控制与决策
Control and Decision

嵌入重采样技术的C4.5决策树集成分类算法的临床医学预测

许召召， 申德荣， 寇月， 聂铁铮

引用本文:
许召召, 申德荣, 寇月, 等. 嵌入重采样技术的C4.5决策树集成分类算法的临床医学预测[J]. 控制与决策, 2021, 36(6): 1342–1350.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1247

嵌入重采样技术的C4.5决策树集成分类算法的临床医学预测

许召召, 申德荣†, 寇 月, 聂铁铮

(东北大学 计算机科学与工程学院，沈阳 110169)

摘 要: 决策树作为一种经典的分类算法,因其分类规则简单易懂被广泛应用于医学数据分析中. 然而,医学数据的样本不平衡问题使得决策树算法的分类效果降低. 数据重采样是目前解决样本不平衡问题的常见方法,通过改变样本分布提升少数类样本的分类性能. 现有重采样方法往往独立于后续学习算法,采样后的数据对于弱分类器的构建不一定有效. 鉴于此,提出一种基于C4.5算法的混合采样算法. 该算法以C4.5算法为迭代采样的评价准则控制过采样和欠采样的迭代过程,同时依据数据的不平衡比动态更新过采样的采样倍率,最终以投票机制组合多个弱分类器预测结果. 通过在9组UCI数据集上的对比实验,表明所提出算法的有效性,同时算法也在稽留流产数据上实现了准确的预测.

关键词: 不平衡数据；数据重采样；决策树；集成学习；混合采样；稽留流产

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1247 **开放科学(资源服务)标识码(OSID):**

引用格式: 许召召,申德荣,寇月,等. 嵌入重采样技术的C4.5决策树集成分类算法的临床医学预测 [J]. 控制与决策, 2021, 36(6): 1342-1350.

Clinical prediction of C4.5 decision tree classification algorithm with embedded resampling technique

XU Zhao-zhao, SHEN De-rong†, KOU Yue, NIE Tie-zheng

(College of Computer Science and Engineering，Northeastern University，Shenyang 110169，China)

Abstract: As a classical classification algorithm, the decision tree algorithm is widely used in medical data analysis because its classification rules are easy to understand. However, the unbalanced sample of medical data reduces the classification effect of the decision tree algorithm. Data resampling is a common method for solving the problem of sample imbalance. It mainly improves the classification performance of minority samples by changing the sample distribution. The existing resampling methods are often independent of the subsequent learning algorithms and the sampled data may not be effective for the construction of weak classifiers. Based on the above observations, we propose a hybrid sampling algorithm based on C4.5. Specifically, this algorithm controls the iterative process of oversampling and undersampling with the evaluation criteria of iterative sampling based on the C4.5. In addition, we dynamically update the sampling ratio of the oversampling based on the unbalanced ratio of the data and eventually combine multiple weak classifiers to predict the results with a voting mechanism. The effectiveness of the proposed algorithm is proved by the comparison experiments on 9 UCI datasets, and the algorithm also achieves accurate predictions on the missed abortion data.

Keywords: unbalanced data；data resampling；decision tree；ensemble learning；hybrid sampling；mission abortion

0 引 言

信息技术在医疗领域的广泛应用促进了医疗机构的信息数字化,同时,电子病历和病案在医院的普及使得医院数据库产生大量医学数据,这些医学数据对于疾病的诊断、治疗和医学研究都非常有价值[1]. 分类作为数据挖掘领域中的重要技术之一,通过对医学数据中的正常样本和病例样本进行构建模型,并利用分类模型实现对未知医学数据的诊断. 决策树作为一种经典的分类算法,因其分类精度高、运转速度快等优点广泛应用于医学数据分析中[2]. 然而

收稿日期: 2019-09-04；**修回日期:** 2019-12-24.

基金项目: 国家自然科学基金项目 (61672142, 61472070, 61602103, 62072084, 62072086)；国家重点研发计划项目 (2018YFB1003404).

责任编委: 阳春华.

†通讯作者. E-mail: shenderong@cse.neu.edu.cn.

有研究表明[3],医学数据的高度不平衡性极大地损害了决策树的分类性能.

数据重采样[4-5]是解决数据不平衡问题的常见方法,依据采样方式不同分为欠采样、过采样和混合采样. 欠采样中以随机欠采样最为常见,通过随机删除多数类样本达到样本均衡的目的. 然而,随机删除样本容易导致重要信息被舍弃,具有一定的盲目性,有学者提出编辑最近邻(edited nearest neighbours, ENN)[6],删除那些周围为少数类样本的多数类样本,该算法虽然可以有效删除多数类样本,但多数类样本周围往往都是多数类样本,删除的样本十分有限[7-8]. 因此,有学者[9]基于k近邻中心并依据聚类中心执行欠采样,也有学者[10]基于聚类欠采样的思想将其与集成学习方法相结合使用,以提升对少数类样本的识别率.

与欠采样相反,过采样通过合成少数类样本提升样本数目. 少数类样本合成技术(synthetic minority oversampling technique, SMOTE)[11]是最为经典的过采样算法,通过随机选择少数类样本和其近邻进行线性插值,生成无重复的少数类样本,有效避免了过拟合现象. 但是,该算法容易产生噪声样本和边界样本,这对于分类算法的模型构建极为不利[12]. 为此,相继出现了一系列SMOTE的改进算法,如Borderline-SMOTE[13]和ADASYN-SMOTE[14]等. 近年来,有学者考虑到数据集正负类样本间的类间距离、类内距离与不平衡之间的联系进行过采样[15];也有学者提出将聚类思想引入过采样中,即重新对数据进行聚类分簇,然后使用在簇内进行过采样[16].

基于欠采样和过采样的方法都有各自的优缺点,为了取得良好的采样效果,可以将两种方法结合使用. 首先使用过采样方法对少数类样本进行线性插值;然后使用欠采样方法删除多数类和之前合成的噪声样本. 如Batista等[17]提出了SMOTE与ENN相结合的算法,能很好地克服SMOTE的噪声问题;陶新民等[18]提出基于SVM的ODR-BSMOTE的混合采样算法,在实现样本平衡的同时更多地删除噪声样本和重复样本;Li等[19]使用两个独立的粒子群优化算法,分别对样本进行过采样和欠采样,以较短的时间达到样本平衡的目的. 混合采样方法有效地融合了两种采样方法的优点,但采样过程独立于后续学习算法,混合采样的过程具有一定的盲目性[20].

基于上述描述可知,混合采样在一定程度上解决了两者的缺点,但混合采样后的样本质量并未得到保证[21]. 对于医学数据分类而言,极小的分类错误率的后果也是十分严重的. 因此,本文在混合采样的基础上提出一种基于C4.5算法的混合采样算法(SMOTE and ENN based on C4.5 decision tree, CSE). 该算法的核心原理是混合采样的多次迭代采样过程,即首先使用改进SMOTE算法对少数类样本进行线性插值;然后使用ENN欠采样对合成后的样本进行去噪. 该算法的优势在于:针对SMOTE算法依据样本不平衡比而设置的过采样倍率,提出一种动态更新的过采样倍率,当两类样本差距较大时,过采样倍率依据两类不平衡设置;当两类样本差距较小时,过采样倍率依据增加一定百分率设置. 对于ENN欠采样模块,与传统欠采样模块不同的是,本文提出使用C4.5算法检测每轮欠采样后样本的分类性能. 为了删除更多的噪声样本,通过多轮混合采样过程,使得两类样本的类别动态变化. 此外,在CSE混合采样的迭代过程中,增加了一种新的停止策略,即若使用欠采样后的样本数目未发生改变,则停止迭代采样过程. 通过在多个高度不平衡的UCI数据集与已有采样算法的对比实验,验证了所提出混合采样算法的可靠性. 最后,将所提出采样算法应用于私有医学数据中,实现了稽留流产的准确预测.

1　集成决策树

1.1　决策树

决策树是一种以实例归纳为基础生成树状型分类规则的算法,因其分类规则简单易懂而被广泛应用于医疗领域[22]. 本文选取C4.5算法[23]为弱分类器,从一个无次序的实例集合中归纳出分类规则,C4.5算法的模型构建过程可以定义如下.

假设样本集为S,其中每个样本由一个包含m项的属性向量表示,假设类别属性A_m具有k个不同取值,那么根据不同取值可以将样本集S划分为k个子集,进而得出样本集S对分类的平均信息量为

$$H(S)=-\sum_{p=1}^{k}P(C_p)\log_2 P(C_p). \tag{1}$$

假设条件属性A_i中有t个不同的取值,那么根据A_i的取值可以将样本集S划分为t个子集. 此外,还可以将$c_i(i=1,2,\dots,k)$进一步划分为k个子集,每个子集C_{pq}表示在$A_i=a_q$的条件下属于第p类的样本集合. 由此,对属性进行划分后,样本集S对分类的平均信息量为

$$H(S/A_i)=-\sum_{q=1}^{t}P(C_q)\Big[-\sum_{p=1}^{k}P(C_{pq})\log_2 P(C_{pq})\Big]. \tag{2}$$

其中: $P(C_q) = |C_p|/|S|(1 \ll p \leqslant k)$, $|C_p|$和$|S|$分别为C和S中实例的个数. 利用A_i对S进行划分的信息增益为

$$G(S, A_i) = H(S) - H(S/A_i). \tag{3}$$

由于使用属性A_i对S进行划分的信息增益率等于信息增益量与分割信息量(split information)之比, 可以得到信息增益率为

$$\mathrm{GR}(S, A_i) = G(S, A_i)/S_p(S, A_i), \tag{4}$$

其中$S_p(S, A_i)$为样本集S的分裂信息量, 定义为

$$S_p(S, A_i) = -\sum_{t=1}^{t}(|S_t|/|S|)\log_2(|S_t|/|S|).$$

1.2 集成学习

决策树在属性分裂过程中存在较高的方差, 数据集微小的波动就会产生完全不同的分裂, 导致分类规则过于复杂, 包含了太多的噪声, 即产生过拟合现象[24]. 因此, 有学者提出使用Bagging方法组合多个决策树的集成方法[2]. 给定决策树和训练集, 通过有放回抽样k次生成k个训练子集. 使用决策树对k个子集进行训练得到分类器$\{C_1, C_2, \ldots, C_k\}$, 通过投票机制输出预测结果$C^*(x)$. 本文选取Bagging方法对C4.5算法进行集成学习, 算法流程如下.

算法1 Bagging-C4.5算法.

输入: 训练集S, 学习算法C, 迭代次数k;

输出: 投票结果$C^*(x)$.

step 1: 从训练集S中进行k次有放回的抽样, 每次共抽取n个训练样本, 得到k个训练子集$S = \{S_1, S_2, \ldots, S_k\}$.

step 2: 对训练子集S_i进行预处理, 根据式(1)~(4)计算所有属性的信息增益率, 选择信息增益率最大的属性为根节点的分裂属性.

step 3: 根据根节点属性值的不同, 采取与step 2同样的方法递归地建立后继分枝, 选择分枝中信息增益率最大的属性进行分裂, 如此训练下去, 直到所有分枝节点的样本都属于同一类别.

step 4: 采用后剪枝策略对生成的决策树进行剪枝操作, 防止生成的决策树过拟合.

step 5: 对生成的决策树进行剪枝, 消除噪声和孤立点等随机因素的影响, 得到简化的决策树.

step 6: 将训练子集S_i生成的分类规则进行提取, 并生成对应的弱分类器C_i. 返回step 2重复进行训练, 直至生成k个弱分类器$\{C_1, C_2, \ldots, C_k\}$.

step 7: 将step 6得到的k个弱分类器对测试集进行组合预测, 采用投票的方法输出结果

$$C^*(x) = \arg\max_{y \in Y}\sum_{i=1}^{k} I(C_i(x) = y), \tag{5}$$

其中$y \in Y$为单个决策树分类的输出结果.

2 基于C4.5决策树的混合采样算法

2.1 传统采样算法的缺点

SMOTE是一种少数类样本合成技术, 其核心思想是对少数类样本进行分析, 并按照线性插值的方法合成少数类样本, 从而解决随机过采样采取简单复制样本导致模型过拟合的问题. SMOTE算法广泛应用于不平衡医学数据中[25], 如患病样本通常只有几十或几百, 而正常样本却高达上万甚至数十万.

设样本集为$S = \{x_1, x_2, \ldots, x_n\}$, 选择少数类样本$S_{i_\min}$, 以欧氏距离计算$S_{i_\min}$与其他少数类样本的距离, 得到$k$个最近邻样本. 依据多数类样本和少数类样本的数目比设置过采样倍率N, 在其k个近邻中随机选取最近N个样本$S_{k_\min}$. 在少数类样本$S_{i_\min}$和其N个最近邻之间进行随机线性插值, 合成新的少数类样本S_{new}, 有

$$S_{\mathrm{new}} = S_{i\_\min} + \mathrm{rand}(0, 1)(S_{k\_\min} - S_{\min}), \tag{6}$$

其中$\mathrm{rand}(0, 1)$表示0~1之间的一个随机数.

将上述合成方法进行多次合成, 直至达到两类数目均衡, 将这些合成的少数类样本与原始少数类样本组合成新的少数类样本. SMOTE算法在一定程度上解决了随机过采样的盲目性, 但会造成样本重叠、边界样本以及噪声样本等问题.

ENN是一种经典的欠采样算法, 其通过删除多数类样本中的噪声样本和边界样本, 达到两类样本平衡的目的. 该算法的流程是对样本S_{i_maj}按照k近邻规则进行分类, 若不能被正确分类, 则将S_{i_maj}从样本集中移除.

假设多数类样本为S_{i_maj}, 以欧氏距离计算与其他多数类样本的距离, 得到k个最近邻样本, 从k个最近邻样本中找到数量最多的那个类别, 并将类别标签赋S_{i_maj}, 样本类别的判别函数可以定义为

$$C_z = \arg\max_{v \in C}\sum_{y \in N_k(S)} I(v = \mathrm{class}(C_y)), \tag{7}$$

其中$I(\cdot)$为k近邻样本的判别函数.

如果S_{i_maj}与其$k/2$个近邻类别不同, 则删除. 虽然ENN算法可以删除多数类中的噪声样本和类边界样本, 但是多数类样本的近邻往往都属于多数类, 能删除的样本个数很少. 此外, 经过欠采样后的样本可能会面临新的样本边界等问题.

2.2 基于C4.5算法的混合采样算法

单一过采样和欠采样均能使数目均衡,但采样后的样本对集成学习的弱分类器构建不一定有效[26]. 此外,传统的混合采样算法也未考虑样本合成的质量问题. 基于此,本文提出将SMOTE和ENN相结合使用. 与传统的混合采样不同的是,所提出混合采样通过动态设置过采样倍率对数据样本进行多次混合采样,设置一个阈值,当欠采样后的多数类样本数目少于该阈值时,停止混合采样的迭代过程. 此外,在混合迭代采样的过程中使用C4.5算法的分类性能检测最佳采样比.

CSE算法的工作原理可以分为两部分:首先使用动态更新采样倍率的SMOTE算法对少数类样本进行过采样;然后使用ENN算法对合成后的样本进行欠采样,用于删除合成后的噪声样本和边界样本. 此外,增加一种CSE算法迭代采样的终止策略,即若欠采样后多数类样本的数目未发生改变,则终止迭代采样. 为了解决ENN算法删除多数类中噪声样本有限的问题,动态过采样倍率可以分为两种:

1) 当少数类样本远少于多数类样本时,过采样倍率使用多数类和少数类样本的数目比作为采样倍率. 定义如下:

$$N_{\text{first}}=\frac{\sum_{i=1}^{m}S_{i\_\text{maj}}-\sum_{j=1}^{n}S_{j\_\min}}{\sum_{j=1}^{n}S_{j\_\min}}\times 100\,\%. \tag{8}$$

其中:S_{i_maj}为多数类样本,$S_{i_\min}$为少数类样本. 通过多数类和少数类样本的采样倍率可以设置需要合成的少数类样本数目$S_{j_\text{syn}}=\sum_{j=1}^{n}S_{j_\min}\times N_{\text{first}}$.

2) 当少数类的样本数目接近于多数类样本或多次迭代采样后的两类样本数目接近时,采取一种动态更新的过采样倍率方法,根据欠采样后的多数类和少数类的错分比进行采样,定义如下:

$$N_{\text{CSE}}=\frac{\sum_{i=1}^{m'}S'_{i\_\text{maj}}}{\sum_{j=1}^{n'}S'_{i\_\min}}\times 100\,\%. \tag{9}$$

其中:S'_{i_maj}为多数类样本中错误分类的样本;$S'_{j_\min}$为少数类样本中错误分类的样本. 由此,可以得出需要合成的少数类样本数目

$$S'_{j\_\text{syn}}=\sum_{j=1}^{\overline{n}}S'_{j\_\text{maj}}\times N_{\text{CSE}}.$$

CSE算法的具体流程如下.

算法2　CSE算法.

输入:　数据集S,最近邻个数k,少数类样本$S_{i_\min}$,多数类样本S_{j_maj};

输出: 混合采样后的数据集S'.

step 1: 初始化过采样倍率N,如果少数类样本远少于多数类样本,则由式(8)设置过采样,否则由式(9)设置过采样倍率.

step 2: 根据过采样倍率N,计算需要合成的样本总数$S_{\text{syn}}(N\times S_{\min})$,遍历每一个少数类样本$S_{i_\min}$.

step 3: 设置合成样本计数C_{syn},根据欧氏距离查找$S_{i_\min}$的k个最近邻样本,并将其索引存入$K_{\min}[\,]$中.

step 4: 随机生成一个0～1之间的实数G_{ap},根据式(6)合成少数类样本,并将合成的样本加入样本集S中.

step 5: $C_{\text{syn}}=C_{\text{syn}}+1$,如果$C_{\text{syn}}<S_{\text{syn}}$,则返回step 2,否则继续执行. 结束对$S_{i_\min}$的操作,将合成的少数类样本加入数据集$S$中.

step 6: 使用C4.5算法对过采样的样本集进行建模训练,并将其存入$\text{Eva}[\,]$. 比较$\text{Eva}[\,]$与$\text{Eva}_{\text{best}}[\,]$的大小,将最大值存入$\text{Eva}_{\text{best}}[\,]$中.

step 7:　设置删除样本计数$C_{\text{del}}=0$,在多数类中查找S_{i_maj}的k个最近邻样本,并将其索引存入$K_{\text{maj}}[\,]$中.

Step 8:　根据类别判别函数判断多数类样本S_{j_maj}与其近邻样本的类别,若S_{j_maj}与其紧邻样本的类别有$k/2$个不同,则删除S_{j_maj}.

step 9: $C_{\text{del}}=C_{\text{del}}+1$,如果$C_{\text{del}}<S_{\text{maj}}$,则返回step 7,否则结束对$S_{j_\text{maj}}$的操作.

step 10: 使用C4.5算法对欠采样后的样本集进行建模训练,将其分类结果存入$\text{Eva}[\,]$中. 比较$\text{Eva}[\,]$与$\text{Eva}_{\text{best}}[\,]$的大小,将最大值存入$\text{Eva}_{\text{best}}[\,]$中.

step 11: 统计欠采样后多数类样本的数目,如果未减少,则终止迭代采样,得到最终平衡的样本集,否则返回step 1.

3 实验结果与分析

3.1 评估指标

传统的评估指标已不适用评估不平衡数据的分类结果,有学者提出使用类别分类结果来评估其分类效果[27]. 如果用TP表示正确预测的多数类样本数目,TN表示正确预测的少数类样本数目,FP表示将少数类预测成多数类的数目,FN表示将多数类预测成少数类的数目,则有:

少数类样本准确性

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}). \tag{10}$$

多数类样本准确性

$$\text{Specificity} = \text{TN}/(\text{FP}+\text{TN}). \tag{11}$$

F-measure是一种不平衡数据的评价指标,主要针对多数类的分类性能进行评价,定义为

$$F\text{-measure} = \frac{2\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}. \tag{12}$$

除F-measure指标外,MCC是一种综合考虑数据集中少数类样本准确性和多数类样本准确性的指标,MCC指标可以定义为

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TN}+\text{FN})(\text{TN}+\text{FP})(\text{TP}+\text{FN})(\text{TP}+\text{FP})}}. \tag{13}$$

由式(13)可知,当数据不平衡时,MCC的值通常远小于Sensitivity和Specificity. 因为此时TN和FP是一个数量级,远大于TP和FN,选取MCC可以显著观察不平衡数据的分类效果.

3.2 实验数据

为了验证CSE算法的有效性,选取9组UCI数据集和私有医学数据集进行实验. 其中数据可分为验证数据和医学数据两部分,这些数据都是高度不平衡的,对于多类别数据,将多个少数类合并为一个类别. 表1给出了各数据集的样本数目、属性个数、少数类样本数目以及两类样本比. 表1中missed abortion为稽留流产数据集. 稽留流产[28]是指因胚胎死亡滞留宫腔不能及时自然排出机体,胎盘溶解产生溶血活酶进入母体血液循环,导致患者易合并凝血功能障碍,甚至危及患者生命. 有研究表明[29],甲状腺功能异常是导致孕妇稽留流产的主要原因之一,因此在发病之前的检测是避免稽留流产发生的有效手段之一. 本文选取某医院的孕妇的甲状腺体检样本作为应用实验数据集.

表1 UCI数据集描述

数据集	样本数目	属性个数	少数类样本数目	多数类样本数目	不平衡比
spambase	4 601	57	1 813	2 788	1.54
abalone	4 177	8	391	3 786	9.68
eighthr	2 534	72	160	2 374	14.8
diabetes	768	8	268	500	1.87
balance	625	4	49	576	11.8
wdbc	569	30	212	357	1.68
ionosphere	351	34	126	225	1.79
haberman	306	81	81	225	2.78
wpbc	198	33	47	151	3.21
missed abortion	361	8	112	249	2.22

3.3 实验结果

3.3.1 CSE算法的迭代采样

本节主要讨论CSE算法的迭代采样实验,通过设置程序中断记录每一轮采样后数据样本的分类性能,实验使用C4.5算法作为评估算法测试采样后的样本集. 实验共分为两部分,首先使用改进SMOTE算法对少数类样本进行线性插值;然后使用ENN算法对多数类样本进行筛选. 图1和图2记录了CSE算法迭代采样过程中数据样本的分类性能.

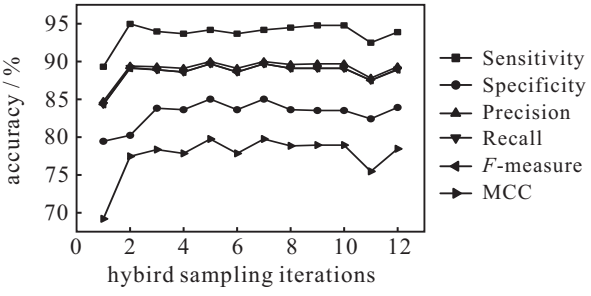

图1 CSE算法在balance数据集上的迭代采样

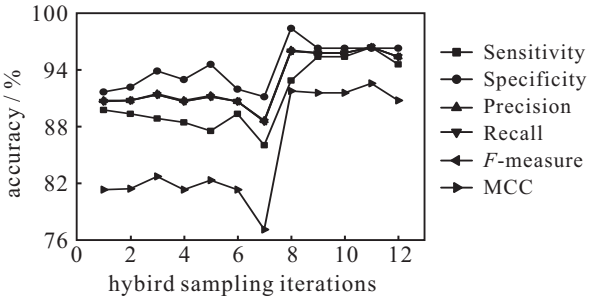

图2 CSE算法在ionosphere数据集上的迭代采样

由图1和图2可以看出,随着迭代采样的递增,Sensitivity和Specificity指标的走势是相同的. 而F-measure值和MCC值变化趋势是不同的,其中F-measure值随迭代次数的递增有增有减,观察发现其递减往往发生在过采样后,而在欠采样后又有所提升,这可能是由于在过采样时合成了较多的边界样本导致少数类的分类性能受到影响. MCC值随迭代次数的增加基本呈现逐步递增的趋势,因此在迭代过程

中多数类的分类性能也有所提升. 综合考虑,选择以MCC指标作为评估标准选择迭代次数最优的采样比例.

3.3.2 采样算法对比实验

在UCI数据集上将CSE算法与传统算法进行对比,对比算法分别为: 过采样算法SMOTE(SM)[11]、Borderline-SMOTE(BSM)[13]、ADASYN-SMOTE(ASM)[14]; 欠采样算法ENN[6]、Tomek Links(TL)[8]、Instance Hardness Threshold (IHT)[30]; SMOTE和ENN(SM+ENN) 以及Tomek Links (SM+TL)[17] 相结合的混合采样算法. 其中,3种过采样算法的最近邻个数统一设置为5, Borderline-SMOTE算法的参数m设置为10. 此外,使用C4.5算法对采样后的数据进行测试,实验结果如表2～表4所示.

由表2～表4可知,经过CSE采样后样本的分类性能均普遍高于对原始数据直接分类的C4.5算法. 对比CSE算法和传统采样算法可知,CSE算法Sensitivity、Specificity和MCC在各个数据集上的平均值分别为95.67%、94.51%、91.11%,显著优于其他采样算法的平均值. 在abalone数据集上不如IHT算法,主要是由于其根据实例硬度能够删除多数类样本. 因此,abalone数据集存在过多的多数类噪声样本,单纯地合成少数类样本并不合适. 混合采样中SMOTE + ENN算法明显优于SMOTE + TOMEK, 且

表2　采样算法采样后UCI数据集的Sensitivity值

数据集	original	欠采样			过采样			混合采样		
		SM	BSM	ADS	ENN	TL	IHT	SM-ENN	SM-TL	CSE
spambase	90.8	93.7	93.7	93.2	94.4	97.7	91.4	96.0	93.3	97.9
abalone	0.00	91.1	92.7	92.1	61.6	0.00	99.0	96.7	91.6	97.8
eighthr	33.8	90.6	96.0	94.3	38.8	31.9	97.5	95.0	91.8	95.0
diabetes	59.7	79.9	83.6	80.8	83.2	70.9	84.0	93.2	80.1	95.7
balance	0.0	89.4	82.6	89.6	0.0	0.0	63.3	95.1	93.5	94.9
wdbc	92.5	93.8	95.8	95.0	92.5	94.6	97.2	96.6	94.1	97.5
ionosphere	82.5	89.7	88.8	90.9	84.9	82.5	90.5	89.3	89.7	96.3
haberman	29.6	70.5	80.0	76.8	56.8	49.4	74.1	93.6	74.3	96.0
wpbc	44.7	78.7	82.4	84.2	52.2	54.3	87.0	91.2	83.7	89.9
average	48.18	86.38	88.40	88.54	62.71	53.44	87.11	94.08	88.01	95.67

表3　采样算法采样后UCI数据集的Specificity值

数据集	original	欠采样			过采样			混合采样		
		SM	BSM	ADS	ENN	TL	IHT	SM-ENN	SM-TL	CSE
spambase	94.4	93.9	94.0	93.1	95.6	97.3	94.8	96.4	93.0	97.8
abalone	100.0	86.7	89.2	87.9	96.9	100.0	99.7	92.9	87.2	97.4
eighthr	96.6	88.8	91.7	92.0	96.3	96.9	100.0	91.0	89.1	94.4
diabetes	81.4	73.4	68.8	70.2	85.8	83.1	88.1	80.4	71.1	96.5
balance	100.0	79.5	92.4	80.6	100.0	100.0	86.3	80.3	78.8	83.6
wdbc	93.6	94.4	95.5	95.2	95.9	95.7	97.2	95.6	94.6	98.3
ionosphere	96.4	91.6	89.8	89.3	92.6	96.4	93.7	92.1	91.6	96.2
haberman	87.1	71.1	64.4	61.8	89.9	84.7	98.8	78.3	69.8	96.3
wpbc	85.4	81.5	77.7	72.3	72.6	84.2	80.4	58.5	77.0	90.1
average	92.77	84.54	84.83	82.49	91.73	82.14	93.32	85.06	83.58	94.51

表4　采样算法采样后UCI数据集的MCC值

数据集	original	欠采样			过采样			混合采样		
		SM	BSM	ADS	ENN	TL	IHT	SM-ENN	SM-TL	CSE
spambase	85.3	87.6	87.7	86.3	90.0	95.0	86.3	93.5	86.3	95.7
abalone	/	77.8	82.0	80.0	62.3	/	98.7	88.9	79.0	95.6
eighthr	33.0	79.5	87.9	86.4	37.6	32.6	97.5	86.2	80.9	89.4
diabetes	41.7	53.4	53.0	51.2	68.9	54.2	82.1	74.4	51.5	92.2
balance	/	69.2	75.4	70.5	/	/	51.0	77.5	73.6	79.0
wdbc	85.5	88.2	91.3	90.2	88.6	89.8	94.3	92.2	88.7	95.8
ionosphere	81.3	81.3	78.6	80.3	77.9	81.3	84.2	81.4	81.3	92.5
haberman	19.5	41.6	45.0	39.1	50.5	35.6	75.2	73.3	44.2	92.3
wpbc	31.1	60.2	60.2	57.0	25.0	38.6	67.5	52.9	60.9	79.9
average	53.91	70.26	71.96	70.07	62.64	61.01	83.87	79.13	70.54	91.11

SMOTE+ENN在大部分数据集上的3个性能指标都很高,过采样中ADASYN算法的各项性能指标均优于其他采样算法,ADASYN算法在SMOTE算法的基础上进行了改进,明显减少了噪声样本的合成. 欠采样中的IHT算法则相对优于ENN和Tomek Links算法. 通过纵向对比3类方法可知,混合采样和过采样相对优于欠采样算法,而进行多次ENN欠采样可以有效地删除样本中的噪声样本.

3.3.3 集成方法对比实验

通过与传统采样算法对比可知,CSE算法取得了更好的效果. 然而以C4.5算法为评估准则的混合采样算法可能存在一定的局限性,即采样后的数据可能只适用于C4.5决策树. 因此,本文使用另外3种决策树算法进行对比验证. 此外,实验中还对比了Bagging方法对决策树进行集成后的效果. 4种决策树算法在9个UCI数据集上3种性能指标的结果如表5和表6所示,最后一行为算法在所有数据集上的平均值.

由表5可知,在原始UCI数据集上,4种决策树以及Bagging算法的分类性能都是较差的. 如在balance数据集上,所有算法的MCC值都是不存在的,因此数据的不平衡严重影响了算法的分类性能. 观察eighthr和wpbc数据集可知,虽然Bagging方法可以有效地提升不稳定算法的分类性能,但对于不平衡数据而言,数据的不平衡度越高,Bagging方法的分类性能反而越差. 此外,通过对比9个UCI数据集的平均值可知,C4.5算法的分类性能要好于其他算法,这也是本文选取C4.5算法为分类算法的重要原因.

由表6可知,经过CSE算法采样后的数据,4种决策树算法以及Bagging方法的分类性能均有显著提升. 未进行集成学习前,C4.5算法在大部分数据集上都取得了很好的效果,平均值为90.27%;而经过Bagging集成后的分类结果是REPTree算法,可知. 选取C4.5算法作为混合采样的评估指标合成后的样本集同样适用于其他分类算法. 此外,经过混合采样后的数据集在C4.5算法以及Bagging方法上有了显著提升,分别为27.66%和25.63%,在另外3种决策树算法上具有同样的结果.

3.3.4 医学数据诊断预测

医学数据中存在严重的样本不平衡性,正常样本往往是病例样本的几倍甚至数十倍,因此,医学数据是一种典型的不平衡数据集. 本文选取某医院近一年孕妇的甲状腺功能体检样本作为训练样本,通过专家经验和特征选择选取甲状腺功能异常、临床甲减、亚临床甲减、甲状腺抗体阳性和单纯甲状腺抗体阳性等8个离散属性的稽留流产数据集. 使用采样算法对稽留流产数据进行采样,结果如表7所示.

表 5 Bagging方法在原始UCI数据集的MCC值

数据集	未使用Bagging				使用Bagging			
	C4.5	REPTree	Randomtree	Hoeffdingtree	C4.5	REPTree	Randomtree	Hoeffdingtree
spambase	85.3	85.1	81.1	55.8	87.8	87.4	88.3	67.2
abalone	/	/	18.9	/	6.10	11.4	14.7	/
eighthr	33.0	18.8	26.8	/	30.4	29.6	33.4	/
diabetes	41.7	44.4	31.9	46.4	43.3	44.1	43.0	46.5
balance	/	/	/	/	/	/	/	/
wdbc	85.5	83.4	85.7	84.5	89.5	89.5	91.4	84.5
ionosphere	81.3	76.9	70.8	74.1	85.1	80.7	82.7	82.5
haberman	19.5	6.40	18.2	24.6	28.2	26.4	8.40	18.9
wpbc	31.1	15.4	18.7	/	28.5	30.2	31.4	/
average	62.66	59.24	57.54	57.08	66.78	65.62	62.76	59.92

表 6 Bagging方法在CSE算法采样后UCI数据集的MCC值

数据集	未使用Bagging				使用Bagging			
	C4.5	REPTree	Randomtree	Hoeffdingtree	C4.5	REPTree	Randomtree	Hoeffdingtree
spambase	95.7	93.7	93.1	76.0	97.1	96.1	97.4	77.1
abalone	95.6	93.9	94.5	83.1	96.8	95.5	97.4	85.4
eighthr	89.4	87.3	88.1	76.0	93.9	92.3	94.2	77.5
diabetes	92.2	90.1	90.6	87.1	92.2	91.6	93.9	87.3
balance	79.0	74.0	85.5	17.4	86.0	87.9	92.8	19.0
wdbc	95.8	95.0	96.7	94.4	97.5	95.0	97.2	94.4
ionosphere	92.5	83.3	88.7	83.7	95.0	93.3	97.1	84.6
haberman	92.3	87.4	91.0	60.3	94.6	91.1	95.1	60.7
wpbc	79.9	70.8	76.1	46.8	78.6	77.9	88.6	46.1
average	90.27	86.17	89.37	69.42	92.41	91.19	94.86	70.23

表 7　采样算法采样后稽留流产数据集的分类性能指标

性能指标	original	欠采样			过采样			混合采样		
		SM	BSM	ADS	ENN	TL	IHT	SM-ENN	SM-TL	CSE
Sensitivity	46.4	64.1	70.7	61.9	57.1	46.4	82.1	89.8	64.7	98.5
Specificity	90.8	82.3	80.7	79.1	94.1	90.8	98.2	100.0	83.9	100.0
MCC	42.4	47.2	51.7	41.7	57.3	42.4	81.4	92.7	49.5	98.5

表 8　Bagging方法在稽留流产数据集的MCC值

数据集	未使用Bagging				使用Bagging			
	C4.5	REPTree	Randomtree	Hoeffdingtree	C4.5	REPTree	Randomtree	Hoeffdingtree
original	42.4	40.8	40.6	30.0	38.3	40.1	42.1	30.3
CSE	98.5	97.3	99.2	64.0	98.1	98.1	99.2	64.1

由表7可知,CSE算法和传统采样算法的3种性能指标都普遍高于对原始数据集进行分类的C4.5算法. 对比CSE算法和其他采样算法可知,CSE算法采样后的稽留流产数据具有更高的分类精度,尤其是Specificity指标提升的幅度最大. 与在UCI数据集上的对比实验一样,混合采样算法的性能优于其他采样算法,SMOTE+ENN的MCC值为92.7%. 对比CSE算法和SMOTE+ENN算法可知,CSE算法显著提升了预测稽留流产检测的准确性.

Bagging方法在稽留流产数据上的对比如表8所示. 由表8可知,经过CSE算法处理后的稽留流产数据在4种决策树算法上的MCC值都有了显著提升. 与决策树算法相比,Bagging方法提升的幅度较小,这是因为Bagging方法能够显著提升不稳定分类算法的分类性能,虽然决策树是不稳定的学习算法,但经过CSE算法处理后的激流数据变得较为稳定.

4　结　论

医学数据挖掘是数据挖掘领域的重要研究方向,对于发现疾病发病规律、疾病危险因素以及寻找治疗方案具有十分重要的研究价值. 传统的决策树算法以总体分类精度为优化目标,往往会倾向于比例较大的类,从而淹没了少数类样本的信息. 然而,医学数据是不平衡的,鉴于此,本文提出了一种基于C4.5算法的混合采样算法(CSE). 该算法主要分为两个部分,首先对不平衡数据进行过采样,使用SMOTE算法合成少数类样本;然后使用ENN算法对合成的噪声样本进行删除. 以C4.5算法的分类性能作为混合采样的迭代停止准则. 此外,针对决策树存在的不稳定现象,提出使用Bagging算法对C4.5算法进行集成,通过在多个UCI数据集上的对比验证得出所提出混合采样算法的有效性和优越性. 最后,将所提出的算法应用于真实医学数据集,取得了较好的效果.

参考文献(References)

[1] Alinejad-Rokny H, Sadroddiny E, Scaria V. Machine learning and data mining techniques for medical complex data analysis[J]. Neurocomputing, 2018, 276: 1.

[2] Lee S J, Xu Z Z, Li T, et al. A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making[J]. Journal of Biomedical Informatics, 2018, 78: 144-155.

[3] Siers M J, Islam M Z. Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem[J]. Information Systems, 2015, 51: 62-71.

[4] 李艳霞, 柴毅, 胡友强,等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34(4): 673-688.
(Li Y X, Chai Y, Hu Y Q, et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019, 34(4): 673-688.)

[5] García V, Sánchez J S, Mollineda R A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance[J]. Knowledge-Based Systems, 2012, 25(1): 13-21.

[6] Wilson D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1972, 2(3): 408-421.

[7] Kang Q, Chen X S, Li S S, et al. A noise-filtered under-sampling scheme for imbalanced classification[J]. IEEE Transactions on Cybernetics, 2016, 47(12): 4263-4274.

[8] Tomek I. Two modifications of CNN[J]. IEEE Transactions Systems, Man, and Cybernetics, 1976, 6(11): 769-772.

[9] 魏力, 张育平. 一种改进型的不平衡数据欠采样算法[J]. 小型微型计算机系统, 2019, 40(5): 1094-1098.
(Wei L, Zhang Y P. Improved under-sampling algorithm for imbalanced data[J]. Journal of Chinese Computer Systems, 2019, 40(5): 1094-1098.)

[10] Lin W C, Tsai C F, Hu Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409/410: 17-26.

[11] 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(1): 321-357.

[12] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. DBSMOTE: Density-based synthetic minority over-sampling technique[J]. Applied Intelligence, 2012, 36(3): 664-684.

[13] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]. International Conference on Intelligent Computing. Berlin: Springer Heidelberg, 2005: 878-887.

[14] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C] IEEE International Joint Conference on Neural Networks. Hong Kong: IEEE, 2008: 1322-1328.

[15] 黄海松, 魏建安, 康佩栋. 基于不平衡数据样本特性的新型过采样SVM分类算法[J]. 控制与决策, 2018, 33(9): 1549-1558.
(Huang H S, Wei J A, Kang P D. New over-sampling SVM classification algorithm based on unbalanced data sample characteristics[J]. Control and Decision, 2018, 33(9): 1549-1558.)

[16] Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information Sciences, 2018, 465: 1-20.

[17] Batista G E A P A, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[18] 陶新民, 童智靖, 刘玉, 等. 基于ODR和BSMOTE结合的不均衡数据SVM分类算法[J]. 控制与决策, 2011, 26(10): 1535-1541.
(Tao X M, Tong Z J, Liu Y, et al. SVM classifier for unbalanced data based on combination of ODR and BSMOTE[J]. Control and Decision, 2011, 26(10): 1535-1541.)

[19] Li J Y, Fong S, Wong R K, et al. Adaptive multi-objective swarm fusion for imbalanced data classification[J]. Information Fusion, 2018, 39: 1-24.

[20] Zhang X, Zhu C, Wu H G, et al. An imbalance compensation framework for background subtraction[J]. IEEE Transactions on Multimedia, 2017, 19(11): 2425-2438.

[21] Peng M L, Zhang Q, Xing X Y, et al. Trainable undersampling for class-imbalance learning[J]. AAAI Conference on Artificial Intelligence, 2019, 33: 4707-4714.

[22] Mirza S, Mittal S, Zaman M. Decision support predictive model for prognosis of diabetes using SMOTE and Decision tree[J]. International Journal of Applied Engineering Research, 2018, 13(11): 9277-9282.

[23] 徐鹏, 林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报, 2009, 20(10): 2692-2704.
(Xu P, Lin S. Internet traffic classification using C4.5 decision tree[J]. Journal of Software, 2009, 20(10): 2692-2704.)

[24] 张元鸣,陈苗,陆佳炜, 等. 基于MapReduce的Bagging决策树优化算法[J]. 计算机工程与科学, 2017, 39(5): 841-848.
(Zhang Y M, Chen M, Lu J W, et al. An optimized bagging decision tree algorithm based on MapReduce[J]. Computer Engineering and Science, 2017, 39(5): 841-848.)

[25] 许召召, 李京华, 陈同林, 等. 融合SMOTE与Filter-Wrapper的朴素贝叶斯决策树算法及其应用[J]. 计算机科学, 2018, 45(9): 65-69.
(Xu Z Z, Li C H, Chen T L, et al. Naive bayesian decision treealgorithm combining SMOTE and filter-wrapper and It's application[J]. Computer Science, 2018, 45(9): 65-69.)

[26] 冯宏伟, 姚博, 高原, 等. 基于边界混合采样的非均衡数据处理算法[J]. 控制与决策, 2017, 32(10): 1831-1836.
(Feng H W, Yao B, Gao Y, et al. Imbalanced data processing algorithm based on boundary mixed sampling[J]. Control and Decision, 2017, 32(10): 1831-1836.)

[27] Tsoumakas G, Katakis I. Multi-label classification: An overview[J]. International Journal of Data Warehousing and Mining (IJDWM), 2007, 3(3): 1-13.

[28] Fei H, Hou J, Wu Z H, et al. Plasma metabolomic profile and potential biomarkers for missed abortion[J]. Biomedical Chromatography, 2016, 30(12): 1942-1952.

[29] 张立岩, 周晓, 刘爱红. 甲状腺功能异常与稽留流产的相关性研究[J]. 国际妇产科学杂志, 2015, 42(2): 207-208.
(Zhang L Y, Zhou X, Liu A H. Exploration of the relationship between thyroid dysfunction and missed abortion[J]. International Journal of Obstetrics, 2015, 42(2): 207-208.)

[30] Smith M R , Martinez T , Giraud-Carrier C . An instance level analysis of data complexity[J]. Machine Learning, 2014, 95(2): 225-256.

作者简介

许召召 (1991−), 男, 博士生, 从事数据挖掘、机器学习等研究, E-mail: zhaozhaotoms@foxmail.com;

申德荣 (1964−), 女, 教授, 博士生导师, 从事分布式数据管理、数据集成等研究, E-mail: shenderong@cse.neu.edu.cn;

聂铁峥 (1980−), 男, 副教授, 博士, 从事数据质量、数据集成等研究, E-mail: nietiezheng@cse.neu.edu.cn;

寇月 (1980−), 女, 副教授, 博士, 从事实体搜索、数据挖掘等研究, E-mail: kouyue@cse.neu.edu.cn.

(责任编辑：郑晓蕾)