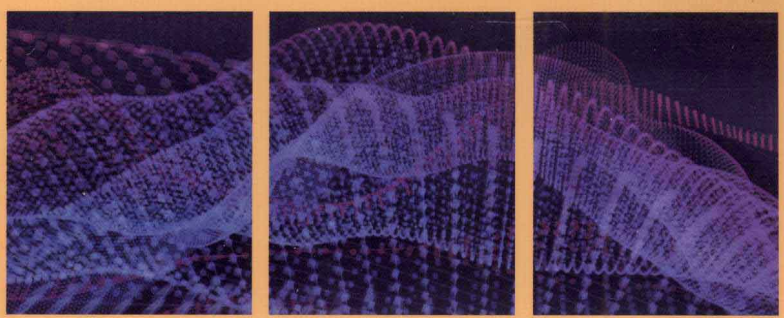
THE RESEARCH ON METHODS AND APPLICATIONS OF COMPLEX DATA ANALYSIS



复杂数据分析方法及其应用研究

熊海涛 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

复杂数据分析方法及其应用研究

熊海涛 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

本书主要从数据挖掘与商务智能的角度，系统地介绍了如何利用复杂数据分析的相关理论和方法来提升复杂事件的识别和预测的效果，同时还结合实际应用问题说明了复杂数据分析的应用过程。主要内容包括复杂数据分析方法综述、基于局部支持向量数据描述的复杂数据分析算法研究、类重叠问题及其处理方法研究、一致性分类方法研究和复杂概念分析应用研究等。

本书可供从事数据挖掘与商务智能研究和应用的科研人员及高等院校信息管理与信息系统专业、管理科学与工程等相关专业师生参考使用。

图书在版编目（CIP）数据

复杂数据分析方法及其应用研究／熊海涛著．—北京：北京理工大学出版社，2013.5

ISBN 978-7-5640-7678-8

Ⅰ．①复…　Ⅱ．①熊…　Ⅲ．①数据处理-研究　Ⅳ．①TP274

中国版本图书馆 CIP 数据核字（2013）第 078412 号

出版发行／北京理工大学出版社

社　　址／北京市海淀区中关村南大街 5 号

邮　　编／100081

电　　话／(010)68914775(办公室)　68944990(批销中心)　68911084(读者服务部)

网　　址／http：// www. bitpress. com. cn

经　　销／全国各地新华书店

印　　刷／北京泽宇印刷有限公司

开　　本／710 毫米×1000 毫米　1/16

印　　张／9

字　　数／116 千字　　责任编辑／廖宏欢

版　　次／2013 年 5 月第 1 版　2013 年 5 月第 1 次印刷　　责任校对／周瑞红

定　　价／28.00 元　　责任印制／王美丽

目　录

第一章 绪 论

1.1 背景介绍

自20世纪80年代末以来，无论在研究领域还是在实践领域，机器学习与知识发现（Machine Learning and Knowledge Discovery）都是最热的话题之一，它有力地推动了商务智能概念的普及与商务智能技术的发展。分类学习（Classification）是机器学习领域的核心子领域之一，其是通过学习得到一个目标函数f，把每个属性集x映射到一个预先定义的类别y，目的是使预测的类别和实际的类别之间的差别尽可能小。作为一类非常重要的数据分析技术，分类已经长时间在众多应用领域发挥着重要作用，如在信用卡欺诈识别、客户分析、企业财务困境预测、突发事件监测、邮件过滤等，并在在线拍卖投标者行为分析、消费者行为分析等领域具有美好的应用前景。

然而，传统的分类分析技术在今天面临着越来越多的挑战。首先，研究及实践人员发现，由于传统分类技术通常只返回一个分类结果，该结果往往囿于初始参数设置或者分类方法不够合理而为劣解，而且不同时候在同一数据上得到的分类结果往往差异很大。其次，随着信息系统在各商务实践领域的广泛渗透，商务数据已经进入了大数据（Big Data）时代，大数据具有三方面特质，即大量（Volume）、多

样（Variety）、实时（Velocity）。2012 年年初，在瑞士达沃斯举行的世界经济论坛上，“大数据，大影响”成为热议话题。这些大数据的出现给传统的分类技术带来了巨大的挑战。在有些实际应用中，传统分类分析技术变得完全不可行。例如：

● 高维数据：高维数据如文本数据、基因序列数据通常是稀疏的，充斥着无关属性，甚至噪声属性。传统分类算法通常在整个属性空间计算样本距离（子空间分类法除外），这样得到的结果很难保证质量。

● 海量数据：某些分类算法如神经网络、支持向量机等复杂度较高，在海量数据上运算效率很低，甚至内存资源不足以完成分类。

● 异构数据：样本数据可能来自不同的业务部门，记录了同一客户在不同业务中的行为，因数据类型差异较大而很难整合在一起；又或者这些数据物理上相距较远，通信成本较高，也无法整合在一起；再或者数据存在属性不同和类别标签缺失的情况，而收集这些属性和类别标签代价较高。这时传统集成学习算法就无法基于样本的全信息完成分类。

有鉴于此，本书聚焦于复杂数据学习研究，尝试构建一个复杂数据学习算法框架，该框架可以实现面向多样化效用函数和大数据的复杂学习，并通过系统原型应用于商务实践问题。具体而言，本项目首先把某些传统分类方法的高效性和多效用函数的适应性结合起来，在分类理论研究的基础上，建立复杂数据学习的理论基础和基本算法；然后系统地研究复杂数据学习的效用函数选择问题和基础分类分量的生成策略；最后通过增量和迭代过程开发可用于并行计算的系统原型，在商务实践领域的复杂特征数据上做深入的应用案例研究。

1.1.1 类不均衡问题

稀有事件永远是人们关注的焦点[1]。典型的稀有事件包括金融欺诈、网络入侵、上市公司财务危机、通信设备故障等。其特点在于正常情形下出现的概率很小，但一旦发生将会产生巨大的影响，如金融卡欺诈检测中，可能在 10 万起信用卡交易中只有 1 起交易是欺诈行为，但这 1 起交易可能会给当事人带来巨大的损失；反垃圾邮件处理中遇到的大部分邮件是有意义的邮件，垃圾邮件只是少数；故障监视中，正常实

例很容易获取，而异常实例就很难获得。以金融卡欺诈为例，英国支付联盟（APACS）的数据表明，2007 年全英金融卡欺诈损失率增长了 25%，达到惊人的 5.35 亿英镑。这在很大程度上应“归功于”“金融卡不在场”（card-not-present，CNP）的欺诈方式（以网上银行遭网络入侵为主），其导致的损失较 2006 年增长了 37%，为 2000 年的 398%。虽然目前还很难获得我国金融卡欺诈事件的确切数字，但从中国互联网信息中心（CNNIC）发布的统计数据来看，我国的网络支付发展十分迅速，2008 年使用规模已达 5 200 万人，年增长率为 57.6%[2]，这意味着 CNP 欺诈的潜在威胁是十分巨大的。事实上，欧洲安全与合作组织的专家斯特劳斯指出，每年网络犯罪引发的经济损失预计在 1 000 亿美元左右，其中还涉及非常严重的国家信息安全问题。在目前的研究热点突发事件应急管理方面，同样也存在稀有事件，即恶意恐慌信息。Web 信息资源是突发事件预警分析的重要依据。在网络化信息时代，除了制度化渠道的信息之外，非制度性渠道中的 Web 信息资源汇集了各类事件的新闻报道和舆论评价等信息，是突发事件的重要信息平台，对于突发事件应急管理与控制具有广泛应用，其中的恶意恐慌信息仅占少数，属于稀有事件范畴，它们能影响公众，给社会稳定带来影响。因此，针对稀有事件展开预测分析，是一项非常重要且非常紧迫的工作，具有很重要的现实意义。

数据挖掘中基于稀有事件和普通事件构成的样本数据进行预测分析，即所谓的“稀有类分析”（Rare Class Analysis）[3]。通常稀有类分析是与数据中各类样本数量的不均衡性（Class Imbalance）以及数据挖掘的分类（Classification）学习联系在一起的，被认为是分类的一个子领域[4]，但它首次被明确提出来却是近几年的事情，并迅速发展成为数据挖掘领域一个重要的研究方向[1]。我们通常将训练数据中的稀有事件称为稀有样本，将普通事件称为普通样本；将稀有事件构成的类称为稀有类，也常称为正类，将普通事件构成的类称为普通类，也常称为负类。

实践表明，在绝大多数情况下，稀有事件的预测难度要远远大于普通事件[5]，因此稀有类分析吸引了很多学者的研究兴趣。近年来，在实

际生活中，大量稀有类分析的应用被提出来，如：信用卡欺诈预测[6,7]、公司破产预测[8]、网络入侵检测[9]、通信设备失效预测[10]、通信风险管理[11]、生物信息学[12]、文本分类[13]、语音识别[14]、医疗预测[15,16]、石油泄漏图片检测[17]等。这些都是稀有类分析的研究领域。在诸多实际应用中，研究者们均指出了数据不均衡对分类学习带来的困难和挑战，其中最主要的方面就是分类器性能大大降低，尤其是稀有类识别能力的降低。事实上，由于稀有类分析如此重要，自2000年以来国际顶级学术会议如AAAI[18]、ICML[19,20]和ACM SIGKDD[21]上出现了为数不少的针对稀有类分析的专题研讨会，讨论稀有类分析问题，以及解决和提高稀有类预测的方法、算法等。毫无疑问，稀有类分析已成为数据挖掘与商务智能领域的热点研究方向。

分类问题是数据挖掘领域的重要研究内容之一，现有的一些分类算法都已经相对成熟，用它们来对均衡数据进行分类一般都能取得较好的分类性能。现有的分类器的设计都是基于类分布大致均衡这一假设的，通常假定用于训练的数据集中各类样本的数量是均衡的，即各类所含的样本数大致相当。但是这一假设在很多现实问题中是不成立的。现实情况中经常存在的是，数据集中于某个类别的样本数量可能会远远少于其他类别。而稀有类分析是基于不均衡数据来进行分析和预测的，具有一系列传统分类算法所没有考虑到的特点。所谓不均衡分类问题，是指训练样本数量在类间分布不均衡的分类问题。具体地说就是，某些类的样本数量远远少于其他类。具有少量样本的类为稀有类，而具有大量样本的类为普通类。物以稀为贵，稀有的信息，往往能获得人们更多的关注，稀有类的正确分类比普通类的正确分类更有价值。

前面的研究背景介绍表明，在许多实际的分类问题中存在着大量的稀有类，这些稀有类的识别具有重要意义。当传统的分类算法用于解决这些不均衡分类问题，即稀有类问题时，往往出现分类器性能的大幅度下降，得到的分类器具有很大的偏向性等问题。最常见的表现是稀有类的识别率远远低于大类，此时本属于稀有类的样本往往被错分到普通类，因此很难获得较好的分类效果。正是由于稀有类分析的这些特点引

发了一系列传统分类方法难以解决的问题，这些问题主要包括稀有类数量稀少问题、噪声问题、决策面偏移问题和评价指标问题四个方面。

1. 稀有类数量稀少问题

任何存在类不均衡分布的数据都可称为不均衡数据。根据这个类是相同的类还是不同的类可分为类间不均衡和类内不均衡，这是两种不同的不均衡，带来的问题也各不相同。

（1）类间不均衡。类间不均衡主要是指训练样本分布的不均衡，其容易导致稀有类样本的贫乏，具体地说，这种贫乏包括绝对贫乏和相对贫乏。绝对贫乏是指稀有类训练样本数量绝对过少，导致该类样本信息无法通过训练样本充分表示，即稀有类样本的数量太少以致无法满足基本的分类学习要求。文献［5］通过生成人工数据的实验指出，绝对贫乏的类的分类错误率要比一般类高出许多。对于绝对贫乏，文献［2］推荐采用过抽样方法，但同时也指出该方法具有较大的局限性。因此，很有必要针对能处理绝对贫乏的稀有类分析算法展开研究。首先，必须严格区分两类绝对贫乏：第一类是“真正的”绝对贫乏（True Absolute Rare，TAR），比如 SARS 这样的突发事件，其样本数量的确很少；第二类则是“虚假的”绝对贫乏（False Absolute Rare，FAR），比如信用卡样本中已经暴露的欺诈样本可能很少，但潜在的未暴露的欺诈样本可能较多，这些潜在的欺诈样本与正常样本混在一起未被标识出来。必须分别针对这两类不同的绝对贫乏展开研究。相对贫乏中稀有样本虽然显著少于普通样本，但其绝对数量仍能满足分类学习的基本要求。此时稀有类样本本身数量并不过少，只是相对于大类来说占有的比例过小。在这种情况下，基于启发式的贪心搜索方法效果将会变差[2]。文献［22］通过改变训练集的概念复杂度、样本不均衡度和训练集规模发现，当总样本数量足够多时，相对贫乏并不一定引起分类器性能下降。相反，绝对贫乏导致的稀有样本分布不集中且数量过少才容易引起分类器性能下降。

（2）类内不均衡。类内不均衡主要是指同一类数据的分布不均衡，因此该类数据具有比较复杂的数据结构。此时，如果该类数据过于稀

缺，则容易在特征空间中形成小的数据区域，从而引发小析取项（Small Disjuncts）问题。文献［23］通过30个实际数据集的测试结果表明，分类错误大部分集中在小析取项上。小析取项之所以有很高的分类错误率，其中很大的原因在于它和重叠数据以及噪声数据难以区分。然而许多分类算法为了防止过学习的产生，需要进行统计显著性检测，如决策树分类算法的剪枝，关联规则分类算法的规则筛选等，只有覆盖足够多样本的决策规则和关联规则才能被保留下来。小析取项的数据经常无法顺利通过这类显著性检测，但如果为了使它们通过检测而降低检测的阈值，又将无法有效地去除噪声。

2. 噪声问题

噪声问题是另外一个影响稀有类分析效果的重要因素。噪声分为属性噪声和类别噪声[24]。类别噪声有两种来源[25]：第一种是不一致样本，例如同一个样本出现在不同的类别里面；第二种就是错误标签，即对训练样本进行人工标注的时候出现误标，这种情况在大规模训练集的标注中经常出现。噪声数据的存在不可避免，并在一定程度上影响到分类器性能。在不均衡数据集中，噪声数据对稀有类的影响要远大于对普通类的影响，由于稀有类样本本来就很少，很多分类器无法区分特殊的稀有类和噪声。如果训练系统降低其通用性，将会导致把噪声数据也包含到训练中，因而必须引入防止过度拟合的技术（如剪枝），以便能够减少因噪声数据引入的小析取项，代价是一些真正的稀有类样本无法得到很好的训练[2]。只要在稀有类的决策域存在少数的噪声样本，就会影响该稀有类决策面的学习。也就是说，稀有类的抗噪能力较弱，容易与噪声数据混淆，并且分类器难以区分稀有类样本和噪声样本[5]。如果分类器采用一些防止过拟合的技术去除噪声，则会将一些稀有类样本信息一并去除。此时稀有类样本信息总量本来就比较少，这种去除严重影响了稀有类的信息含量。如果不去除噪声，分类性能也难以提高。因此噪声数据对大部分分类器的性能影响较大。

3. 决策面偏移问题

传统的分类方法大都建立在训练样本数量均衡的基础上。当将其用

于解决不均衡分类问题时，它们的分类性能往往有不同程度的下降，也就是稀有类导致了决策面的偏移。

基于特征空间决策面进行类别划分的分类器，如支持向量机，目标在于寻找一个最优的决策面。为了降低噪声数据的影响和防止过学习的产生，最优决策面必须兼顾训练分类准确率和决策面的复杂度，即采用结构风险最小化原则。然而，如果训练集不均衡，则支持向量的个数也不均衡。在结构风险最小化原则下，支持向量机会忽略稀有类中少量支持向量对结构风险的影响，而扩大决策边界，最终导致训练的实际超平面与最优超平面不一致。

基于概率估计的分类器，如贝叶斯分类器，分类准确率依赖于概率分布的准确估计。当稀有类样本过少时，概率估计的准确率将远小于普通类，稀有类的识别率也因此下降。基于规则的分类器，如决策树和关联规则分类器，需要对规则进行筛选。其中，支持度和可信度是规则筛选的重要指标。但是，当训练集不均衡时，基于上述指标的筛选会变得困难且不合理[9]。

4. 评价指标问题

分类器评测指标的科学性直接影响着分类器的性能，因为分类器训练的目标是实现最高的评测指标，总体目标的优劣直接决定了分类算法优化的过程。目前，多数传统的分类方法都是建立在训练样本集均衡的假设之上的，即用分类错误率来评价其分类性能。这些分类方法一般是以准确率作为分类器的评测指标的。然而在训练样本集呈现不均衡分布时，以准确率作为评测指标的分类器倾向于降低稀有类的分类效果[26]，而且准确率不重视稀有类对分类性能评测的影响。例如，假设有一个训练样本数量为1:99的两类问题，考虑一种极端情况下的分类效果，即分类器将所有样本都预测为普通类，此时分类器仍可以得到99%的训练准确率。由于稀有类的存在及其重要性，稀有类分析更关心稀有类样本的分类准确率。因此稀有类分析的评价指标不能简单地采用一般分类的评价指标。

1.1.2 类重叠问题

分类方法是机器学习领域的重要研究内容之一，现有的一些分类方

法都已经相对成熟，它们都是基于各类的样本数目大致均衡这一假设，即各类所含的样本数大致相当，并以分类精度作为评价目标，因此用它们来对均衡数据进行分类一般都能取得较好的分类效果。而将其应用于稀有类分析时，则倾向于忽视稀有类数据，因此需要针对稀有类分析的特点进行研究。尽管稀有类分析在实践领域如此重要，并且吸引了众多专家、学者的研究兴趣。但不可否认，近年来稀有类分析技术的研究进展略显缓慢。

随着研究的深入，陆续有学者提出数据集的不平衡问题并不是影响分类效果的关键因素，而类重叠现象的存在才是导致分类精度不高的主要原因。数据重叠问题逐渐成为新的研究热点，并受到越来越多研究者的关注。已有的研究大都关注已有分类器在类重叠时的表现，尚未提出有效的针对不均衡分类的重叠问题的解决策略。本书就是针对这个问题进行研究的。

本书将对不均衡分类的重叠问题进行系统研究。本书提出了存在类重叠时的不均衡分类的学习策略，通过系统比较来研究类重叠和样本量不均衡对分类效果的影响，并从剔除类别噪声入手，对数据进行预处理，提高了不平衡数据的整体分类效果。本书的研究具有显著的理论意义和现实意义，有助于弥补稀有类分析在存在类重叠时的一些空白。

1.1.3 集成学习问题

一个直观的想法是，可以在一定范围内设置初始参数，并多次分类返回多个分类结果，最后从中挑选较优者，这正是集成学习（Ensemble Learning）的基本思想[27]。文献［28］指出传统集成学习方法对于各分类器的结果往往通过融合方法（Fusions Methods）或者选择某个分类器（Selecting Classification）的结果获得最终分类结果。因此传统集成学习方法缺乏全局考虑，它只能在一定程度上处理复杂数据问题，还无法解决诸如样本不一致、结构复杂、样本重叠、类别标签缺失等问题。然而，真实的复杂数据往往具有复杂的概念模型。概念模型是指训练数据在属性空间的分布情况。数据中的复杂概念（Complex Concept，也称复杂的数据固有结构）正是其中隐含的且亟待解决的难点问题。如类不均衡问题，即不同类别样本之间

数量相差较大[29]；同时由于数据采集的客观条件，样本会不一致，并且不同类别的样本往往在某些属性上具有相似的取值而产生重叠[30]；另外，在积累的过程中往往不可避免地包含噪声和类别缺失进而影响集成学习模型的精确性[31,32]。因此在处理复杂数据的实际应用中，传统集成学习方法便变得不可行[33,34,28]。

组合集成学习（Combined Ensemble Learning）应运而生。组合集成学习隶属于集成学习方法，但是其目的是通过融合来自多个分类的结果而达到效用最大化，以得到更高质量和更好鲁棒性的一个分类结果。组合集成学习通常可被形式化为如式（1.1）所示的 NP 完全（NP-complete）问题：

$$\max_{\pi} \sum_{i=1}^{r} w_i U(\pi, \pi_i) \tag{1.1}$$

式中：π——组合集成学习结果；

π_i——第 i 个基础分类分量；

w_i——权重；

U——效用函数。

研究表明，组合集成学习通常可以获得比传统集成学习方法更好的分类结果[35]。因此，组合集成学习迅速成为了机器学习中分类领域的热点问题，大量算法如权重法、投票法、随机森林法、基于遗传算法的分类器集成法等被纷纷提出。尽管如此，组合集成学习尚有大量的问题亟待解决，首先是组合集成学习的理论体系问题，这也是组合集成学习的基础。目前组合集成学习研究文献普遍存在重组合、轻理论体系的特点，例如对组合集成学习的理论基础缺乏研究，没有对组合集成学习的效用函数选择问题展开研究，也没有对大数据问题进行理论层面的思考。此外，已有文献也没有系统研究基础分类分量的生成、评价、选择和加权策略，成熟的组合集成学习应用系统及应用案例仍极为罕见。

1.2 相关研究分析

1.2.1 复杂数据研究分析

复杂数据分析理论目前仍停留在对样本复杂性的研究上。研究的

思路主要有两个：一个是从具体分类算法的归纳偏置（Inductive Bias）入手来探讨样本复杂性的影响[35,36]；另一个则尝试从独立于算法的角度来研究样本复杂性的影响。复杂数据分析算法的一些重要方法则主要包括重抽样、成本敏感学习、集成学习方法、划分方法、调整归纳偏置、单类学习以及特征选择方法等。这些算法主要针对复杂数据分析的特点进行处理，从而达到提高复杂数据分析的效果的目的。复杂数据分析的评价指标可用于引导建模的搜索过程以及最终分类结果的评价。由于稀有类的存在及其重要性，复杂数据分析的评价指标不能简单地采用一般分类的评价指标。

通过以上对复杂数据分析问题的介绍可以看到，从整体上来说，目前对复杂数据分析的研究多为将复杂数据通过一定的处理转换为较为简单的数据再进行处理，或者改进已有的适用于简单数据分类的算法来进行分类，许多有针对性的解决方法在近年来被陆续提出，但对于复杂数据分析中的样本复杂性如何对分类算法产生影响，我们仍然缺乏清晰和系统的认识，针对复杂数据分析效果不佳的深层次原因的研究极少。事实上，对于某些分类器如支持向量机，在普通样本和稀有样本显著线性可分的情况下，样本的复杂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。

具体来说，现有的研究存在以下几种问题。

① 数据固有结构，或者说是数据的复杂概念对复杂数据分析的影响是普遍存在的，但是以往的复杂数据分析理论研究对此考虑得并不充分，将数据比例的不均衡认为是影响分类效果的原因，但是研究表明，通常是数据的复杂固有结构对复杂数据分析产生影响。那么这种复杂数据结构是如何对复杂数据分析问题产生影响的呢？我们有必要对其作用机理进行深入研究。

② 目前复杂数据分析算法研究主要从归纳偏置和样本的分布均衡两个角度来考虑问题，其关注的核心为各类数据的复杂结构问题，缺乏对数据整体概念的全局考虑，对数据固有结构带来的问题研究很少。传统分类方法则期望达到一种均衡，没有偏向稀有类，如何提高其稀有类识

别精度是一个需要解决的问题。单类学习方法、聚类方法和集成学习方法由于其特点能够适应数据固有结构的变化，从一定层面揭示数据的固有结构，但其应用于复杂数据分析时容易将一些普通类划分到稀有类中，如何准确地识别普通类也是一个需要解决的问题。目前还没有将数据的固有结构同分类方法相结合，从而有效提高稀有类预测能力的研究。

③ 现有的某些点评价指标应用于复杂数据分析中时，没有将稀有类和普通类分别对待，容易受参数影响（如训练样本中的两类样本比例），故存在问题。基于曲线的评价指标存在着分类错误成本难于确定和对不同的分类器采取不同的标准获得的评价指标无法准确地表明算法的优劣性等问题。

④在复杂数据分析甚至整个数据挖掘领域的研究中，缺乏完整的、丰富的案例研究始终是一个非常突出的问题。由于复杂数据分析具有很强的实践背景，并且具有很强的现实意义，因此有必要在实践环节展开案例研究，如 Web 突发事件应急管理中的恶意恐慌信息、C2C 电子商务欺诈识别。

1.2.2 类重叠问题研究分析

随着研究的深入，人们意识到样本的不均衡性的确会给一些分类学习的归纳过程带来影响，但这并不是导致不均衡分类如此困难的全部原因。另一个不容忽视的问题是类重叠的影响。重叠数据的分类已经成为数据挖掘和机器学习中的难点之一。错分现象经常发生在类边界上，这也是经常出现重叠的地方[37]。目前，很难找到一个解决重叠部分分类的好的方法[38]。现有研究大都是关注已有分类器在类重叠时的分类效果，尚未找到一个有效的针对不均衡分类的重叠问题的解决策略。且已有研究的实验大都建立在人工数据集上，尚未对真实数据集进行验证。

Prati[39]等人于 2004 年对类不平衡和类重叠（Class Overlapping）进行了比较研究，他们指出，分类器性能的下降不能只归咎于类不平衡因素的存在，在一些类不平衡比较严重的分类学习中，分类器仍然具有良好的性能，这是类重叠并不严重的结果。2005 年，Prati[40]等人通过在人工数据集上进行实验，指出数据重叠程度对于不均衡分类的效果有很