONIONS project: Applying ontologies to the integration of medical terminologies[J]. Data and Knowledge Engineering, 1999, 31(2): 183–220.

[15] García V, Mollineda R A, Sánchez J S. On the k-NN performance in a challenging scenario of imbalance and overlapping[J]. Pattern Analysis & Applications, 2008, 11(3/4): 269–280.

[16] Liu C L. Partial discriminative training for classification of overlapping classes in document analysis[J]. International Journal on Document Analysis and Recognition, 2008, 11(2): 53–65.

[17] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003(3): 1157–1182.

[18] Tax D. DD_tools[EB/OL]. http://www-ict.ewi.tudelft.nl/~davidt/dd_tools.html.

[19] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(6): 1145–1159.

[20] Juszczak P, Tax D, Pe E, et al. Minimum spanning tree based one-class classifier[J]. Neurocomputing, 2009, 72(7/9): 1859–1869.

[21] Blake C L, Merz C J. UCI repository of machine learning databases[EB/OL]. http://kdd.ics.uci.edu.