分类中的类重叠问题及其处理方法研究①

熊海涛[1,2]，吴俊杰[2]，刘洪甫[2]，刘 鲁[2]
(1. 北京工商大学计算机与信息工程学院，北京 100048;
2. 北京航空航天大学经济管理学院，北京 100191)

摘要: 类重叠问题是数据挖掘与机器学习领域的瓶颈问题之一. 如果其中还存在类不均衡问题时,情况变得更加复杂. 有鉴于此,本文在已有文献基础上归纳了三种类重叠学习算法及提出一种新的方法: 分隔法,并首次将支持向量数据描述算法用于实际数据的重叠样本识别,对类重叠问题及其与类不均衡问题的相互影响进行了系统研究. 在真实数据上采用五种分类器的实验结果表明: 1) 多数情况下"分隔法"是表现最佳的类重叠学习算法; 2) 分隔法通常对基于分界面而非规则的分类器更为有效; 3) 分隔法在类不均衡问题中表现很好,当基础分类器为支持向量机时尤为突出. 最后针对支持向量机的实验结果给出了理论分析.

关键词: 数据挖掘; 分类; 类重叠; 类不均衡; 支持向量数据描述
中图分类号: TP181 **文献标识码:** A **文章编号:** 1007-9807(2013)04-0008-14

0 引 言

在信用卡欺诈识别、客户流失分析、企业财务困境预测、设备失效预测等大量管理问题中,数据挖掘的分类(classification)技术被广泛使用[1-4]. 然而,在这些实际问题中,由于数据采集的客观条件,不同类别的样本往往在某些属性上具有相似的取值,从而使得分类效果不佳. 由于这些样本位于属性空间的重叠区域,因此被称为重叠样本,由其引发的问题也被称为"类重叠"(class overlapping)问题. 类重叠问题是数据挖掘和机器学习领域的瓶颈问题之一. 研究表明,分类器划分错误往往集中在数据的边界区域,而这正是类重叠经常存在的区域,很难找到一种行之有效的方法来解决类重叠问题[5,6]. 考虑到上述信用卡欺诈识别等问题中,往往还存在异常样本和正常样本数量相差较大的类不均衡(class imbalance)问题[1,7],此时异常样本的预测难度要远大于正常样本,因此情况变得更加复杂.

鉴于类重叠问题和类不均衡在管理中的重要性和复杂性,许多学者致力于研究两者的解决方法. 对于类不均衡问题,目前的研究比较多,大量的算法被提出来如重抽样、成本敏感学习、调整归纳偏置等[8,9]. 本文也对类不均衡问题提出了一种基于局部聚类的稀有类分析算法,对普通类进行局部聚类,然后再进行分类,能够获得较好的效果[1]. 而对于类重叠问题的研究则较少. Prati 和 Batista 等通过人工生成重叠数据对类重叠问题进行系统研究,结果表明类重叠的程度与类不均衡有紧密的联系[10]. 此外,Visa 和 Ralescu 使用模糊分类器进行实验,发现类重叠对分类效果的影响比类不均衡大[11]. 同时 García 和 Mollineda 等则将多个不同类型的基本分类器,分别是 k-NN、MLP、NB、RBF 和 C4.5,应用于在重叠区域具有不同的类比例的人工数据集上,实验结果表明基于全局学习的分类器有利于类重叠区域的大类样

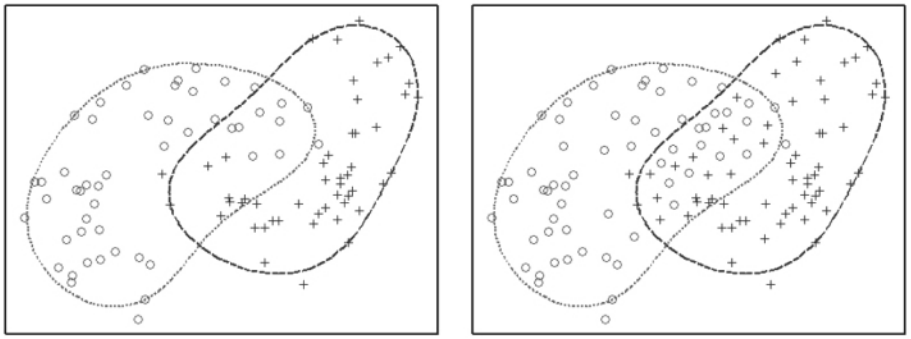
① 收稿日期: 2011-03-16; 修订日期: 2012-01-06.
基金项目: 国家自然科学基金资助项目(71201004; 70901002); 国家自然科学基金重大研究计划资助培育项目(90924020); 北京市教育委员会科技发展计划面上项目(km201310011009); 北京市大学生科学研究与创业行动计划建设项目(pxm2012_014213_000067).
作者简介: 熊海涛(1983—),男,江西九江人,博士. Email: xionghtao@gmail.com

本的划分[12]. 然而上述研究主要停留在利用人工生成的重叠数据研究类重叠是否对最终分类效果产生影响,关注的是基本分类器在处理类重叠问题时的效果,并没有真正解决类重叠问题,很难应用于实际数据. 因此从实践的角度来说,对发现实际数据中的类重叠区域、解决类重叠问题的方法进行系统研究,具有十分重要的现实意义. Tang 和 Gao 在处理类重叠问题时,尝试通过 k-NN 寻找重叠区域的样本[13],但 k-NN 在处理复杂数据时很难有效发现不同类别数据的边界特征. 支持向量数据描述算法(support vector data description,SVDD)是一种单类学习方法[14]. SVDD 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,能够刻画数据的空间分布,从而在一定程度上揭示数据在分布空间的固有结构,并用来进行分类. 然而其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. 此外,对于类重叠问题与类不均衡问题的相互影响还缺乏系统研究,没有明确两者是如何影响最终的分类效果,究竟是何者起了真正影响.

针对上述不足,本文系统地研究了分类中的类重叠问题,主要贡献体现在如下几个方面. 首先,本文在前续研究成果[15]的基础上首次将 SVDD 算法用于实际数据的重叠样本识别. 其次,本文采用舍弃法、合并法、层次法和分隔法来解决类重叠问题,并对这四种方法的优劣和适用情况进行了分析. 前三种方法主要对已有的类重叠处理方法[13,16]进行归纳得到,同时借鉴了文字识别[5,17-20]和多标签学习领域[21,22]中处理重叠问题的方法(必须指出,这些研究中重叠样本是通过已知的多标签信息给定的,而本文研究则利用 SVDD 算法来寻找真实数据中无法通过标签信息得到的重叠样本);第四种方法即分隔法则是本文的创新. 再次,本文在实际数据上对类重叠问题及其与类不均衡问题的相互影响进行了系统研究. 实验结果表明: 1) 多数情况下分隔法是表现最佳的类重叠学习算法; 2) 分隔法通常对基于分界面而非规则的分类器更为有效; 3) 分隔法在类不均衡问题中表现很好,当分类器为支持向量机时尤为突出. 最后,本文以支持向量机为例对分隔法在不均衡数据分析上的作用进行了理论探讨.

现实数据中往往存在着一定的类重叠现象,因此类重叠问题吸引了大批学者的研究兴趣. n 维空间 $\boldsymbol{R}^n$ 中的类重叠区域可以描述为: 至少有两个不同的类 C_a 和 C_b 同时出现的概率大于 0 的区域 A,即对于任意的重叠区域样本 $x \in A$,$p(x \mid C_a) > 0$ 且 $p(x \mid C_b) > 0$[23]. 显然,如果重叠区域较大,那么分类器的精度就会比较差[24].

类重叠问题可分为概念重叠(concept overlapping)和样本重叠(sample overlapping)两个层次[5,6,25]. 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,属于宏观层面的重叠. 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,属于微观层面的重叠. 图 1 给出了类重叠的两个例子,其中图 1a 是概念重叠的例子,图 1b 则是样本重叠的例子,虚线表示的是类的边界也就是概念. 可以发现,当两类数据较为接近时概念重叠就容易发生,此时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.



(a) 概念重叠例子　　(b) 样本重叠例子

图 1　类重叠示例

Fig. 1 Class overlapping examples

已有研究表明分类器划分错误往往集中在数据的边界区域[1,3]，这正是概念重叠经常存在的区域. 严格意义上的样本重叠是很少发生的，除非出现了严重的数据采集问题或者数据属性缺失，这在基于样本独立假设的数据分析中很难得到解决. 真实数据通常表现为存在广泛的概念重叠，因此若无特别说明，下文所提的类重叠问题均为概念重叠问题. 此时分布稀疏的样本中的非重叠部分可能距边界所在的重叠区域较远，对边界的确定影响较小甚至无关. 而传统的分类器使用所有样本训练，其中包含的非重叠区域样本可能导致决策面偏移，从而使得分类误差尤其是稀有类的分类误差较大. 因此，重叠样本对边界的学习有着至关重要的作用，针对重叠区域进行学习能够提高分类效果 —— 这就是本文研究的一个重要起点.

1 类重叠学习框架

本节对类重叠学习的相关方法进行介绍，并给出完整的类重叠学习框架. 利用该框架可以实现对多种类重叠学习算法和不同分类器的统一调用.

1.1 SVDD: 重叠区域识别方法

首先应识别数据中的概念重叠区域. 为此引入支持向量数据描述(SVDD) 算法. SVDD 针对单类样本进行学习，寻找一个高维空间的超球体来覆盖尽可能多的数据在该特征空间的映像，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 $X = \{x_i, i = 1, \cdots, n\}$，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 建立如下优化问题[14]

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

其中 v 为对目标类别样本的拒绝度 $0 < v \leqslant 1$，nv 则为外点数量的上限和支持向量的下限. 引入拉格朗日函数可得

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

根据 KKT 条件，样本数据因此分为三类: 内点，位于超球体内部，其 $\| \Phi(x_i) - a \|^2 < R$，即 $\alpha_i = 0, \beta_i = \frac{1}{nv}$; 支持向量，位于超球体边界，其 $\| \Phi(x_i) - a \|^2 = R$，即 $0 < \alpha_i < \frac{1}{nv}, \beta_i > 0$; 外点，位于超球体外部，其 $\| \Phi(x_i) - a \|^2 > R$，即 $\alpha_i = \frac{1}{nv}, \beta_i = 0$.

决策函数 $f(x)$ 为

$$f(x) = \operatorname{sgn}\left[R^2 - \sum_{i,j=1}^{n}\alpha_i\alpha_j K(x_i, x_j) + 2\sum_{i=1}^{n}\alpha_i K(x_i, x) - K(x, x)\right] \tag{3}$$

相应的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. 本文利用 SVDD 算法识别重叠区域. 具体做法是，利用 SVDD 算法对样本中的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比较. 如果一个样本在至多一个单类模型上的决策函数值大于或等于 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大于或等于 0，那么它位于概念重叠区域. 如图 1 所示，红色和蓝色虚线分别为两个类的边界，同时落入两个边界内的样本数据即是位于重叠样本.

1.2 类重叠学习算法

在类重叠问题研究中，Chen 和 Ma 等提出来一种多标签的分类方法对重叠区域的数据引入新标签[16]. 而 Tang 和 Gao 则将重叠区域样本分别与两类合并进行分类，然后对重叠区域再进行细分[13]. 文字识别文献中对类重叠问题的处理方法主要有两个策略，分别是合并策略和精炼策略[5]. 合并策略是将重叠的类别合并为一个新类别，并忽略原来类别的边界之间的差别. 精炼策略则是对重叠的类别进行特别处理，从而对重叠

类别的边界进行精炼.

本文归纳已有类重叠问题的处理策略,并借鉴文字识别中对类重叠问题的处理方法,系统地总结了三种不同的类重叠学习算法: 舍弃法(discarding scheme)、合并法(merging scheme)、层次法(hierarchical scheme),并提出了一种类重叠学习算法: 分隔法(separating scheme). 其中,舍弃法简单地舍弃重叠区域的样本; 合并法基于合并策略,归纳自文献[13]; 层次法和分隔法基于精炼策略,其中层次法归纳自文献[16]. 不失一般性,假设在二分类数据集中存在两个类别"A"和"B",对上述四种学习算法详述如下:

舍弃法　此方法舍弃重叠区域的样本,只利用非重叠区域样本进行学习以获得唯一一个分类模型 M. 每一条测试样本都通过该分类模型进行分类. 具体过程如图 2 所示.

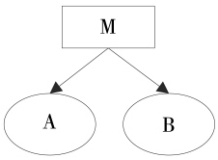


图 2　舍弃法处理过程

Fig. 2 Procedure of the discarding scheme

合并法　此方法将位于重叠区域的样本合并为一个新的类,标记为"O". 接着利用添加了"O"类别的样本进行三分类学习以获得第一个分类模型 M1. 然后对分入重叠区域的样本继续进行"A"、"B"二分类学习以获得第二个分类模型 M2. 每一条测试样本首先通过 M1 进行分类. 如果分类结果为"O"类,则继续通过 M2 进行分类. 具体过程如图 3 所示.

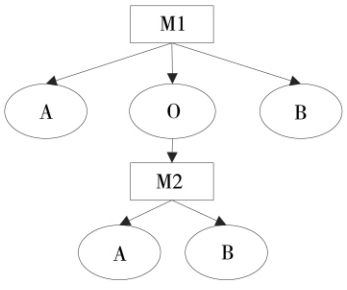


图 3　合并法处理过程

Fig. 3 Procedure of the merging scheme

层次法　此方法进行二分类层次学习. 首先将位于重叠区域的样本合并为一个新的类,标记为"O". 然后将"O"类样本和"A"类样本继续合并,得到一个组合类,标记为"A&O". 学习过程是层次的,首先对类"A&O"和类"B"进行学习获得第一层的分类模型 M1,然后对类"O"和类"A"进行学习获得第二层的分类模型 M2,最后对重叠区域的样本进行学习获得第三层的分类模型 M3. 每一条测试样本依次通过这三个模型进行分类. 具体过程如图 4 所示.

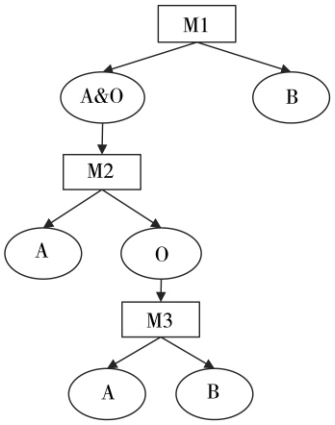


图 4　层次法处理过程

Fig. 4 Procedure of the hierarchical scheme

分隔法　此方法对重叠区域样本和非重叠区域样本进行分隔学习,得到两个分类模型 M1 和 M2. M1 将重叠类样本分为"A"和"B"类,M2 则将非重叠类样本分为"A"和"B"类. 每一条测试样本根据其是否位于重叠区域使用不同的分类模型进行分类. 具体过程如图 5 所示.

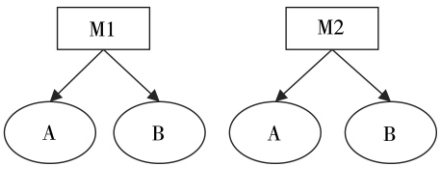


图 5　分隔法处理过程

Fig. 5 Procedure of the separating scheme

1.3　类重叠学习框架

基于上述类重叠学习算法及支持向量数据描述算法,本文提出类重叠学习框架如图 6 所示. 这里以分隔法为例对类重叠学习框架进行说明. 这个框架简单实用,并为使用不同的类重叠学习算法提供了很好的灵活性.

```
输入: 训练集 D,测试集 T,SVDD 算法,分类器 C
输出: 分类结果 p
Begin
//Phase 1: SVDD 学习,重叠区域获取
    for class i = 1 to n        //n 是训练样本中的类的数目
        M_i = train( SVDD,D( i) )
    end for
    for class i = 1 to n
        for class j = i to n
            Do_ij = findOverlapRegion( D,M_i,M_j) ;
        end for
    end for
  Do = ∪_{1≤i<j≤n} Do_ij;
//Phase 2: 类重叠模型学习,采用一对多的方式进行学习
    for class i = 1 to n
        Mo_i = train( SVDD, C, Do, i) ;        // 重叠区域分类模型
        Mn_i = train( SVDD, C, D-Do, i) ;      // 非重叠区域分类模型
    end for
//Phase 3: 利用综合模型预测
    p = predictLabel( T, Mo, Mn) ;
End
```

图 6 类重叠学习框架

Fig. 6 A framework for overlapping class learning

第一阶段,对于样本中的每个类别 $i(i = 1, \cdots, c)$,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. 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 Do.

第二阶段,利用类重叠学习算法进行类重叠模型学习,这里使用的是分隔法. 首先利用第一阶段的重叠区域获取所得到的结果,将数据集 D 划分为重叠区域 Do 和非重叠区域 $D-Do$. 然后采用一对多(one-against-all)的学习方式,利用分类算法 C 在重叠区域和非重叠区域分别进行学习获得重叠区域分类模型 Mo 和非重叠区域分类模型 Mn.

第三阶段,利用综合模型进行预测. 每一条测试样本根据其是否位于重叠区域使用不同的分类模型进行分类. 如果该测试样本位于重叠区域,则使用第二阶段获得的重叠区域分类模型 Mo 对其类别进行预测; 如果该测试样本位于非重叠区域,则使用二阶段获得的非重叠区域分类模型 Mn 对其类别进行预测. 对于多分类数据集最后对预测的结果采取投票法(voting)获得最终的测试样本预测结果.

在运行时间方面,由于四种类处理策略共享重叠样本发现方法,因此时间复杂度没有本质差别. 其中,舍弃法由于没有考虑重叠样本比分隔法表现稍快,而合并法和层次法由于需要建立具有层次结构的多个分类模型比分隔法表现稍慢.

2 实验结果

实验目的是研究各种类重叠学习算法的效果,并对类重叠问题及其与类不均衡问题的关系进行系统分析. 为了进行比较,选取了 UCI[26] 和 LIBSVM[27] 数据库中的七个二分类数据集和三个多分类数据集. 通过 SVDD 算法寻找这十个数据集的类重叠区域,并计算其重叠样本在所有样本中所占的比重. 表 1 给出了实验数据集的相关信息.

表 1　数据集主要特征

Table 1 Main characteristics of data sets

数据集	来源	类别数	样本数	属性数	重叠比例
Inosphere	UCI	2	351	34	41.3%
Fourclass	LIBSVM	2	178	12	31.1%
Sonar	UCI	2	208	60	20.2%
Transfusion	UCI	2	748	10	19.3%
Splice	LIBSVM	2	1 000	60	12.6%
Diabetes	UCI	2	768	8	7.6%
German	LIBSVM	2	1 000	24	5.7%
Glass	UCI	3	175	9	32.8%
Balance	UCI	3	625	4	17.4%
Vehicle	UCI	4	946	18	4.8%

实验分为三部分. 第一部分实验比较四种类重叠学习算法的绩效,其中使用了五种分类器: 决策树 C4.5、规则分类器 RIPPER、朴素贝叶斯(NB)、最近邻法(k-NN,k = 5) 和线性支持向量机(SVMs)[28]. 在这个基础上,第二部分实验探索影响类重叠学习算法分类精度的重要因素. 第三部分实验则关注不均衡数据,通过随机抽样的方法构造不均衡数据集,研究类重叠问题和类不均衡问题的相互影响.

实验使用 10 折交叉验证方法,即数据集被划分为 10 份,每次都选取其中的 9 份作为训练集,剩余的 1 份作为测试集. 实验选择被广泛使用的 F 值(F-measure)[29] 作为评价指标,每个类别都被视为正类计算其 F 值. F 值的计算公式如下

$$F = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

其中 TP,FP 和 FN 分别是真正、假正和假负的样本数目. F 值越大,模型分类效果越好.

2.1　实验 I: 四种类重叠学习方法的比较

本实验旨在寻找在实际数据上效果最优的类重叠学习算法. 表 2 给出了四种类重叠学习算法在不同分类器和二分类数据集上的分类效果. 其中,数据集名称下的值表示样本重叠比例,黑体表示四种方法中的 F 值最大者.

表 3 则给出了四种类重叠学习算法在不同分类器和多分类数据集上的分类效果. 对于多分类数据由于篇幅有限,本文将所有类别的 F 值累加并求平均. 其中,数据集名称后的值表示样本重叠比例,黑体表示四种方法中的平均 F 值最大者.

讨论: 从表 2 和表 3 的实验结果看,在绝大多数情况下分隔法比其它三种方法具有更好的分类效果; 在样本重叠比例较大的数据集上,前者的优势甚至非常明显. 这是由于合并法和层次法在学习的过程中引入了额外的重叠类别,增加了学习的复杂度. 事实上,合并法适用于重叠区域的样本占所有样本比重较大,并且与非重叠区域样本有较明显的区别的情况; 而层次法也适用于重叠区域的样本占所有样本比重较大,但是重叠区域样本与某类样本特征较为相似的情况. 舍弃法在舍弃重叠样本时会丢失一些重要边界信息,而这些信息对于预测未知样本的类别可能是至关重要的. 如果数据集的样本重叠比例较高,丢失的重要信息可能就越多,最终会导致分类精度的降低. 因此,舍弃法适用于重叠区域样本数量不太多,并且重叠区域样本对分类边界无太大影响的情况,此时舍弃重叠区域样本能够减少重叠区域样本对整体分类边界的干扰,反而能够提高最终分类效果. 与上述三种方法相比,本文提出的分隔法将数据集合理地划分为非重叠区域样本和重叠区域样本. 这一方面有利于降低非重叠区域的分类难度,另一方面则可以聚焦于重叠区域寻找分类边界,有利于进行更加精确的学习和降低学习的时间复杂度. 分隔法适用于数据集中存在一定量的重叠样本,并且与非重叠区域分隔较不明显的情况. 真实数据通常表现为这种情况,因此把重叠区域与非重叠区域分隔开来进行学习能够获得较好的分类效果.

表 2 四种类重叠学习算法在二分类数据集上的分类结果

Table 2 Classification results by different learning schemes on binary data sets

数据集	分类器	*F* 值(类 1)				*F* 值(类 2)			
		舍弃	合并	层次	分隔	舍弃	合并	层次	分隔
Inosphere (41.3%)	C4.5	0.766	0.809	0.814	**0.863**	0.788	0.902	0.875	**0.925**
	RIPPER	0.728	0.809	0.819	**0.848**	0.730	0.892	0.877	**0.918**
	NB	0.607	0.849	0.852	**0.861**	0.407	0.909	0.891	**0.914**
	k-NN	0.722	0.668	0.743	**0.805**	0.797	0.873	0.884	**0.910**
	SVMs	0.629	0.760	0.766	**0.786**	0.612	0.897	0.891	**0.905**
Fourclass (31.1%)	C4.5	0.795	0.904	0.852	**0.986**	0.897	0.919	0.952	**0.991**
	RIPPER	0.726	0.913	0.905	**0.956**	0.711	0.886	0.917	**0.980**
	NB	0.533	0.686	0.629	**0.721**	0.685	0.776	0.793	**0.823**
	k-NN	0.862	0.974	**1.000**	**1.000**	0.953	**1.000**	**1.000**	**1.000**
	SVMs	0.584	0.706	0.690	**0.757**	0.641	0.787	0.809	**0.838**
Sonar (20.2%)	C4.5	0.620	0.582	0.597	**0.652**	0.593	0.559	0.604	**0.701**
	RIPPER	0.672	0.613	0.646	**0.678**	0.576	0.512	0.545	**0.620**
	NB	0.689	0.560	0.611	**0.715**	0.479	0.673	**0.675**	0.663
	k-NN	0.642	0.615	0.634	**0.651**	0.569	0.623	0.620	**0.645**
	SVMs	**0.699**	0.663	0.668	0.676	0.665	0.689	0.691	**0.719**
Transfusion (19.3%)	C4.5	0.381	0.487	**0.532**	0.522	0.649	0.718	0.796	**0.864**
	RIPPER	0.388	0.425	0.421	**0.489**	0.696	0.763	0.764	**0.830**
	NB	0.410	0.454	0.476	**0.517**	0.689	0.793	0.845	**0.878**
	k-NN	0.509	0.522	0.574	**0.606**	0.651	0.797	0.803	**0.857**
	SVMs	0.323	0.471	0.495	**0.504**	**0.758**	0.714	0.707	0.742
Splice (12.6%)	C4.5	**0.936**	0.926	0.917	0.928	**0.939**	0.929	0.922	0.930
	RIPPER	0.936	0.878	0.883	**0.941**	0.939	0.880	0.895	**0.943**
	NB	0.821	0.826	0.823	**0.829**	0.833	0.831	0.834	**0.837**
	k-NN	0.775	0.726	0.738	**0.784**	0.723	0.551	0.669	**0.726**
	SVMs	0.787	0.790	**0.792**	0.783	0.803	**0.804**	0.801	0.794
Diabetes (7.9%)	C4.5	**0.622**	0.604	0.604	0.606	**0.804**	0.797	0.781	0.795
	RIPPER	**0.632**	0.625	0.627	0.631	**0.816**	0.802	0.803	**0.816**
	NB	0.623	0.620	0.622	**0.625**	0.824	0.817	0.823	**0.832**
	k-NN	0.608	0.600	0.605	**0.611**	0.807	0.796	0.802	**0.809**
	SVMs	**0.629**	0.617	0.614	**0.629**	**0.832**	0.824	0.826	**0.832**
German (5.7%)	C4.5	0.523	0.469	0.481	**0.537**	**0.828**	0.795	0.796	0.822
	RIPPER	0.483	0.536	**0.538**	0.521	0.823	0.829	0.815	**0.830**
	NB	0.548	0.553	0.549	**0.557**	**0.835**	0.808	0.814	0.830
	k-NN	0.409	0.396	0.407	**0.423**	0.797	0.798	0.785	**0.801**
	SVMs	0.514	0.514	0.519	**0.541**	**0.836**	**0.836**	0.827	0.832

表 3　四种类重叠学习算法在多分类数据集上的分类结果

Table 3 Classification results by different learning schemes on multi-class data sets

数据集	分类器	平均 F 值			
		舍弃	合并	层次	分隔
Glass (32.8%)	C4.5	0.713	0.737	0.746	**0.794**
	RIPPER	0.718	0.789	0.803	**0.820**
	NB	0.682	0.691	0.708	**0.775**
	k-NN	0.754	0.816	0.815	**0.889**
	SVMs	0.749	0.762	0.797	**0.863**
Balance (17.4%)	C4.5	0.502	**0.573**	0.556	0.565
	RIPPER	0.546	0.540	0.528	**0.571**
	NB	0.697	0.742	0.728	**0.766**
	k-NN	0.705	0.717	0.723	**0.782**
	SVMs	0.821	0.894	0.906	**0.921**
Vehicle (4.8%)	C4.5	0.704	0.658	0.642	**0.716**
	RIPPER	**0.701**	0.588	0.603	0.684
	NB	0.427	0.373	0.369	**0.431**
	k-NN	0.719	0.685	0.677	**0.725**
	SVMs	0.786	0.724	0.715	**0.810**

2.2　实验 II: 分隔法细析

实验 I 表明分隔法是提出的四种类重叠学习算法中的最优方法. 本实验研究影响分隔法分类效果的可能因素. 这里原始方法指不进行任何类重叠处理,直接采用分类器进行学习的方法.

图 7 给出了分隔法与原始方法的 F 值,其中 F 是将数据集中每个类都作为正类计算得到的 F 值累加并求平均, "o" 和 "s" 则分别表示原始方法和分隔法的结果. 图中的横坐标为数据集名称的前两个字母,从左到右按样本重叠比例升序排列. 从图中可以发现,在绝大部分分类器和数据集上,分隔法比原始方法具有更好的分类效果. 而且随着样本重叠比例的增加,分隔法的优势通常更为明显. 比如在高重叠比例的 Inosphere 和 Glass 数据集上使用 NB 时,分隔法就显著优于原始方法. 而对于样本重叠比例较低的数据集,由于只有少数样本落入重叠区域,这时通过分隔法很难有效提高分类效果.

另一方面可以观察基础分类器选择对分隔法分类效果的影响. 实验中使用的分类器按照性质不同可以分为两类. C4.5 和 RIPPER 是基于规则的分类器,而 NB、k-NN 和 SVMs 是基于分界面的分类器. 表 4 给出了各分类器在不同数据集上的 10 折交叉验证结果,即以原始方法为基准进行成对比较的双边 t-检验绝对值. 由于 $n = 10$,显著性水平 $\alpha = 0.05$,因此 t-检验的拒绝域为 $|t| \geqslant t_{\alpha/2}(n-1) = 2.26$. 表中数据集从左到右按样本重叠比例升序排列,并使用数据集名称的前两个字母进行表示; 黑体数值表示 t-检验值大于 2.26,即显著优于原始方法. 从表中的结果可以看出,当使用基于分界面的分类器时,t-检验值在绝大多数情况下都大于 2.26,即分隔法表现更佳. 在重叠比例较高的 Sonar 和 Inosphere 数据集上,分类效果的提升尤为显著. 需要说明的是,在 Fourclass 数据集上,由于 k-NN 的平均 F 值已经超过 0.95,分隔法很难显著提高其分类绩效,此时 t-检验值仅为 0.38. 但对于基于规则的分类器,t-检验值在绝大多数情况下都小于 2.26,分隔法的优势不是那么明显且缺乏一致性. 如在 Sonar 上采用 RIPPER 时,分隔法甚至降低了 RIPPER 的分类绩效. 这主要是因为,分隔法中重叠区域的寻找为进一步精炼分类模型的分界面提供了帮助; 而样本向重叠区域和非重叠区域的分流,则会加大遗失一些重要规则的证据的可能性,从而降低了基于规则的分类器的分类绩效.

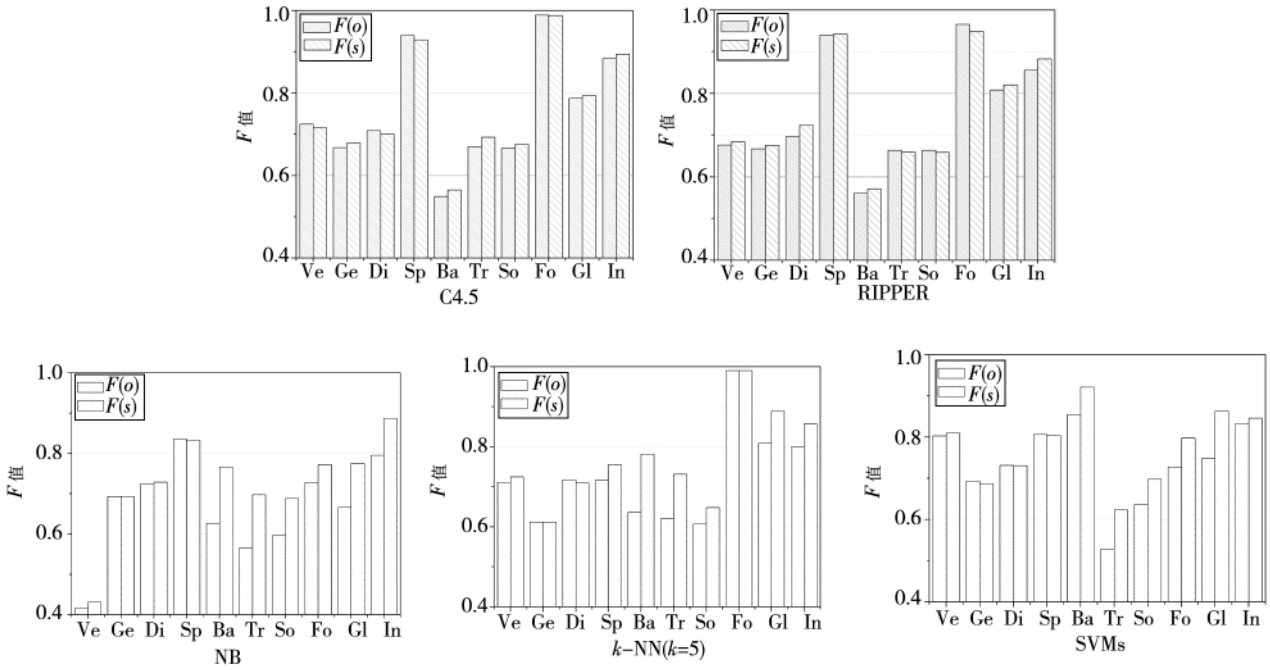


图 7　分隔法与原始方法的分类效果

Fig. 7 Performances of the separating and original schemes

表 4　实验结果的 t-检验值

Table 4 t-test values of the experiment results

分类算法	数据集									
	Ve	Ge	Di	Sp	Ba	Tr	So	Fo	Gl	In
C4.5	1.75	1.23	**2.66**	1.64	2.05	2.15	0.78	0.29	0.57	0.62
RIPPER	1.60	1.08	1.87	0.75	1.98	0.56	0.32	0.44	1.83	**2.47**
NB	2.13	0.59	0.81	**4.86**	**23.64**	**22.06**	**14.74**	**8.12**	**13.65**	**15.23**
k-NN	1.44	0.85	1.06	**5.73**	**24.36**	**19.61**	**10.11**	0.38	**17.42**	**13.77**
SVMs	1.92	1.67	0.48	1.23	**9.59**	**17.96**	**5.41**	**16.50**	**23.81**	**3.54**

2.3　实验 III: 分隔法用于类重叠及类不均衡共存问题

本实验关注分隔法及基于分界面的分类器在类重叠问题和类不均衡问题共存情况下的表现. 为了构造不均衡数据集，实验选择了类重叠比例大于 10% 的二分类数据集 Inosphere、Fourclass、Sonar、Transfusion 和 Splice，并通过随机抽样方法对二分类数据集中的少数类进行欠抽样 (under-sampling)，抽样比例分别为 20%，40%，60% 和 80%，每个抽样比例重复 10 次实验返回平均 F 值. 结果如图 8 所示，其中 "s" 表示分隔法结果，"o" 表示原始方法.

从图 8 的实验结果可以发现，在样本重叠比例较高的数据集上，类不均衡虽然同时降低了分隔法和原始方法的分类精度，但分隔法相对于原始方法的优势仍然显著存在. 这充分说明，类不均衡只是降低了基础分类器如 NB 的分类绩效，对分隔法在类重叠问题上的作用没有太大的负面影响. 特别地，对于分类器 SVMs 而言，随着类不均衡程度的增加，即少数类抽样比率的减少，采用分隔法的分类结果与原始方法的分类结果相比，优势越来越明显. 这说明，对于某些分类器而言，分隔法能够有效地减少类不均衡对分类器的负面影响. 下节针对 SVMs 的理论分析对这个现象将做进一步探讨.

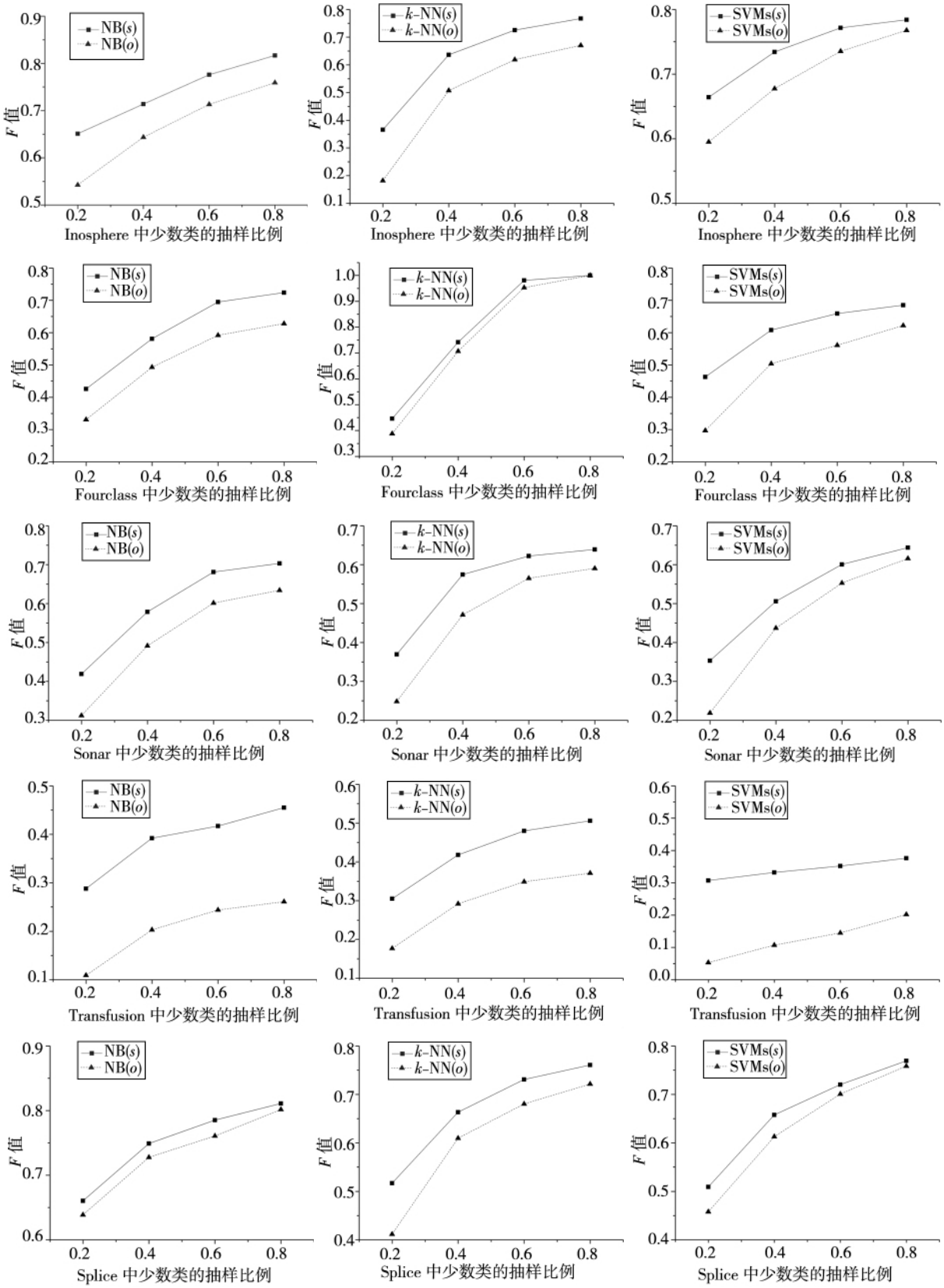


图 8　类不均衡对分隔法和原始方法的影响

Fig. 8 The effect of class imbalance on the separating and original schemes

3 基于 SVMs 的分析

下面就实验 III 的结果进行理论分析. 实验中的 SVMs 是使用线性核函数的 C-SVM,它通过寻找最大间隔超平面来划分数据. 给定包含两个类别的训练集 $X = \{x_i, i = 1, \cdots, l\}$,令 y_i 表示类别标号,且 $y_i \in \{-1, +1\}$,C-SVM 解决如下最优化问题

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i \\ &\text{s. t. } y_i((w \cdot x_i) + b) \geqslant 1 - \xi_i, \\ &\quad \xi_i \geqslant 0, i = 1, \cdots, l \end{aligned} \tag{5}$$

式(5) 的拉格朗日函数如下

$$\begin{aligned} L(w, b, \xi) = &\frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i - \\ &\sum_{i=1}^{l} \alpha_i (y_i((w \cdot x_i) + b) - 1 + \xi_i) - \sum_{i=1}^{l} \beta_i \xi_i \end{aligned} \tag{6}$$

其约束条件为

$$\alpha_i \geqslant 0, \beta_i \geqslant 0, i = 1, \cdots, l \tag{7}$$

将 L 分别对 R,a 和 ξ 求偏导,并令其为 0,可得

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l} \alpha_i y_i x_i = 0 \tag{8}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{9}$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \tag{10}$$

按照 y_i 的取值 +1 和 −1,可将样本划分为正类样本和负类样本,因此式(9) 可转换为

$$\sum_{i=1}^{l_p} \alpha_i = \sum_{j=1}^{l_n} \alpha_j = \Delta \tag{11}$$

其中 l_p 和 l_n 分别是正类和负类样本的数目. 令 $\Delta = C\Delta'$,式(11) 可继续转换为

$$\sum_{i=1}^{l_p} \alpha_i = \sum_{j=1}^{l_n} \alpha_j = C\Delta' \tag{12}$$

且由式(7) 和式(10) 有

$$0 \leqslant \alpha_i \leqslant C \tag{13}$$

根据 KKT 条件,样本数据可分为三类:

1) 如果 $\alpha_i = 0$,x_i 是非支持向量,它被正确划分,不影响最大间隔的确定;

2) 如果 $0 < \alpha_i < C$,x_i 是支持向量,它位于间隔区域的边缘上;

3) 如果 $\alpha_i = C$,x_i 也是支持向量,它被错误划分或者落入间隔区域.

第 3) 类样本造成了所谓的“间隔错误”[30]. 在利用 C-SVM 分类时,间隔错误严重与否在很大程度上决定了分类的准确性. 假设正类和负类样本中第 2) 类样本在式(12) 中的和为 A_p 和 A_n,第 3) 类样本数分别为 d_p 和 d_n,根据式(12) 有

$$\begin{aligned} &\sum_{i=1}^{l_p} \alpha_i = A_p + Cd_p = C\Delta' \Rightarrow d_p \leqslant \Delta', \\ &\sum_{i=1}^{l_n} \alpha_i = A_n + Cd_n = C\Delta' \Rightarrow d_n \leqslant \Delta' \end{aligned} \tag{14}$$

式(14) 表明,正类和负类的间隔错误样本的上界均为 Δ'. 这意味着在间隔错误最大的极端情况下,$d_p = d_n = \Delta'$. 令 v_p、v_n 分别为间隔错误样本在各类样本中所占比例的上界,则由 $d_p = v_p l_p$ 和 $d_n = v_n l_n$ 可以得到

$$v_p / v_n = l_n / l_p \tag{15}$$

在类不均衡问题中,稀有类通常记为正类,多数类被记为负类,正类数目远小于负类,即 $l_p << l_n$. 由式(15) 可得 $v_p >> v_n$,即正类的间隔错误样本比例的上界远大于负类,这意味着最大间隔划分超平面存在向正类偏移的趋势. 如果两类样本相隔较远,这种偏移不会影响 C-SVM 的分类效果. 反之,这种向正类的偏移则会降低稀有类的分类精度. 当同时存在类重叠问题时,这种偏移带来的影响将更为明显. 图 9 很好地展示了这一点. 如图所示,当“+”所代表的正类与“o”所代表的负类具有相同的样本量时,尽管有类重叠的影响,C-SVM 的最大间隔超平面 h1 仍在二类样本的中间位置. 然而,若只保留正类中的部分样本(如红色“+”所示) 与负类样本构成不均衡数据,C-SVM 的最大间隔超平面 h2 则由 h1 的位置明显向正类偏移,这将导致正类也就是稀有类的分类错误率上升.

图 10 则展示了为什么分隔法能够帮助提高 C-SVM 在不均衡及重叠数据上的分类绩效. 如图所示,对于图 9 中的不均衡数据,对非重叠样本和

重叠样本分别学习可以依次得到最大间隔超平面 h3 和 h4. 由于去除了重叠样本的影响,C-SVM 的间隔错误样本量大大减少,h3 的偏移量因此也较 h2 明显减少,更接近图9 中的 h1. 另外,由于重叠样本中正类负类样本的比例没有初始数据那么悬殊,h4 的位置通常也不会大幅偏离 h1. 正是在这两种效果的联合作用下,分隔法往往能提高 C-SVM 的分类绩效.

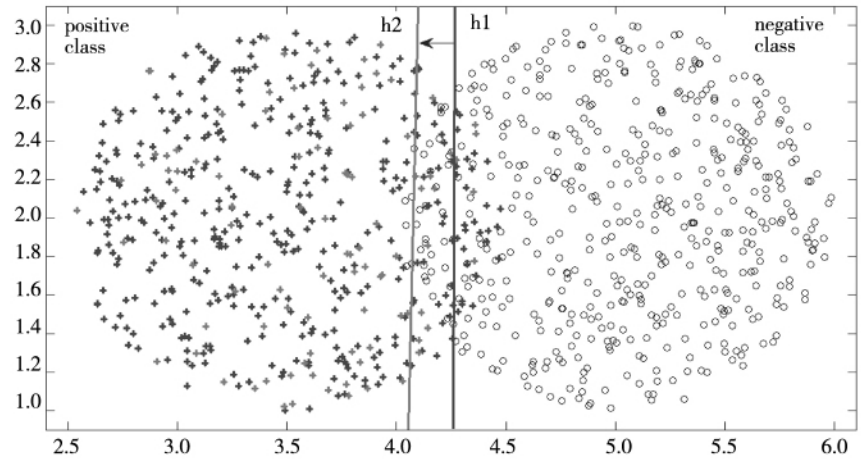


图 9　最大间隔超平面的偏移

Fig. 9 The deviation of maximal margin hyperplane

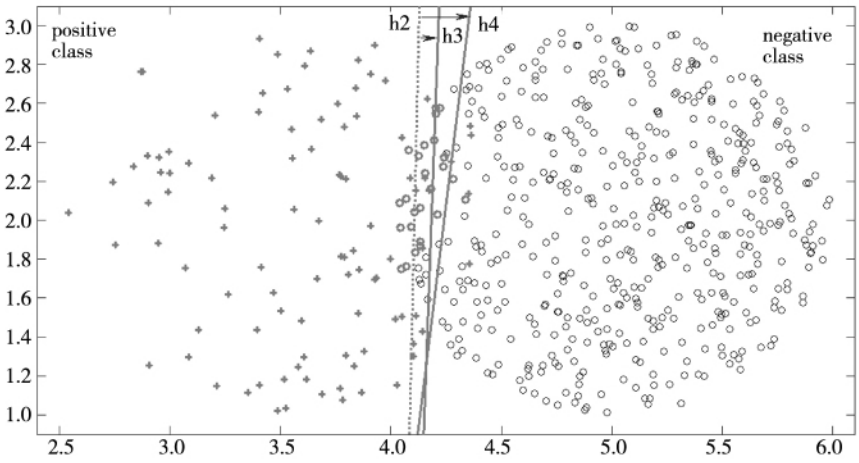


图 10　使用分隔法后最大间隔超平面的复位

Fig. 10 The reset of maximal margin hyperplane after using the separating scheme

4　结束语

本文对类重叠学习问题进行了系统研究. 首先利用 SVDD 算法寻找真实数据中的重叠区域. 然后通过在四种类重叠学习策略上进行的一系列比较实验,发现了本文提出的将重叠区域和非重叠区域分别进行学习的分隔法具有最佳分类效果. 实验进一步比较了五种基本分类器在分隔法中的表现,结果表明 SVMs 等基于分界面的分类器比基于规则的分类器对分隔法更为敏感. 实验还研究了类重叠问题和类不均衡问题的综合影响,结果表明当类不均衡程度加剧时,分隔法能够明显提高分类器尤其是 SVMs 的分类精度. 文末对此给出了初步的理论分析和探讨. 未来研究将继续关注重叠样本分类的理论分析,尝试从根本上揭示类重叠学习的基本规律.

参 考 文 献:

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

[2] 邹 鹏, 李一军, 郝媛媛. 基于代价敏感性学习的客户价值细分 [J]. 管理科学学报, 2009, 12(1): 48 – 56.
Zou Peng, Li Yijun, Hao Yuanyuan. Customer value segmentation based on cost-sensitive learning [J]. Journal of Management Sciences in China, 2009, 12(1): 48 – 56. (in Chinese)

[3] 夏国恩. 基于满意属性选择的客户流失预测 [J]. 管理学报, 2010, 7(6): 48 – 56.
Xia Guoen. Customer churn prediction on satisfactory attribute selection [J]. Chinese Journal of Management, 2010, 7(6): 48 – 56. (in Chinese)

[4] 杨海军, 太 雷. 基于模糊支持向量机的上市公司财务困境预测 [J]. 管理科学学报, 2009, 12(3): 102 – 110.
Yang Haijun, Tai Lei. Predicting financial distress of listed corporations based on fuzzy support vector machine [J]. Journal of Management Sciences in China, 2009, 12(3): 102 – 110. (in Chinese)

[5] 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.

[6] García V, Alejo R, Sánchez J S, et al. Combined effects of class imbalance and class overlap on instance-based classification [C]//In Intelligent Data Engineering and Automated Learning, 2006, 371 – 378.

[7] Jo T, Japkowicz N. Class imbalances versus small disjuncts [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 40 – 49.

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

[9] 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 Yeqing, Liu Sanyang, Gu Mingtao. Improved learning algorithm with transductive support vector machines [J]. Systems Engineering: Theory & Practice, 2009, 29(7): 113 – 118. (in Chinese)

[10] Prati R C, Batista G E, Monard M C. Class imbalance versus class overlapping: An analysis of a learning system behavior [C]//In Proceedings of the Mexican International Conference on Artificial Intelligence, 2005, 312 – 321.

[11] Visa S, Ralescu A. Learning from imbalanced and overlapped data using fuzzy sets [C]//In Proceedings of International Conference on Machine Learning 2003 workshop: learning with imbalanced data sets II, 2003, 97 – 104.

[12] 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.

[13] Tang Y, Gao J. Improved classification for problem involving overlapping patterns [J]. IEICE Transactions on Information and Systems E Series D, 2007, 90(11): 1787 – 1795.

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

[15] 熊海涛, 吴俊杰, 刘 鲁. LSVDD: 基于局部支持向量数据描述的稀有类分析算法 [J]. 系统工程理论与实践, 2012, 32(8): 1784 – 1792.
Xiong Haitao, Wu Junjie, Liu Lu. LSVDD: Rare class analysis based on local support vector data description [J]. Systems Engineering: Theory & Practice, 2012, 32(8): 1784 – 1792.

[16] Chen B, Ma L, Hu J. An improved multi-label classification method based on svm with delicate decision boundary [J]. International Journal of Innovative Computing, Information and Control, 2010, 6(4): 1605 – 1614.

[17] Lu B L, Ito M. Task decomposition and modular combination based on class relations: A modular neural network for pattern classification [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1244 – 1256.

[18] Podolak I T. Hierarchical classifier with overlapping class groups [J]. Expert Systems with Applications, 2008, 34(1): 673 – 682.

[19] Zhou J, Krzyzak A, Suen C Y. Verification-a method of enhancing the recognizers of isolated and touching handwritten numerals [J]. Pattern Recognition, 2002, 35(5): 1179 – 1189.

[20] Bellili A, Gilloux M, Gallinari P. An mlp-svm combination architecture for onine handwritten digit recognition: Reduction of recognition errors by support vector machines rejection mechanisms [J]. International Journal on Document Analysis and Recognition, 2003, 5(4): 244 - 252.
[21] Boutell M R, Luo J, Shen X, et al. Learning multilabel scene classification [J]. Pattern Recognition, 2004, 37(9): 1757 - 1771.
[22] Tsoumakas G, Katakis I. Multi-label classification: An overview [J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1 - 13.
[23] Kretzschmar R, Karayiannis N B, Eggimann F. Handling class overlap with variance-controlled neural networks [C] // Proceedings of the International Joint Conference on Neural Networks, 2003, 517 - 522.
[24] Jain A K, Duin R P W, Mao J. Statistical pattern recognition: A review [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4 - 37.
[25] 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.
[26] Blake C L, Merz C J. Uci repository of machine learning databases [EB/OL]. 2011, 01, 18. http://kdd.ics.uci.edu.
[27] Chang C C, Lin C J. Libsvm: A library for support vector machines [EB/OL]. 2011, 01, 18. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[28] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining [M]. Boston: Addison-Wesley Longman Publishing Co., Inc, 2005.
[29] He H, Garcia E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 20(2): 191 - 220.
[30] Scholkopf B, Smola A, Williamson R C, et al. New support vector algorithms [J]. Neural Computation, 2000, 12(5): 1207 - 1245.

Towards classification with class overlapping

XIONG Hai-tao[1,2], *WU Jun-jie*[2], *LIU Hong-pu*[2], *LIU Lu*[2]

1. School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;
2. School of Economics and Management, Beihang University, Beijing 100191, China

Abstract: Classification with class overlapping (CWCO) has long been regarded as one of the toughest yet pervasive problems in data mining and machine learning communities. When it is combined with the well-known class imbalance problem, the situation becomes even more complicated, and few works in the literature addresses this problem. To meet this critical challenge, in this paper, we make a systematic study on the CWCO problem and its interrelationship with the class imbalance problem. Specifically, we first introduce the support vector data description (SVDD) algorithm for capturing overlapping objects, and then introduce three learning schemes and propose a separating scheme for solving the CWCO problem. Extensive experiments on various real-world data sets using five different classifiers show that the separating scheme: 1) performs the best among the four schemes for CWCO, 2) is more suitable for classifiers using decision boundaries, and 3) performs well for class imbalance data, in particular with the support vector machines (SVMs). Finally, we provide theoretic explanations for the superior performance of the separating scheme using SVMs.

Key words: data mining; classification; class overlapping; class imbalance; support vector data description (SVDD)