不平衡数据的集成分类算法综述*

李 勇[1,2]，刘战东[1]，张海军[1]

(1. 新疆师范大学 网络信息安全与舆情分析重点实验室，乌鲁木齐 830054；2. 南京航空航天大学 计算机科学与技术学院，南京 210016)

摘 要：集成学习是通过集成多个基分类器共同决策的机器学习技术，通过不同的样本集训练有差异的基分类器，得到的集成分类器可以有效地提高学习效果。在基分类器的训练过程中，可以通过代价敏感技术和数据采样实现不平衡数据的处理。由于集成学习在不平衡数据分类的优势，针对不平衡数据的集成分类算法得到广泛研究。详细分析了不平衡数据集成分类算法的研究现状，比较了现有算法的差异和各自存在的优点及问题，提出和分析了有待进一步研究的问题。

关键词：不平衡数据；集成学习；分类；代价敏感；数据采样

中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2014)05-1287-05

doi:10.3969/j.issn.1001-3695.2014.05.002

Review on ensemble algorithms for imbalanced data classification

LI Yong[1,2], LIU Zhan-dong[1], ZHANG Hai-jun[1]

(1. *Key Laboratory of Network Information Security & Public Opinion Analysis, Xinjiang Normal University, Urumqi* 830054, *China*; 2. *College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 210016, *China*)

Abstract: Ensemble learning by integrating multiple base classifiers that trained different set can effectively improve the classification accuracy. In the base classifier training process, imbalanced data set can be processed by either cost-sensitive or data sampling technology. Due to the advantages of ensemble learning in imbalanced data classification, ensemble algorithms for imbalanced data classification have been widely research. This paper surveyed the state of the art of imbalanced data ensemble classification algorithms, including the mechanisms and features of major existing learning algorithms, their advantages and disadvantages, highlighted the open research issues and future research directions.

Key words: imbalanced data; ensemble learning; classification; cost-sensitive; data sampling

分类是机器学习和数据挖掘中重要的知识获取手段之一。常见的分类算法如决策树、贝叶斯网络、支持向量机和神经网络等已经得到了广泛应用[1]。现有的分类算法通常假定用于训练的数据集是平衡的，即各类所含的样例数大致相等。当遇到类数据不平衡时，以总体分类精度为学习目标的传统分类算法会过多地关注多数类，而使少数类样本的分类性能下降[2,3]。然而在实际应用中，少数类样例被误分的代价要比多数类被误分的代价大。例如在软件缺陷预测中，有缺陷的样本数要远远小于无缺陷样本数，但分类的目标是识别出有缺陷的少数类样例；类似的还有医疗诊断、石油泄漏监测、网络入侵监测、信用卡欺诈等领域[4]。不平衡数据分类关注的是类数据不平衡或未被充分表达情况下学习算法的性能，该问题的研究已经成为机器学习领域的热门课题之一[2,5~7]。根据现有的研究成果，解决不平衡数据的分类问题可以引入代价敏感技术或通过采样技术使数据重平衡进行处理[8]。

随着集成学习技术的发展，越来越多的研究将集成学习技术引入不平衡数据的分类学习[9~15]。集成学习是通过训练集成多个弱分类器提高最终学习效果的一种技术[16,17]。采用集成学习进行类不平衡数据分类具有以下优势：a)不平衡数据的最优类分布和最优类代表样例的寻找可以与集成学习中的多次采样技术融合在一起，避免额外的学习代价；b)多个分类器的集成可以防止过拟合，降低单分类器在处理不平衡数据时可能产生的偏差。

1 不平衡数据与集成分类算法概述

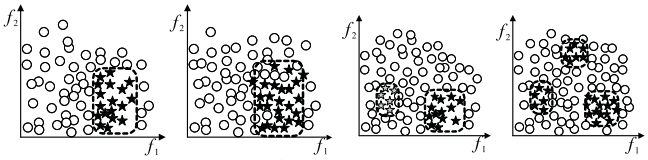
1.1 不平衡数据的特征

不平衡数据主要表现为类间不平衡和类内不平衡[4]两种类型。类间不平衡是指类之间呈现了不相等的数据分布，如图1(a)(b)所示。在一些实际应用中，数据呈现了极端的类与类之间的数据不平衡，不平衡率有些能达到 1000 ∶ 1 或者更高[18]。类内不平衡是指某个类与其子类的样本数量不平衡或者某类数据呈现多个小的分离项，如图1(c)(d)所示。大量研究表明，类间的数据不平衡不是影响分类学习的唯一因素，类内的数据不平衡才是影响分类效果的关键因素[3,4,19]。所以不平衡数据分类问题主要是数据分布的复杂性，如图1中(b)所示的数据重叠[20]、(c)所示的存在少数类的子类问题、(d)所示的小分离项问题[19]，这些问题都直接影响到分类器的学习

收稿日期：2013-09-14；**修回日期**：2013-10-28 **基金项目**：新疆自治区高校科研计划资助项目(XJEDU2012S28)；国家教育部人文社会科学研究青年基金资助项目(11YJC870014)；新疆师范大学重点实验室基金资助项目(WLYQ2012108)；国家自然科学基金资助项目(61163045)

作者简介：李勇(1983-)，男，讲师，博士研究生，主要研究方向为机器学习、软件智能(liyong@live.com)；刘战东(1982-)，男，讲师，硕士，主要研究方向为模式识别；张海军(1973-)，男，副教授，博士，主要研究方向为机器学习、自然语言处理.

效果。



(a)类间不平衡数据　(b)不平衡数据的重叠　(c)少数类的子类问题　(d)类数据的小分离项问题

图1　不平衡数据的特征

1.2　集成分类学习

集成分类学习是通过集成多个基分类器共同决策的机器学习技术,通过调用简单的分类算法,获得多个不同的基分类器,然后采用某种方式将基分类器组合成一个分类器。基分类器的准确性和差异性是影响最终集成分类器学习效果的两个重要因素[21]。Boosting[22]和 Bagging[23]算法是使用最广泛的集成学习算法,主要思想是通过对训练集进行不同的处理方式训练得到有差异基分类器,从而提高集成分类器的学习效果。

AdaBoost 算法[16]是 Boosting 算法的具体实现,是指在迭代过程中提高前一轮被错误分类的样本权值,降低被正确分类的样本权值,从而通过不同的训练样本集获得有差异的基分类器,最后通过加权表决集成最终分类器。AdaBoost. M1 和 AdaBoost. M2 算法[17]是 AdaBoost 算法在处理多分类问题的改进版本,在处理二分类问题时与 AdaBoost 算法基本上是相同的。AdaBoost 算法的伪代码如下所示:

输入:训练数据集$\{(x_1,y_1),\cdots,(x_m,y_m)\}$,其中 x_i 是输入空间 X 中的实例,y_i 是输出分类 Y 中相应的分类标签,$Y=\{-1,+1\}$。

初始化 $D^1(i)=1/m$;

对于 $t=1,2,\cdots,T$

a)使用 D^t 训练基分类器 $h_t\rightarrow Y$;

b)计算权值参数 α_t;

c)更新训练数据集的权值 $D^{(t+1)}(i)=D^t(i)\exp(-\alpha_t y_i h_t(x_i))/Z_t$,$Z_t$ 为规范化因子,$Z_t=\sum_i D^t(i)\exp(-\alpha_t y_i h_t(x_i))$;

输出:最终分类器 $H(x)=\mathrm{sing}(\sum_{(t=1)}^{T}\alpha_t h_t(x))$。

在 Bagging 算法中,从原始训练集有放回地随机选取若干样例组成的各基分类器训练集,通过多次选取不同的训练集增加了基分类器的差异度,从而提高最终集成分类器的泛化能力。Bagging 算法的伪代码如下所示:

输入:训练数据集$\{(x_1,y_1),\cdots,(x_m,y_m)\}$,其中 x_i 是输入空间 X 中的实例,y_i 是输出分类 Y 中相应的分类标签,$Y=\{-1,+1\}$。

对于 $t=1,2,\cdots,T$

a)可放回地随机抽取样例形成训练集 D^t;

b)使用 D^t 训练基分类器 $h_t\rightarrow Y$;

输出:最终分类器 $H(x)=\mathrm{sing}(\sum_{(t=1)}^{T}h_t(x))$。

1.3　不平衡数据的集成分类技术

集成分类算法是以总体学习精度为目标,不能直接用于处理不平衡数据的分类学习。根据现有的研究成果,在采用集成算法对不平衡数据进行分类时,可以在算法或数据两个层面进行处理。

算法处理是指在集成分类算法的训练过程中引入代价因子,根据不同类样例被错误分类的代价不同而赋予不同的代价因子,形成代价敏感的集成分类算法。由于 AdaBoost 算法是通过改变训练样本的权值而得到不同的基分类器训练集,通常在其样例权值更新中引入代价因子形成代价敏感的集成分类算法,相关算法在 2.1 节进行讨论。

数据处理是指在构建基分类器的过程中结合使数据重新平衡的采样技术,使集成算法可以在不影响学习性能的平衡训练数据上构建分类器。不同的重采样数据平衡策略和集成分类算法的结合形成了基于数据处理的 Boosting 集成分类算法、基于数据处理的 Bagging 集成分类算法和基于数据处理的混合集成分类算法。

1.4　不平衡数据分类算法评价体系

不平衡数据分类需要分类器能够在不影响多数类学习精度的前提下,对少数类样例有较高的分类精度。机器学习中常用的以总体分类精度为指标的评价准则不适合于不平衡数据分类算法的评价。在现有研究中通常采用可以提供更多信息的评价准则,如基于混淆矩阵的单评估指标、精确度—召回度曲线、ROC 曲线和成本曲线等,本节对使用最广泛的单评估指标和 ROC 曲线进行介绍。

1.4.1　单评估指标

不平衡数据的分类算法学习结果可以用混淆矩阵来表示,在本文中将分类学习任务关注的少数类定义为正类,多数类定义为负类。混淆矩阵的定义如表 1 所示。

表 1　二分类问题的混淆矩阵

	预测为正类	预测为负类
正类(positive)样例	正确正例 *TP*	错误负例 *FN*
负类(negative)样例	错误正例 *FP*	正确负例 *TN*

相关评估指标如下:

$$\text{准确率(precision)}=\frac{TP}{TP+FP}$$

$$\text{召回率(recall)}=\frac{TP}{TP+FN}$$

$$F\text{ 值}(F\text{-measure})=\frac{(\beta^2+1)\,\text{recall}\times\text{precision}}{\beta^2\times\text{recall}+\text{precision}},\beta\text{ 为参数}$$

$$\text{G-均值(G-mean)}=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}}$$

准确率又称为正确率,表示正类样例被正确分类的比例,对数据的分布敏感;召回率又称为查全率,表示正类样例被正确分类的完整度,是指分类器对正例分类的“能力”度量,对数据的分布不敏感。理想的分类结果是准确率和召回率越高越好,但在有些情况下这两个指标是矛盾的。例如极端情况下,只对不平衡数据中的一个正例进行了正确分类,则准确率为 100%,但召回率很低。因此在不同研究中可以通过绘制准确率—召回率曲线(precision-recall curves)[24]来进行分析。

F 值是准确率和召回率的加权调和平均,当参数 $\beta=1$ 时,就是最常见的 *F*1 值。*F*1 综合了准确率和召回率的结果,当 *F*1 值较高时说明结果比较理想。G-均值表示正例分类准确率和负例分类准确率的均衡值。*F* 值和 G-均值的关注点仍然是分类准确率,在分类器整体性能评估方面还存在欠缺。

1.4.2　ROC 曲线

为了解决单评估指标的不足,文献[24~26]研究了 ROC 曲线评估指标对不平衡数据的分类算法并进行评价。ROC 曲线是显示分类模型真正率和假正率之间折中的一种图形化方法。ROC 曲线如图 2 所示。

真正率(TPR)也称为灵敏度,表示正例被正确分类的样本数与正例样本数的比率,即 $TP/(TP+FN)$;假正率(FPR)也称为误报率,表示负例被错分为正类的样本数与负例样本数的比率,即 $FP/(FP+TN)$。

关于 ROC 曲线可以通过图 2 中的几个关键点进行解释:*A* 点(TPR = 0,FPR = 0)表示把每个样例都预测为负类的模型;*B* 点(TPR = 1,FPR = 1)表示把每个样例都预测为正类的模型;*C*

点(TPR=1,FPR=0)为理想模型,也就是说一个好的分类模型应该尽可能靠近图形的左上角,即 L_2 代表的模型要优于 L_1 代表的模型;点 E 位于连接点 A 和点 B 的主对角线上,表示为随机猜测模型。

ROC 曲线下方的面积(AUC)[27]提供了评价模型平均性能的另一种方法,AUC 值越大,模型越好。理想模型的 AUC 值为 1,随机猜测模型的 AUC 值为 0.5。

$$AUC = (1 + TPR - FPR)/2$$

ROC 曲线没有提供分类器性能的置信度,不能推断出不同分类器性能的统计特性,而且没有提供不同分类器执行时多个类的概率或错分代价,有研究提出采用代价曲线来解决上述问题,限于篇幅,可以参考文献[28]。

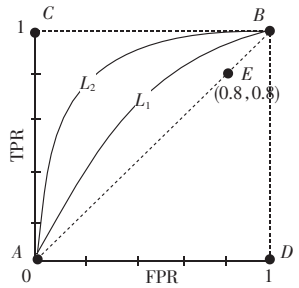


图 2　ROC 曲线

2　不平衡数据的集成分类算法

2.1　代价敏感的集成分类算法

在不平衡数据分类问题中,结合代价敏感技术的集成分类算法主要是在 AdaBoost 算法迭代更新样本权值的过程中引入代价因子。文献[29]指出,数据分布进行代价修正后,分布会偏向高代价类,使得分类算法可以更关注到少数类的样例。

在 AdaBoost 算法中,通过更新样本权值形成新的基分类器训练集,在每一次迭代中训练样本的权重更新为 $D^{(t+1)}(i) = D^t(i)\exp(-\alpha_t y_i h_t(x_i))/Z^t$,其中 $\alpha_t = 1/2\ \ln((1-\varepsilon_t)/\varepsilon_t)$ 为权值更新参数,$\varepsilon_t = \sum_{i,y_i \neq h_t(x_i)} D^t(i)$ 为训练集上的分类错误,Z^t 为归一化因子。

文献[11]将代价值 $C_i(C_i \in [0, +\infty))$ 引入 AdaBoost 算法的权值更新公式中,根据位置的不同,可以形成 $D^{(t+1)}(i) = D^t(i)\ \exp(-\alpha_t C_i y_i h_t(x_i))$、$D^{(t+1)}(i) = C_i D^t(i)\ \exp(-\alpha_t y_i h_t(x_i))$ 和 $D^{(t+1)}(i) = C_i D^t(i)\exp(-\alpha_t C_i y_i h_t(x_i))$,分别对应为 AdaC1、AdaC2 和 AdaC3 算法。代价值 C_i 表示将 x_i 误分的代价,通过提高少数类样例的代价值,使得算法可以在每次迭代生成基分类器过程中更加关注少数类样例,提高其分类精度。

类似算法还有 AdaCost[10]、CSB1 和 CSB2[9]。AdaCost 算法是将 AdaBoost 算法中样本的权重更新为 $D^{(t+1)}(i) = D^t(i) \times \exp(-\alpha_t y_i h_t(x_i)\beta_{\mathrm{sgn}(h_t(x_i),y_i)})$,其中 β 为代价调节函数,对于较高代价样例,函数值较高。文献[10]中给出代价调节函数的建议值,错误分类为 $\beta_+ = -0.5C_i + 0.5$,正确分类为 $\beta_- = 0.5C_i + 0.5$。其中 C_i 为第 i 个被错分类样例的代价,并且可以根据实际应用调整该值。在该算法中,如果给少数类和多数类样例赋予相同代价调节函数值,并不能蜕化为 AdaBoost 算法。CSB1 算法是将 AdaBoost 算法的权值更新方式变为 $D^{(t+1)}(i) = D^t(i)\,C_{\mathrm{sgn}(h_t(x_i),y_i)} \times \exp(-y_i h_t(x_i))$,CSB2 算法的权值更新为 $D^{(t+1)}(i) = D^t(i) \times C_{\mathrm{sgn}(h_t(x_i),y_i)} \times \exp(-\alpha_t y_i h_t(x_i))$。当 x_i 被正确分类时,$C_{\mathrm{sgn}(h_t(x_i),y_i)} = 1$,当 x_i 被错误分类时,$C_{\mathrm{sgn}(h_t(x_i),y_i)} = \mathrm{cost}(y_i, h_t(x_i)) \geqslant 1$,其中 $\mathrm{cost}(i,j)$ 为 i 类的样本被错误分为 j 类时的代价。

文献[11]采用 C4.5 作为基分类器对上述算法进行了实验比较。实验表明,AdaC2 和 AdaC3 的召回率要高于 AdaC1 算法,且对代价参数的设置较为敏感,AdaC2 算法的总体性能要优于 AdaC1 和 AdaC3 算法;CSB2 在代价参数改变时,性能波动较为明显;AdaCost 是将 AdaC1 的代价因子改为代价函数,召回率要高于 AdaC1,对代价参数的变化也不太敏感。

上述算法在实际应用中存在的问题是代价值或代价函数不容易定义,也有研究者针对该问题进行了研究。文献[30]提出了一种基于权重采样的 Boosting 算法,通过采样函数调整原始 Boosting 损失函数,使得分类器侧重于少数类样例的有效判别;作者同时给出了等同过采样、误分过采样和分界面过采样三种函数形式,实验表明分界面过采样分类性能最好。还有文献[13]提出的 RareBoost 算法,该算法是通过改变 AdaBoost 每一次迭代过程中权值参数 α_t 的计算方式来解决数据不平衡问题。在第 t 次迭代过程中利用混淆矩阵计算两个不同的 α_t 值:$\alpha_t^p = \frac{1}{2}\ln\frac{TP_t}{FP_t}$,$\alpha_t^n = \frac{1}{2}\ln\frac{TN_t}{FN_t}$。其中,$\alpha_t^p$ 用来预测为正类的样例更新权值,α_t^n 用来预测为负类的样例更新权值。在 RareBoost 算法中并没有使用代价因子,而是根据每个实例所预测的类标签通过 α_t^p 和 α_t^n 分别进行权值更新。该算法同样要保证正例的分类精度大于 50%。

在代价敏感的集成分类算法中只是在算法层次进行了修改,没有增加算法的开销,效率较高,有效提高了不平衡数据的分类效果。存在的问题是代价值或代价函数不容易定义,在标准实验数据集中,代价值也只能主观给出。

2.2　数据处理的集成分类算法

2.2.1　数据处理的 Boosting 集成分类算法

在 Boosting 集成分类算法中,通过样本权值的更新迭代生成不同的基分类器训练集,在此过程中使用采样技术实现数据的重新平衡。通过使用不同的数据重平衡采样策略产生了一系列处理不平衡数据的集成分类算法。

文献[31]提出合成正类样例的过采样技术(SMOTE)[32]和 AdaBoost.M2 算法相结合的 SMOTEBoost 算法[31]。基本思想是在每次提升迭代中引入 SMOTE 过采样技术,通过加入合成正例使得每个基分类器能更加关注正类样例。SMOTE 技术根据特征空间相似性在正类样本中构造合成样例,首先寻找某个正例的同类 K-近邻样例,随机选择其中一个,按照欧式距离在其之间插入合成正类样例。在 SMOTEBoost 算法中通过在基分类器不同权值的训练集迭代中实现数据的平衡,提高基分类器的差异性和最终的分类精度。但该算法通过插值的方式加入合成样例,使得新增合成样例只分布在原始样例的连线上,不能很好地反映数据的实际分布,容易造成过泛化[33]。文献[34]提出 MSMOTEBoost 算法对 SMOTEBoost 算法进行改进,在迭代过程中使用改进的合成正类样例过采样技术(MSMOTE)进行不平衡数据的处理。按距离将正类样例分为三组,即安全样例、边界样例和潜在的噪声样例。在 MSMOTE 中,安全样例的合成算法与 SMOTE 相同;对于边界样例则选择其距离最近的样例;而对于潜在的噪声样例则不进行任何操作。文献[34]的实验表明,MSMOTEBoost 算法的准确率和 F

值要优于 SMOTEBoost 算法。

文献[15]提出的 PCBoost 算法是在迭代过程中通过随机过采样的方式生成少数类的合成样例平衡数据集,并及时提高被当前基分类器错分的“困难样例”,使之能够得到更多的关注,同时修正扰动数据,消除错误添加的合成样例对最终集成分类结果的影响。实验结果表明,该算法具有处理不平衡数据的优势。

与上述采用过采样集成算法相对应的是结合欠采样的集成算法。文献[35]提出的 RUSBoost 算法是在 AdaBoost 算法的迭代过程中采用随机欠采样技术(RUS)从多数类中随机选择样例,不给其分配新的权重,从而使算法更加关注少数类样例。该算法与 SMOTEBoost 相比,具有实现简单、训练时间短等优势,但在欠采样中有可能移除潜在的有用多数类样例。为了避免该问题,在文献[12]中提出了 EusBoost 算法,该算法采用进化下采样技术[36],选择多数类中最具代表性的样例,实现与正类样本的数据平衡,并引入适应度函数保证基分类器的差异性,最终提高不平衡数据的集成分类精度。

与前述算法不同,文献[14]提出的 DataBoost-IM 算法是在 AdaBoost. M1 算法中采用文献[37]提出的数据合成技术。根据在迭代过程中多数类和少数类被误分类的样本数量比例生成合成样本,保证类间权值之和的平衡。由于多数类和少数类都找到了代表性样例,所以在分类时,可以做到在不牺牲多数类样例分类精度的前提下,获得少数类样例较高的学习精度。但是该算法同时生成的是两类样例,并没有及时修正错误添加的合成样例,造成其迭代过程可能面临过多的训练数据,影响到算法的执行效率。

2.2.2 基于数据处理的 Bagging 集成分类算法

由于 Bagging 算法实现简单、泛化能力强,有研究提出采用 Bagging 集成算法处理数据不平衡问题。基于数据处理的 Bagging 集成分类算法由于不需要计算与更新权重,比 Boosting 集成分类算法要简单。在这些算法中,对每一次有放回随机抽取基分类器训练集时,引入过采样和欠采样技术实现数据的重新平衡,保证了基分类器的差异性和集成分类器的学习精度。

文献[38]提出的 OverBagging 算法是将 Bagging 算法中有放回的随机采样技术替换为随机过采样技术来处理数据的不平衡问题。通过对少数类样本的过采样实现数据的平衡。由于每次迭代过采样都是面对所有的多数类样本,会导致基分类器训练集过大,影响分类学习的效率。文献[38]提出的 SMOTEBagging 算法,不同于 OverBagging 算法的随机过采样,在 OverBagging 中采样样本数取决于多数类的样本数,而在 SMOTEBagging 中,每次迭代时选择多数类样本数逐步倍增的方式,在此过程中数量不足的少数类样本通过 SMOTE 算法生成,同时在每次迭代中选择不同数量的多数类样本也可以提高基分类器的差异性。

同样也有文献提出在 Bagging 算法中采用欠采样的方式实现数据的平衡。文献[39]提出 UnderBagging 算法,该算法与 OverBagging 相反,在迭代生成基分类器训练集的过程对多数类欠采样,以达到数据的平衡。对少数类有放回地重复采样也能获得差异性大的基分类器,而且与 OverBagging 相比,由于是参照少数类样本进行采样,每个包的样例数会较少,效率较高。但在欠采样过程中容易忽略有用的多数类样例,造成分类结果的不精确。另外还有 Asymmetric bagging 算法[40]、Roughly balanced bagging 算法[41]等,算法的实现思想与 UnderBagging 基本上是类似的。

2.2.3 基于数据处理的混合集成分类算法

前面讨论的算法分别基于 Boosting 或 Bagging 算法实现不平衡数据的分类学习。而文献[42]提出的 EasyEnsemble 和 BalanceCascade 这两个算法是基于欠采样数据处理的 Boosting 和 Bagging 混合集成分类算法。EasyEnsemble 算法的基本思想是随机采样生成多数类样例的若干个与少数类样例数相等的子集,每个多数类样例子集和少数类样例构成若干个“平衡数据包”,然后采用 AdaBoost 算法训练生成若干个基分类器,最后进行集成。在该算法中采用 Bagging 作为主要的集成方式,但是每个基分类器采用 AdaBoost 方式训练生成,可以理解为集成算法的集成。

在 EasyEnsemble 算法中生成的“平衡数据包”在迭代过程中样例数量是不变的,而 BalanceCascade 算法与 EasyEnsemble 的并行训练方式不同,是在每一次迭代过程中删除被上一轮基分类器正确分类的多数类样例,进行基分类器的串行训练。不仅可以避免随机欠采样中忽略有用的多数类样例,而且综合了 AdaBoost 算法有效降低模型偏差和 Bagging 算法有效降低模型方差的优点,提高了不平衡数据分类算法的性能。

基于数据处理的集成分类算法在不平衡数据处理方面取得了不错的学习效果。从统计意义上讲,代价敏感和数据处理集成分类算法在处理不平衡数据时其性能是不相上下的[43],所以在实际应用中需要根据各种算法的优势及局限性选择最适合数据特征的算法。

3 进一步研究的问题

在代价敏感的集成分类算法中,存在的问题是代价值或代价函数不容易定义。在实际应用研究中,如何根据数据特征确定代价值或给出代价值的标准是进一步需要研究的问题;在数据处理集成分类算法中,重采样时如何确定数据集的最优分布是一个关键问题。另外,类内不平衡容易导致的小分离项、类间数据重合和噪声数据的处理等问题直接影响到最终的分类效果,如何结合集成算法的优势处理这类问题也是要研究的问题。

本文的关注点是不平衡数据分类的集成算法特性,而在集成分类算法中,基分类器算法的选择和差异性度量对不平衡数据分类的影响也是现在研究的热点。随着机器学习算法的研究,越来越多的技术可以用于不平衡数据的集成分类处理,如核方法、半监督方式的主动学习技术等,对这些算法进行更深入的理论分析,并在更多的数据集上进行实验来评价其性能和价值,是值得进一步探讨的问题。

4 结束语

不平衡数据的分类学习在许多领域都有广泛的应用,如软件缺陷预测、网络入侵检测等领域。由于集成技术在处理不平衡数据学习方面的优势,所以是近年来机器学习领域的研究热点。本文主要讨论不平衡数据集成分类算法的实现思路、研究进展及其算法分析,并探讨了其在理论和应用中有待进一步研究的问题。

参考文献:

[1] WU Xin-dong, KUMAR V, QUINLAN J R, *et al*. Top 10 algorithms in

data mining[J]. Knowledge and Information Systems,2008,14(1):1-37.
[2] CHAWLA N V,JAPKOWICZ N,KOTCZ A. Editorial:special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter,2004,6(1):1-6.
[3] 陶新民,郝思媛,张冬雪,等. 不均衡数据分类算法的综述[J]. 重庆邮电大学学报:自然科学版,2013,25(1):101-110.
[4] HE Hai-bo,GARCIA E A. Learning from imbalanced data[J]. IEEE Trans on Knowledge and Data Engineering,2009,21(9):1263-1284.
[5] 徐淑坦,王朝勇,孙延凤. 一种不均衡数据的改进蚁群分类算法[J]. 吉林大学学报:理学版,2011,49(4):733-739.
[6] 陶新民,郝思媛,张冬雪,等. 基于样本特性欠取样的不均衡支持向量机[J]. 控制与决策,2013,28(7):978-984.
[7] 夏战国,夏士雄,蔡世玉,等. 类不均衡的半监督高斯过程分类算法[J]. 通信学报,2013,34(5):42-51.
[8] 林智勇,郝志峰,杨晓伟. 不平衡数据分类的研究现状[J]. 计算机应用研究,2008,25(2):332-336.
[9] TING K M. A comparative study of cost-sensitive boosting algorithms[C]//Proc of the 17th International Conference on Machine Learning. 2000:983-990.
[10] FAN Wei,STOLFO S J,ZHANG Jun-xin,*et al*. AdaCost: misclassification cost-sensitive boosting[C]//Proc of the 16th International Conference on Machine Learning. 1999:97-105.
[11] SUN Yan-min,KAMEL M S,WONG A K C,*et al*. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12):3358-3378.
[12] GALAR M,FERNÁNDEZ A,BARRENCHEA E,*et al*. EUSBoost:enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling[J]. Pattern Recognition,2013,46(12):3460-3471.
[13] JOSHI M V,KUMAR V,AGARWAL R C. Evaluating boosting algorithms to classify rare classes:comparison and improvements[C]// Proc of IEEE International Conference on Data Mining. Washington DC:IEEE Computer Society,2001:257-264.
[14] GUO Hong-yu,VIKTOR H L. Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach[J]. SIGKDD Exploration Newsletter,2004,6(1):30-39.
[15] 李雄飞,李军,董元方,等. 一种新的不平衡数据学习算法 PCBoost[J]. 计算机学报,2012,35(2):2202-2209.
[16] FREUND Y,SCHAPIRE R. A desicion-theoretic generalization of online learning and an application to boosting[J]. Journal of Computer & System Sciences,1997,55(1):119-139.
[17] FREUND Y,SCHAPIRE R E. Experiments with a new boosting algorithm[C]//Proc of the 13th International Conference on Machine Learning. 1996.
[18] PEARSON R,GONEY G,SHWABER J. Imbalanced clustering for microarray time-series[C]//Proc of the 20th International Conference on Machine Learning. 2003.
[19] JO T,JAPKOWICZ N. Class imbalances versus small disjuncts[J]. SIGKDD Explorations Newsletter,2004,6(1):40-49.
[20] PRATI R C,BATISTA G E,MONARD M C. Class imbalances versus class overlapping: an analysis of a learning system behavior[C]//Advances in Artificial Intelligence. Berlin:Springer,2004:312-321.
[21] DIETTERICH T. Ensemble methods in machine learning[C]//Proc of the 1st International Conference on Multiple Classifier Systems. London:Springer,2000:1-15.
[22] SCHAPIRE R E. The strength of weak learnability[J]. Machine Learning,1990,5(2):197-227.
[23] BREIMAN L. Bagging predictors[J]. Machine Learning,1996,24(2):123-140.
[24] DAVIS J,GOADRICH M. The relationship between precision-recall and ROC curves[C]//Proc of the 23rd International Conference on Machine Learning. New York:ACM Press,2006:233-240.
[25] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters,2006,27(8):861-874.
[26] FAWCETT T. ROC graphs:notes and practical considerations for researchers[R]. Palo Alto:HP Lab,2004:1-38.
[27] 汪云云,陈松灿. 基于 AUC 的分类器评价和设计综述[J]. 模式识别与人工智能,2011,24(1):64-71.
[28] DRUMMOND C,HOLTE R C. Cost curves:an improved method for visualizing classifier performance[J]. Machine Learning,2006,65(1):95-130.
[29] ZADROZNY B,LANGFORD J,ABE N. Cost-sensitive learning by cost-proportionate example weighting[C]//Proc of the 3rd IEEE International Conference on Data Mining. 2003:435.
[30] 李秋洁,芳耀斌,王执铨. 基于 Boosting 的不平衡数据分类算法研究[J]. 计算机科学,2011,38(12):224-228.
[31] CHAWLA N,LAZAREVIC A,HALL L,*et al*. SMOTEBoost:improving prediction of the minority class in boosting[C]//Knowledge Discovery in Databases. Berlin:Springer,2003:107-119.
[32] 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(6):321-357.
[33] MEASE D,WYNER A J,BUJA A. Boosted classification trees and class probability/quantile estimation[J]. Journal of Machine Learning Research,2007,8(5):409-439.
[34] HU Sheng-guo,LIANG Yan-feng,MA Lin-tao,*et al*. MSMOTE:improving classification performance when training data is imbalanced[C]//Proc of the 2nd International Workshop on Computer Science and Engineering. Washington DC:IEEE Computer Society,2009:13-17.
[35] SEIFFERT C,KHOSHOFTAAR T M,Van HULSE J,*et al*. RUSBoost:a hybrid approach to alleviating class imbalance[J]. IEEE Trans on Systems, Man and Cybernetics, Part A:Systems and Humans,2010,40(1):185-197.
[36] GARCÍA S,HERRERA F. Evolutionary undersampling for classification with imbalanced datasets:proposals and taxonomy[J]. Evolutionary Computation,2009,17(3):275-306.
[37] GUO Hong-yu,VIKTOR H. Boosting with data generation:improving the classification of hard to learn examples[C]//Innovations in Applied Artificial Intelligence. Berlin:Springer,2004:1082-1091.
[38] WANG Shuo,YAO Xin. Diversity analysis on imbalanced data sets by using ensemble models[C]//Proc of IEEE Symposium on Computational Intelligence and Data Mining. [S. l.]:IEEE Press,2009:324-331.
[39] BARANDELA R,VALDOVINOS R M,SÁNCHEZ J S. New applications of ensembles of classifiers[J]. Pattern Analysis & Applications,2003,6(3):245-256.
[40] TAO Da-cheng,TANG Xiao-ou,LI Xue-long,*et al*. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2006,28(7):1088-1099.
[41] HIDO S,KASHIMA H,TAKAHASHI Y. Roughly balanced bagging for imbalanced data[J]. Statistical Analysis and Data Mining, 2009,2(5-6):412-426.
[42] LIU Xu-ying,WU Jian-xin,ZHOU Zhi-hua. Exploratory undersampling for class-imbalance learning[J]. IEEE Trans on Systems, Man, and Cybernetics, Part B:Cybernetics,2009,39(2):539-550.
[43] LÓPEZ V,FERNÁNDEZ A,MORENO-TORRES J G,*et al*. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification,open problems on intrinsic data characteristics[J]. Expert Systems with Applications,2012,39(7):6585-6608.