不平衡分类问题研究综述

叶志飞[1], 文益民[2], 吕宝粮[1,3]

(1. 上海交通大学 计算机科学与工程系, 上海 200240; 2. 湖南工业职业技术学院 信息工程系, 湖南 长沙 410208;
3. 上海交通大学 智能计算与智能系统教育部微软重点实验室, 上海 200240)

摘 要: 实际的分类问题往往都是不平衡分类问题, 采用传统的分类方法, 难以得到满意的分类效果. 为此, 十多年来, 人们相继提出了各种解决方案. 对国内外不平衡分类问题的研究做了比较详细地综述, 讨论了数据不平衡性引发的问题, 介绍了目前几种主要的解决方案. 通过仿真实验, 比较了具有代表性的重采样法、代价敏感学习、训练集划分以及分类器集成在 3个实际的不平衡数据集上的分类性能, 发现训练集划分和分类器集成方法能较好地处理不平衡数据集, 给出了针对不平衡分类问题的分类器评测指标和将来的工作.

关键词: 机器学习; 不平衡模式分类; 重采样; 代价敏感学习; 训练集划分; 分类器集成; 分类器性能评测
中图分类号: TP181 文献标识码: A 文章编号: 1673-4785(2009)02-0148-09

A survey of imbalanced pattern classification problems

YE Zhi-fei[1], WEN Yi-min[2], LU Bao-liang[1,3]

(1. Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; 2. Department of Information Engineering, Hunan Industry Polytechnic, Changsha 410208, China; 3. MOE-Microsoft Key Lab. for Intelligent Computing and Intelligent Systems, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Imbalanced data sets have always been regarded as presenting significant difficulties when applying machine learning methods to real-world pattern classification problems. Although various approaches have been proposed during the past decade, limitations are imposed by many real-world imbalanced data sets, and as a result a lot of further research is currently being done. In this paper, we provide an up-to-date survey of research on imbalanced pattern classification problems. We first took a deep look into the problems that imbalanced data sets bring, and then we introduced different kinds of solutions in detail with their representative approaches. Finally, using three real imbalanced data sets, we compared the performance of some typical methods including re-sampling, cost sensitive learning, training set partitions, and the performance of classifier ensembles. In addition, topics such as evaluation indexes and future areas of research were also discussed.

Keywords: machine learning; imbalanced pattern classification; re-sampling; cost sensitive learning; task decomposition; classifier ensemble; evaluation matrices

所谓不平衡分类问题, 是指训练样本数量在类间分布不平衡的模式分类问题. 具体地说就是某些类的样本数量远远少于其他类. 本文称具有少量样本的那些类为稀有类, 而具有大量样本的那些类为大类. 物以稀为贵, 稀有的信息, 往往能获得人们更多的关注. 在许多实际的模式分类问题中, 同样存在稀有的类, 它们虽然很重要, 但是用传统的分类方法, 却难以被正确分类. 当传统的机器学习方法用于解决这些不平衡分类问题时, 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. 最常见的表现是稀有类的识别率远远低于大类. 因此, 本属于稀有类的样本往往被错分到大类.

在实际应用中, 不平衡问题很常见. 有些问题其原始数据的分布就存在不平衡, 如通过卫星雷达图片检测海面石油油污[1]、监测信用卡非法交易[2]、

收稿日期: 2008-04-23
基金项目: 国家自然科学基金资助项目 (60375022, 60473040).
通信作者: 吕宝粮. E-mail: blu@cs.sjtu.edu.cn

发掘基因序列中编码信息以及医学数据分类[3-4]等. 这些问题都以稀有类的信息为关注的重点, 如在信用卡非法交易记录的监测问题中, 非法交易记录是监测的目标. 但训练数据中包含大量正常的信用卡交易记录, 只有很少的一部分是非法交易记录, 使用一般的模式分类方法, 非法交易记录的检测率很低. 有些不平衡分类问题源自数据收集过程中人为地造成的不平衡, 比如由于个人隐私或者高昂的数据采集代价等因素导致某些类的样本数量过少. 还有些不平衡问题来自多类(multi-class)问题和多标号(multi-label)问题的分解. 有些分类器如支持向量机(SVM), 无法直接解决多类多标号问题, 必须将原始问题分解成一系列两类问题解决, 这样很容易导致原来平衡的问题变得不平衡, 原来不平衡的问题变得更加不平衡.

不平衡分类问题到底有什么样的特征? 它们为何会导致分类器性能下降? 有什么有效的方法可以解决不平衡分类问题? 通过分析各种不平衡分类问题和已有的解决方案, 希望对上述问题做出回答. 内容安排如下: 第 1节讨论不平衡问题的特征及问题点; 第 2节介绍目前已有的解决不平衡问题的主要方法; 第 3节讨论针对不平衡问题的分类器评价指标; 第 4节通过仿真实验比较主要的几种分类方法在一些不平衡数据上的分类性能; 第 5节总结并讨论未来的研究方向.

1　不平衡分类问题

不平衡分类问题具有一系列传统模式分类方法所没有考虑到的特点, 从而引发了一系列传统模式分类方法难以解决的问题.

1.1　数据稀缺问题

样本分布的不平衡容易导致稀有类样本的稀缺, 具体地说, 稀缺包括绝对稀缺和相对稀缺.

绝对稀缺是指稀有类训练样本数量绝对过少, 导致该类信息无法通过训练样本充分表示. Weiss[5]通过生成人工数据的实验指出, 绝对数据稀缺类的分类错误率要比一般类高出许多. 此外, 当某类数据过于稀缺时, 容易在特征空间中形成小的数据区域, 从而引发小区块(small disjuncts)问题. Weiss和 Hirsh[6]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区块上.

小区块之所以有很高的分类错误率, 其中很大的原因在于它和噪声数据块难以区分. 许多分类器为了防止过学习的产生, 需要进行统计显著性(statistical significance)检测, 如决策树分类器的剪枝, 关联规则分类器的规则筛选等, 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. 小区块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.

相对稀缺是指稀有类样本本身数量并不过少, 但相对大类, 占有的比例过小. 在这种情况下, 稀有类样本的识别好比大海捞针, 使得基于启发式的贪心搜索方法效果变差[7]. Japkowicz和 Stephen[8]通过改变训练集的概念复杂度、样本不平衡度和训练集规模发现, 当总样本数量足够多时, 相对稀缺并不一定引起分类器性能下降. 相反, 绝对稀缺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.

1.2　噪声问题

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. 但是, 对不平衡分类问题, 噪声数据对稀有类将产生更大的影响[7]. 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, 并且分类器难以区分稀有类样本和噪声数据[5]. 如果分类器采用一些防止过学习的技术去除噪声, 则会将一些稀有类样本信息一并去除. 如果不去除噪声, 分类性能也难以提高.

1.3　决策面偏移问题

传统的模式分类方法, 大都建立在训练样本数量均衡的前提下. 当用于解决不平衡分类问题时, 它们的分类性能往往有不同程度的下降.

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(accuracy)和决策面的复杂度, 即采用结构风险最小化原则. 然而, 如果训练集不平衡, 则支持向量的个数也不平衡. 在结构风险最小化原则下, 支持向量机会忽略稀有类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大类,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(support)和可信度(confidence)是规则筛选的重要指标.但是,当训练集不平衡时,基于上述指标的筛选变得困难且不合理[9].

1.4 评测指标问题

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.传统的模式分类方法一般以准确率作为分类器评测指标.但是以准确率为评测指标的分类器倾向于降低稀有类的分类效果[10-11].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,即使分类器将所有样本分到大类,它仍可以得到99%的训练准确率.

2 不平衡分类问题的解决策略

迄今为止,解决不平衡分类问题的策略可以分为两大类.一类是从训练集入手,通过改变训练集样本分布,降低不平衡程度.另一类是从学习算法入手,根据算法在解决不平衡问题时的缺陷,适当地修改算法使之适应不平衡分类问题.平衡训练集的方法主要有训练集重采样(re-sampling)方法和训练集划分方法.学习算法层面的策略包括分类器集成、代价敏感学习和特征选择方法等.

2.1 重采样方法

重采样方法是通过增加稀有类训练样本数的上采样(up-sampling)和减少大类样本数的下采样(down-sampling)使不平衡的样本分布变得比较平衡,从而提高分类器对稀有类的识别率.

最原始的上采样方法是复制稀有类的样本,但是这样做容易导致过学习[12],并且对提高稀有类识别率没有太大帮助[13].较高级的上采样方法则采用一些启发式技巧,有选择地复制稀有类样本,或者生成新的稀有类样本. Chawla等人[14]提出的SMOTE算法是一种简单有效的上采样方法,该方法首先为每个稀有类样本随机选出几个邻近样本,并且在该样本与这些邻近的样本的连线上随机取点,生成无重复的新的稀有类样本. Lee[15]则通过为训练集中的稀有类样本加上随机噪声的方式获取新的正类样本. Kuba等人[16]将稀有类和大类交叉分布区域内的样本重新标定成稀有类样本,以降低不平衡度.

另一方面,下采样通过舍弃部分大类样本的方法,降低不平衡程度. Kubat和Matwin[17]采用单边采样方式,去除大类中的噪音样本、边界样本和冗余样本. Chen等人[18]则通过修剪大类的支持向量,达到平衡支持向量个数的目的,从而提高稀有类的识别率. Raskutti和Kowalczyk[19]同时考虑上采样和下采样,并且扩展到一类学习,即只采用一类的样本作为训练集,因此不存在不平衡分类问题. Estabrooks和Japkowicz[20]同时采用上采样和下采样以及不同的采样率,获得大量的子分类器,并使用混合专家(mixture-of-experts)学习框架将这些子分类器集成.他们的结果显示,这种方法比普通的AdaBoost有更好的分类效果,但并不清楚到底是上采样还是下采样更有效,也不清楚哪种采样率最合适.

虽然重采样在一些数据集上取得了不错的效果,但是这类方法也存在一些缺陷.上采样方法并不增加任何新的数据,只是重复一些样本或增加一些人工生成的稀有类样本,增加了训练时间.更危险的是,上采样复制某些稀有类样本,或者在它周围生成新的稀有类样本,使得分类器过分注重这些样本,导致过学习[12,14].上采样不能从本质上解决稀有类样本的稀缺性和数据表示的不充分性,因此有人指出它的性能不如下采样[12].但是Japkowicz[8]对人工数据的一项系统研究得到了相反的结论.下采样在去除大类样本的时候,容易去除重要的样本信息.虽然有些启发式的下采样方法,只是去除冗余样本和噪声样本,但是多数情况下这类样本只是小部分,因此这种方法能够调整的不平衡度相当有限.

2.2 训练集划分方法

对训练数据集进行划分,是另一种有效的训练集平衡方法. Chan和Stolfo[21]首先根据代价敏感学习的需要,学习一个合理的类别样本分布比例.然后将大类样本随机划分成一系列不相交子集.这些子集的大小由稀有类样本集的数量和预先学习的样本分布比例决定.接下来分别将这些不相交子集跟稀有类样本结合,组成一系列平衡的分类子问题,单独训练成子分类器.最后通过元学习(meta-learning)将这些子分类器的输出进一步学习成组合分类器.这种方法在信用卡非法使用检测问题上大大降低了总代价. Yan等人[21]采用类似的问题分解方式,并将每个子问题用SVM独立训练后采用分类器集成,

得到的组合分类器的性能超过了上采样和下采样方法. 上述训练集划分方法仅考虑了划分后子训练集的规模和分布, 没有对划分规则作进一步考虑. Lu和 Ito[22]提出了最小最大模块化(min-max modular)神经网络模型, 该模型利用最小最大化集成规则, 能有效地将子分类器组合, 使组合分类器容易地实现并列学习和增量学习. 之后 Lu等人[23]将上述模型推广到支持向量机并提出了"部分对部分"(part vs. part)任务分解策略. "部分对部分"任务分解策略可对不平衡两类子问题作进一步分解. 这种分解策略可以自由地控制每个子问题的规模和平衡度, 并且可以根据先验知识和训练集样本的分布特征, 制定有效的分解规则. 实验表明, 该方法比代价敏感学习和重采样方法能更好地解决不平衡问题[24].

2.3 分类器集成方法

上述通过训练集划分得到的子分类器, 利用分类器集成的方法获得了良好的效果. Kotsiantis和 Pintelas[25]将训练集重采样后用 3种学习方法分别训练, 然后将得到的分类器采用多数投票方法给出预测类别. 实验表明, 他们的方法能提高对稀有类样本的识别率. Estabrook等人[26]通过计算发现, 根据训练集的自然分布得到的分类器不一定具有最好的一般化能力. 他们提出通过对原不平衡问题进行重采样, 从而构建多个平衡度不同的训练集, 训练后采用分类器挑选和偏向正类的原则将各个分类器综合. 仿真实验表明, 该方法比单独应用上采样和下采样方法获得了更好的准确率和 ROC(receiver operating characteristic)曲线. Chen等人[27]提出了平衡随机森林的方法, 该方法对正类和反类分别进行重采样, 重采样多次后采用多数投票的方法进行集成学习. Chawla等人[28]将 boosting算法与 SMOTE算法结合成 SMOTEBoost算法, 该算法每次迭代使用 SMOTE生成新的样本, 取代原有 AdaBoost算法中队样本权值的调整, 使得 Boosting算法专注于正类中的难分样本. Liu等人[29]基于人脸识别的级联模型提出了一种基于级联模型的不平衡数据分类方法, 该方法通过逐步筛掉反类样本, 使得级联结构中后面的结点得到更为平衡的训练集. Zhou和 Liu[30]提出了代价敏感神经网络与分类器集成相结合的方法, 他们通过 21个 UCI标准数据集的实验发现, 分类器集成不仅对处理 2类不平衡问题有效, 而且对多类不平衡问题同样有效.

2.4 代价敏感学习方法

在大部分不平衡分类问题中, 稀有类是分类的重点. 在这种情况下, 正确识别出稀有类的样本比识别大类的样本更有价值. 反过来说, 错分稀有类的样本需要付出更大的代价. 代价敏感学习[31]赋予各个类别不同的错分代价, 它能很好地解决不平衡分类问题. 以两类问题为例, 假设正类是稀有类, 并具有更高的错分代价, 则分类器在训练时, 会对错分正类样本做更大的惩罚, 迫使最终分类器对正类样本有更高的识别率.

Domingos[32]提出了一种 Metacost方法, 该方法通过估计训练样本的后验概率密度, 结合代价矩阵(cost matrix)计算每个训练样本的理想类别, 然后根据理想类别修改原训练样本的类别, 得到新的训练集, 最后使用基于错误率的分类器学习这个新的训练集. 仿真实验表明, Metacost比下采样和上采样方法能获得更低的错误代价. Metacost的重要意义在于它能将普通的基于准确率的学习方法容易地改造成对错分代价敏感的学习方法. Chen[27]在平衡随机森林的基础上提出了带权随机森林算法, 该方法赋予每个类一个权值, 训练样本最少的类赋予的权值最大. 在构造决策树的过程中引入权值, 每一棵决策树的决策采用带权多数投票. 最后所有的决策树采用带权投票集成. Che等人[33]通过训练集先验信息的分析, 利用支持向量机为不同类的样本设置惩罚系数.

给不同的训练样本赋予不同的权值也能起到代价敏感学习的作用. Fan等人[34]提出了一种 AdaCost算法, 该算法通过在 Boosting算法的权值更新规则中引入每个训练样本的错分代价, 提高了 Boosting算法对稀有类的查全率和查准率. 该算法的权值更新原则是: 如果错分代价较大的样本被弱分类器错分, 则它对应的权值被"较大"地增加. 如果它被正确分类, 则它对应的权值被"较小"程度地减少. Josh等人[35]通过研究发现, 如果 AdaCost算法中的基分类器能获得较平衡的查全率和查准率, 则 AdaCost能获得对稀有类较平衡的查全率和查准率.

代价敏感学习能有效地提高稀有类的识别率. 但问题是, 一方面, 在多数情况下, 真实的错分代价很难被准确地估计[36]. 另一方面, 虽然许多分类器可以直接引入代价敏感学习机制, 如支持向量机和决策树, 但是也有一些分类器不能直接使用代价敏

感学习,只能通过调整正负样本比例或者决策阈值间接地实现代价敏感学习[37],这样不能保证代价敏感学习的效果.

2.5 特征选择方法

特征选择方法对于不平衡分类问题同样具有重要意义.样本数量分布很不平衡时,特征的分布同样会不平衡.尤其在文本分类问题中,在大类中经常出现的特征,也许在稀有类中根本不出现.因此,根据不平衡分类问题的特点,选取最具有区分能力的特征,有利于提高稀有类的识别率.

通过采用特征选择来解决不平衡分类问题主要集中于自然语言处理领域. Cardie和 Howe[38]以基于事例学习(case based learning)的框架为基础,提出了一种与测试样本相关的动态特征加权方法.该方法首先利用训练集得到一棵决策树,然后计算每个测试样本在测试路径上的信息收益,并以此计算每个特征的权值,最后,从训练集中挑选 k个与测试样本最接近的样本,并对他们测试类别进行投票.该方法在提高正类样本准确率的同时确保了总的准确率不下降. Zheng和 Srihari[39]针对文本分类中存在的不平衡分类问题,按照一个经验性的样本比例,挑选正负 2个样本集,分别从中选择最能表示该类样本的特征集,然后将这些特征集合并作为最后挑选的特征.对不同规模的特征集进行特征挑选的仿真实验表明,该特征挑选方法能有效提高文本分类的 F1测度.

2.6 其他方法

Wu和 Chang[40]提出了一种修改支持向量机核函数矩阵(kernel matrix)方法,该方法通过将核函数矩阵进行保角变换(conformal transformation),扩大稀有类特征向量处的边界,从而增加正负类样本的分离度,减少大类的支持向量数目,起到降低不平衡度的效果.理论分析和仿真试验结果表明,该方法在一些不平衡数据集上有比较好的效果.

Arunasalam和 Chawla[9]提出了一种自上而下的基于联合规则的分类器.他们指出,在处理不平衡分类问题时,传统的支持度和可信度在筛选关联规则时存在缺陷,并提出了补集类支持度(complement class support)作为挑选关联规则的重要指标.通过自上而下地将筛选出的最佳关联规则逐一添加到决策树中,形成最终的分类器.实验结果表明,该方法在不平衡数据上比传统的基于联合规则的分类器对稀有类有更高的识别率.

Hong等人[41]在 ROC曲线下面积 AUC(area under curve)指标的基础上,定义了 LOO-AUC(leave-one-out area under curve). LOO-AUC借鉴交叉验证的方法,每次移除一个样本,利用剩余样本训练的分类器预测该样本.在核分类器模型采用正交形式表示的基础上, Hong等人利用前向回归的更新规则,实现了 LOO-AUC的快速计算.他们提出了一种新的分类器最佳参数估计方法:正规正交带权最小方差估计,并以最大化 LOO-AUC作为模型选择标准,实现了正交前向模型选择.实验表明,该方法在生成数据和实际数据集上都能很好地处理不平衡问题.

一类学习(one-class learning)[42]也被用于处理不平衡问题.当样本数量不平衡时,并且当特征空间中混杂有大量噪音特征时,基于学习单一稀有类样本的产生式模型,相比于学习两类问题的判别式模型具有更好的性能[19].

3 分类器评价指标

鉴于大类对准确率标准的影响大于稀有类,导致稀有类的识别率难以提高,新的分类器评价指标更注重稀有类对性能指标的影响.

最常见的分类器评价指标是 ROC曲线,以及 ROC曲线下覆盖的面积 AUC[43]. ROC曲线和 AUC能够公平地对待稀有类和大类,与查准率和查全率类似, ROC曲线可以在稀有类识别率和大类识别率之间做权衡.

为了定义 ROC曲线,需要用到机器学习方法的基础评价指标——混淆矩阵(如表 1所示).

表 1 两类混淆矩阵

Table 1 A two-class confusion matrix

	预测正类	预测反类
预测正类	TP	FN
预测反类	FP	TN

在一个两类混淆矩阵中,实际为正类,预测也为正类的样本数量称为正确正类 TP(true positive);实际为正类,预测为反类的称为错误反类 FN(false negative);实际为反类,预测为正类的称为错误正类 FP(false positive);实际为反类,预测为反类的称为正确反类 TN(true negative).

利用混淆矩阵可以定义常用的分类器评价指

标, 如表 2所示.

表 2　常用评价指标

Table 2　Common evaluation metrics

名　称	计算方法
平衡准确率 BA	(TP/(TP+FN)+TN/(TN+FP))/2
查全率 TPR	TP/(TP+FN)
查准率	TP/(TP+FP)
误警率 FPR	FP/(FP+TN)
F1测度	2×R×P/(R+P)
几何平均	$\sqrt{R\times P}$

ROC曲线的 X轴表示 FPR, Y轴表示 TPR, ROC曲线上的一组点是通过调整分类器决策阈值得到的, ROC曲线越凸越靠近左上方, 表示对应的分类器一般化能力越强. AUC(area under curve)是指 ROC曲线下面包括的面积, 即 ROC曲线的积分, AUC能以定量的方式表示该 ROC曲线对应的分类器的一般化能力. 然而需要指出的是, ROC和 AUC仅适合于两类问题, 对多类问题, 无法直接应用.

查准率(Precision)和查全率(Recall)是信息检索与数据挖掘中常用的评价指标. 许多系统将两者同时考虑, 如 F1测度和几何平均(G-means), 它们都同等看待 Precision和 Recall对分类器评测的贡献. Joshi[44]比较系统地研究了包括 F1和 G-means在内的多种标准, 通过分析 Precision, Recall和训练样本不平衡度的关系得出结论, 在不平衡度不是很大的情况下, F1和 G-means可以作为较好的评测指标. 但当训练样本很不平衡时, F1比 G-mean要好. Josh指出, 当测试集和训练集不是同分布或不同类的错分代价不同时, 使用 ROC曲线比较不同分类器的性能更合适.

4　几种主要学习方法的性能比较

在实验中, 选取了 4种比较具有代表性的学习方法, 它们分别是, 代价敏感学习的决策树算法(C5.0)、代价敏感的支持向量机(CSVM)、SMOTE[14]采样法、以及结合训练集划分与分类器集成的最小最大模块化支持向量机(M3-SVM)[23].

实验使用了来自 3个不同领域的不平衡数据集. Abalone是 UCI标准数据集中比较难分的一个数据集, 类间重叠程度较大, 各种分类器在它上面的效果都不是很理想. 它总共有 29个类, 4 177个样本. 采用一对其他分解策略把其中第 11类的 487个样本作为正例, 其余作为负例, 构成一个不平衡的两类问题.

Park[45]蛋白质亚细胞定位数据, 是一个典型的生物信息学模式分类问题. 它总共有 7 579条蛋白质序列, 分布在 12个不同的亚细胞位置上, 有些位置上的蛋白质数量很不平衡. 最多的位置上有 1 932条序列, 而最少的位置只有 40条. 同样采用一对其他分解策略, 把其中细胞外的 861条作为正例, 其余的作为负例, 构成一个不平衡的两类分类问题.

Rooftop[46]数据是一个不平衡的场景识别问题, 它总共包括 17 829张图片, 其中只有 781张图片被标注成正例(屋顶照片), 其余的 17 048张图片都是负例.

上述 3个数据集的概况如表 3所示, 其中训练集和测试集是按照 6:4的比例随机划分得到的.

分类器性能的评价指标采用 ROC曲线、AUC以及 ROC曲线与对角线交点处的 TPR, TNR(即 1-FPR)和平衡准确率 BA.

图 1给出了 Park数据上几种方法的 ROC曲线. 从图中可以看出, M3-SVM具有最好的分类性能, CSVM的方法仅次于 M3-SVM, 而 C5.0方法效果最差. 使用 SMOTE采样方法前后, CSVM和 C5.0对应的 ROC曲线非常接近, 说明 SMOTE采样在 Park数据上没有明显的效果.

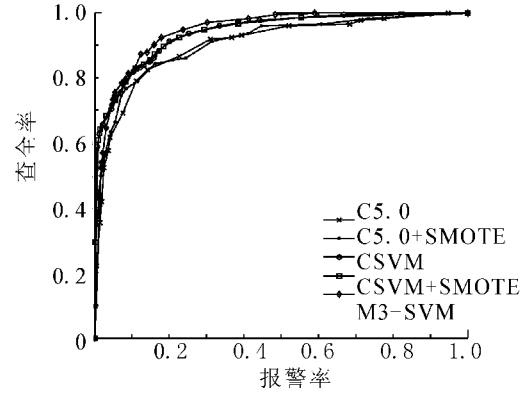

图 1　Park数据的 ROC曲线

Fig. 1　ROC curves for Park data

表 3　3种数据集的概况

Table 3　Three imbalanced data sets

名　称	正类样本数	负类样本数	正负样本数比
Rooftop	781	17 048	1∶21.8
Park	861	6 718	1∶7.8
Abalone	487	3 690	1∶7.6

表 4 各种方法的性能比较

Table 4 Performance comparison of different methods %

数据集	方法	TPR	TNR	BA	AUC
Rooftop	C5.0	78.5	80.2	79.9	87.43
	CSVM	80.3	**81.8**	81.1	87.98
	C5.0 + SMOTE	79.9	80.1	80.0	88.22
	CSVM + SMOTE	81.3	80.4	80.9	87.87
	M3 - SVM	**81.6**	81.4	**81.5**	**89.28**
Park	C5.0	82.6	85.8	84.2	90.39
	CSVM	84.9	85.5	85.2	93.93
	C5.0 + SMOTE	84.3	83.8	84.2	90.96
	CSVM + SMOTE	85.4	85.1	85.3	94.10
	M3 - SVM	**87.2**	**87.7**	**87.5**	**94.54**
Abalone	C5.0	61.5	59.6	60.6	66.84
	CSVM	59.0	58.8	58.9	64.25
	C5.0 + SMOTE	64.5	62.4	63.5	69.53
	CSVM + SMOTE	62.7	63.3	63.0	68.00
	M3 - SVM	**67.5**	**66.4**	**67.0**	**72.67**

表 4给出了在 3个数据集上其他的分类器性能指标. 从该表可以得到下面一些观察结果:

1)仅用代价敏感的 SVM和决策树算法在解决不平衡问题时效果较差, 两者的性能差异不大, 因数据集的不同而变化. 在 Rooftop数据上两者性能相近, 在 Park数据上 CSVM略好于决策树 C5.0, 而在 Abalone数据上 C5.0优于 CSVM.

2)在多数情况下 SMOTE采样的方法对分类精度有所提高. 但有时却没有什么效果甚至导致性能下降, 如在 Rooftop数据上, SMOTE采样后的数据经 CSVM分类比采样前分类效果反而下降.

3)结合数据集划分和分类器集成思想的 M3-SVM表现出了最好最稳定的分类性能.

5 结束语

本文综述了不平衡分类问题的特征、问题点以及已有的几种主要解决方案和新的分类器评测指标. 通过在 3个不同领域的不平衡数据集上的实验, 比较了决策树、支持向量机、代价敏感学习、采样方法以及训练集划分结合分类器集成等方法的性能. 实验结果表明, 训练集划分结合分类器集成的方法在处理不平衡问题时具有最好的效果. 目前研究不平衡模式分类问题都是基于不平衡的两类问题, 即使是不平衡的多类问题, 也是通过将原问题分解成 2类问题的方法解决. 另外目前没有针对多类不平衡分类问题的公认评价指标, ROC和 AUC不能直接运用于多类问题, 因此迫切需要提出针对多类不平衡分类问题的评价指标和相应的学习算法. 迄今为止, 不平衡模式分类问题的理论研究成果很少, 以上的研究多是依据实验的方法, 所得到的结果也多是经验性的. 因此进一步的理论分析非常重要.

参考文献:

[1] KUBAT M, HOLTE R C, MATWIN S. Machine learning for the detection of oil spills in satellite radar images [J]. Machine Learning, 1998, 30(2): 195-215.

[2] CHAN P K, STOLFO S J. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection [C] // Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998: 164-168.

[3] CHOE W, ERSOY O K, BINA M. Neural network schemes for detecting rare events in human genomic DNA [J]. Bioinformatics, 2000, 16(12): 1062-1072.

[4] PLANT C, BOHM C, BERNHARDT T, et al. Enhancing instance-based classification with local density: a new algorithm for classifying unbalanced biomedical data [J]. Bioinformatics, 2006, 22(8): 981-988.

[5] WEISS G M. Learning with rare cases and small disjuncts [C] // Proceedings of the 12th International Conference on Machine Learning. San Francisco: Morgan Kaufmann,

1995: 558-565.

[6] WEISS G M, HIRSH H. A quantitative study of small disjuncts[C] //Proceedings of the 17th National Conference on Artificial Intelligence. Texas: AAAI Press, 2000: 665-670.

[7] WEISS G M. Mining with rarity: a unifying framework[J]. SIGKDD Explorations, 2004, 6(1): 7-19.

[8] JAPKOWICZ N, STEPHEN S. The class imbalance problem: a systematic study[J]. Intelligent Data Analysis Journal, 2002, 6(5): 429-450.

[9] ARUNASALAM B, CHAWLA S. CCCS: a top-down associative classifier for imbalanced class distribution[C] //International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 517-522.

[10] DRUMMOND C, HOLTE R. Explicitly representing expected cost: an alternative to ROC representation[C] //Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2000: 187-207.

[11] PROVOST F, FAWCETT T. Robust classification for imprecise environments[J]. Machine Learning, 2001, 42(3): 203-231.

[12] DRUMMOND C, HOLTE R C. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling[C] //International Conference on Machine Learning. Washington DC, 2003: 152-154.

[13] LING C, LI C. Data mining for direct marketing problems and solutions[C] //Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998: 73-79.

[14] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[15] LEE S S. Noisy replication in skewed binary classification[J]. Computational Statistics and Data Analysis, 2000, 34(2): 165-191.

[16] KUBAT M, HOTLE R, MATWIN S. Learning when negative examples abound[C] //Proceedings of the 9th European Conference on Machine Learning. London: Springer-Verlag, 1997: 146-153.

[17] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C] //Proceedings of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1997: 179-186.

[18] CHEN X W, GERLACH B, CASASENT D. Pruning support vectors for imbalanced data classification[C] //Proceedings of 18th International Joint Conference on Neural Networks. Montreal, Quebec, Canada, 2005: 1883-1887.

[19] RASKUTTI B, KOWALCZYK A. Extreme re-balancing for SVMs: a case study[C] //International Conference on Machine Learning. Washington DC, 2003: 65-71.

[20] ESTABROOKS A, JAPKOWICZ N. A mixture-of-experts framework for learning from unbalanced data sets[C] //Proceedings of the 4th Intelligent Data Analysis Conference. Lisbon, Portugal, 2001: 34-43.

[21] AN R, LIU Y, JIN R, et al. On predicting rare classes with SVM ensembles in scene classification[C] //IEEE International Conference on Acoustics, Speech and Signal Processing. Hong Kong, 2003: 21-24.

[22] LU B L, ITO M. Task decomposition and module combination based on class relations: a modular neural network for pattern classification[J]. IEEE Transaction on Neural Networks, 1999, 10(5): 1244-1256.

[23] LU B L, WANG K A, UTIYAMA M, et al. A part-versus-part method for massively parallel training of support vector machines[C] //Proceedings of 17th International Joint Conference on Neural Networks. Budapest, Hungary, 2004: 735-740.

[24] YE Z F, LU B L. Learning imbalanced data sets with a min-max modular support vector machine[C] //Proceedings of the 20th International Joint Conference on Neural Networks. Orlando, USA, 2007: 1673-1678.

[25] KOTSIANTIS S B, PINTELAS P E. Mixture of expert agents for handling imbalanced data sets[J]. Annals of Mathematics, Computing & Teleinformatics, 2003, 1(1): 46-55.

[26] ESTABROOK A, TAEHO J, JAPKOWICZ N. A multiple resampling method for learning from imbalanced data sets[J]. Computational Intelligence, 2004, 20(1): 18-36.

[27] CHEN C, LIAW A, BREIMAN L. Using random forest to learn imbalanced data[R]. No. 666. Statistics Department, University of California at Berkeley, 2004.

[28] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: improving prediction of the minority class in boosting[C] //Proceedings of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Cavtat-Dubrovnik, Croatia, 2003: 107-119.

[29] LIU X Y, WU J X, ZHOU Z H. A cascade-based classification method for class-imbalanced data[J]. Journal of NanJing University Natural Science, 2006, 42(2): 148-155.

[30] ZHOU Z H, LIU X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transaction on Knowledge and Data Engineering, 2006, 18(1): 63-77.

[31] PAZZANI M, MERZ C, MURPHY P, et al. Reducing misclassification costs[C] //Proceedings of the 11th International Conference on Machine Learning. San Francisco, CA, USA, 1994: 217-225.

[32] DOMINGOS P. METACOST: a general method for making classifiers cost sensitive[C]//Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining. San Diego, CA: ACM Press, 1999: 155-164.

[33] CHE H G, BONGER R E, LIM C C. Dual-nu support vector machine with error rate and training size biasing[C]//Proceedings of the 25th IEEE International Conference on Acoustics, Speech and Signal Processing. Salt Lake City, USA: IEEE Press, 2001: 1269-1272.

[34] FAN W, STOLFO J S, ZHANG J X, et al. AdaCost: misclassification cost-sensitive boosting[C]//Proceedings of the 16th International Conference on Machine Learning. San Mateo, USA: 1999: 97-105.

[35] JOSHI M V, AGARWAL R C, KUMAR V. Predicting rare classes: can boosting make any weak learner strong[C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada: ACM Press, 2002: 297-306.

[36] CHAWLA N V. C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure[C]//International Conference on Machine Learning. Washington DC, 2003: 125-130.

[37] ELKAN C. The foundation of cost-sensitive learning[C]//Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, Washington, 2001: 239-246.

[38] CARDIE C, HOWE N. Improving minority class prediction using case-specific feature weights[C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1997: 57-65.

[39] ZHENG Z H, SRIHARI R. Optimally combining positive and negative features for text categorization[C]//International Conference on Machine Learning. Washington DC, 2003: 241-245.

[40] WU G, CHANG E Y. KBA: kernel boundary alignment considering imbalanced data distribution[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 786-795.

[41] HONG X, CHEN S, HARRIS C J. A kernel-based two-class classifier for imbalanced data sets[J]. IEEE Transaction on Neural Networks, 2007, 18(1): 28-41.

[42] SCHOLKOPF B, PLATT J C, TAYLOR J S, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1472.

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

[44] JOSHI M V. On evaluating performance of classifiers for rare classes[C]//Proceedings of the 2nd IEEE International Conference on Data Mining. Japan, 2002: 641-644.

[45] PARK K J, KANEHISA M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs[J]. Bioinformatics, 2003, 19(13): 1656-1663.

[46] MALOOF M A. Learning when data sets are imbalanced and when costs are unequal and unknown[C]//International Conference on Machine Learning. Washington DC, 2003: 154-160.

作者简介:

叶志飞, 男, 1983年生, 硕士, 主要研究方向为统计机器学习和模式分类.

文益民, 男, 1969年生, 博士后, 副教授, CCF高级会员, 主要研究方向为统计学习理论、生物信息学和图像处理. 发表学术论文20余篇.

吕宝粮, 男, 1960年生, 教授、博士生导师、博士、IEEE高级会员, 主要研究方向为仿脑计算理论与模型、神经网络理论与应用、机器学习、模式识别、脑—计算机接口、生物信息学与计算生物学. 已在 IEEE Trans. Neural Networks, IEEE Trans. Biomedical Engineering, Neural Networks和ICCV等国际期刊和会议上发表学术论文80余篇.