

不平衡数据处理方法对中药不良反应预测的应用研究^{*}

吴东苑¹, 杨伟^{1,2**}, 唐进法^{3**}, 李学林³, 王晓艳³, 刘红梅⁴, 易丹辉⁵

(1. 中央民族大学理学院 北京 100081; 2. 中国中医科学院中医临床基础医学研究所 北京 100700; 3. 河南中医药大学第一附属医院 郑州 450000; 4. 国家康复辅具研究中心 附属康复医院 北京 100176; 5. 中国人民大学统计学院 北京 100872)

摘要:目的:针对中药不良反应数据的不平衡性,探索并应用不平衡数据的处理方法,对中药的不良反应进行预测。本文以使用丹红注射液的患者为研究对象,对来自37家医院集中监测数据进行深度挖掘,在使用了丹红注射液的患者中预测是否发生不良反应。方法:从数据层面采用四种方法:不处理、随机欠采样、随机过采样、SMOTE采样;从算法层面采用四种模型或算法:决策树、随机森林、AdaBoost算法、Gradient Boosting算法,对数据的不平衡性进行处理。两个层面的方法两两结合,对16种方法与模型或算法组合的预测效果进行比较。结果:随机欠采样和AdaBoost算法相结合、随机欠采样和Gradient Boosting算法相结合的预测效果较为理想,recall和G-mean都达到80%以上,AUC指标也高达0.86。结论:初步探索中药不良反应可能适用的不平衡数据处理方法,预测结果结合实际经验,能较准确地预测使用了丹红注射液的患者是否发生不良反应,在临床实际应用中能起到一定的警示作用。同时,根据输出的变量重要性排名,能最大程度地避免用药后的不良反应的发生,为丹红注射液的安全性再评价提供一些科学参考依据。

关键词:不平衡数据 不良反应 集中监测数据 采样 boosting

doi:10.11842/wst.2017.09.008 中图分类号:R33 文献标识码:A

中药注射剂是一种不仅见效快,而且拥有传统中药治疗特色的制剂,它能够对急危重症患者发挥非常重要的治疗作用^[1]。然而这几年,随着中药注射剂的普遍使用,被发现的药品不良反应记录也在逐渐增多^[2]。由于临床应用中更关注中药不良反应的预测,而这方面的研究还很不足,特别是利用来自医院集中监测的观察性数据对不良反应的预测研究涉及更少,本文以丹红注射液为例进行不良反应预测的应用研究,在一定程度上弥补这方面的空白。除此之外,本文

针对中药不良反应数据中,相比庞大的未发生不良反应患者数量,发生不良反应的患者还是占少数的情况,采用不平衡数据处理方法进行不良反应预测的应用研究。通常我们把仅有极少的少数类样本的数据称为不平衡数据^[3],此时,若采用一般的分类学习算法训练预测模型,则模型很难获得让人满意的预测精度^[4]。

丹红注射液是以丹参、红花作为主要提取物的一种中药注射剂,它具有调节血管张力,改善脏器血液循环的作用^[5],随着其在临床上的普遍使用,发生不良反应的病例数正在逐渐上升^[6],丹红注射液的安全性再评价研究也在日益丰富。目前,从统计学角度出发的

收稿日期:2017-05-13

修回日期:2017-08-25

* 国家自然科学基金委青年科学基金项目(81502898):大型观察性医学数据的因果图模型研究,负责人:杨伟;重大新药创制专项子课题(2015ZX09501004-001-007):临床需长期使用的中药口服制剂安全性监测研究,负责人:李学林。

** 通讯作者:杨伟,统计学博士,助理研究员,主要研究方向:大规模观察性数据分析方法及因果推断研究;唐进法,医学博士,副主任药师,主要研究方向:中药合理用药。

表1 混淆矩阵

	预测为正类	预测为负类
实际为正类	TP	FN
实际为负类	FP	TN

丹红注射液不良反应研究尚处于萌芽阶段,多采用描述性统计^[7]、显著性检验^[8]和 Logistic 回归^[9]等传统统计方法对不良反应情况进行描述与分析,尚未进行预测方面的研究。本文使用丹红注射液人群中不良反应发生的数据就呈现不平衡特点,我们将初步探索处理不良反应不平衡数据的方法,并进行不良反应的预测,且分析引起不良反应发生的可能影响因素,这对正确探讨丹红注射液临床安全用药的潜在规律,以及科学地评估中药安全性和临床合理应用提供方法学基础。

1 方法原理

由于使用中药的患者中发生不良反应的人群数量过于稀少,为了达到较为理想的预测效果,本文探索并应用不平衡数据处理方法对其进行预测。处理不平衡数据,一般从两大层面入手:数据层面和算法层面。

在数据层面上,主要依照一定的规律对数据进行重新采样,使其达到平衡。本研究采用如今较为流行几种重新采样方法。随机欠采样是对使用中药患者中未发生不良反应的病例进行简单随机抽样,使其与发生不良反应的病例数持平;随机过采样是对使用中药患者中发生不良反应的病例进行简单随机复制,使其数量扩增至与未发生不良反应的病例数相当;SMOTE^[10]是针对每一个使用中药并发生了不良反应的患者,根据其最近邻的分布,随机生成一些模拟的不良反应样本,使其数量与未发生不良反应的病例数相当。

在算法层面上,一般是将一些传统预测方法与能够适应不平衡数据分布的算法相结合,形成能够处理不平衡数据的预测方法。本研究采用常见的决策树、随机森林^[11],并以决策树为基础结合 AdaBoost 算法^[12]和 Gradient Boosting 算法^[13]对丹红注射液的不良反应进行预测。AdaBoost 算法和 Gradient Boosting 算法都是一种集成学习算法,不同之处在于:AdaBoost 算法是在每一轮迭代中,对被分类预测错误的样本进行加权,以此更新原有样本的分布情况,然后结合新分布下的样本重新构建一个新模型并得到它的对应权重,最后将所有模型进行线性组合;Gradient Boosting 则是在每一次迭代中计算出一个残差,每次迭代新的模型,都是为了让上次迭代构造模型的残差能够往梯度方向上升

低,最后将所有模型进行线性组合。

针对不平衡数据分类问题,常用的评价标准有 *recall* (又称 *sensitivity*)、*precision*、*specificity*、*F-value*、*G-mean* 等,具体公式如下:

$$recall = sensitivity = TP / (TP + FN)$$

$$precision = TP / (TP + FP)$$

$$specificity = TN / (TN + FP)$$

$$F-value = \frac{(1 + \beta^2) \times recall \times precision}{\beta^2 \times recall + precision}$$

$$G-mean = \sqrt{sensitivity \times specificity}$$

可以知道,求取少数类的 *F-value* 是一种非常有效的评价准则,它是 *recall* 和 *precision* 的调和均值,其中的 β 是可调参数,用于反映 *recall* 和 *precision* 的相对重要度,通常情况下取 1。如果同时关注多数类样本和少数类样本的分类性能,可以使用 *G-mean* 来度量效果,它能度量两个类别的平均分类性能。

除此之外,ROC 曲线能够全面地展示分类器在不同阈值下的分类性能,它是令纵坐标 $y = TP / (TP + FN)$,表示少数类被分对的比例,横坐标 $x = FP / (TN + FP)$,表示多数类被分错的比例。一个分类算法的 ROC 曲线如果越靠近坐标平面的左上角,说明这个分类算法的分类效果越好。与此同时,为了对分类性能进行更好地定量分析,可以采用 ROC 曲线下的面积 AUC 来对分类性能进行评估,一个分类算法的 AUC 值越接近于 1,表明这个分类算法的分类效果越理想^[14]。

2 材料与方法

2.1 数据来源

本研究采用丹红注射液患者医院集中监测数据,该数据由 6 省市 37 家医院参与研究,监测医院从 2009 年 4 月至 2013 年 8 月所有使用丹红注射液的住院患者,共计纳入有效病例数 30888 例,年龄跨度为 0 到 99 岁。其中发生不良反应 108 例,未发生不良反应 30780 例,数据呈现极度不平衡性特点。

2.2 特征初筛

原始数据集包含患者基本信息、病症情况、给药情况、综合情况这四大类信息,共有变量 1834 个。其中患者基本信息包含年龄、性别、体重指数、怀疑过敏物等 78 个变量,病症情况包含适应病症、是否中医辨证等 671 个变量,给药情况包含是否首次使用丹红注射液、用药次数、合并用药名称等 970 个变量,综合情况

包含病情变化情况、症状改善情况等 115 个变量。由于原始数据集存在大量特征变量,需要在模型构建之前对特征变量进行初步筛选。采用三种方法综合选取变量:组间比较、特征选择、临床经验判断。

2.2.1 组间比较

将发生不良反应的病例样本归为不良反应组,将未发生不良反应的病例样本归为正常组,针对每一个变量进行组间比较,查看在各个变量上两组别之间是

否有显著差异。最后发现有 14 个变量存在组间差异,分别为性别、体重指数、有无个人过敏史、有无家族过敏史、有无过敏疾病史、是否中医辨证、是否首次使用丹红注射液、用药次数、最后一次用药时间、最后一次静滴速度、病情变化情况、症状改善情况、中医证候疗效判定、体征变化情况。

2.2.2 特征选择

特征选择是从原始数据集的所有特征中,通过选

表 2 14 个存在组间差异的变量情况

变量		[ALL] N=30887	不良反应组 N=108	正常组 N=30779	P 值
性别	男	16883 (54.66%)	37 (34.26%)	16846 (54.73%)	<0.001
	女	14004 (45.34%)	71 (65.74%)	13933 (45.27%)	
体重指数		23.94 [21.80,26.00]	23.50 [20.70,25.50]	23.94 [21.80,26.00]	0.04
个人药物、食物等过敏史	否	25960 (84.06%)	71 (65.74%)	25889 (84.12%)	<0.001
	是	4923 (15.94%)	37 (34.26%)	4886 (15.88%)	
家族过敏史	否	29058 (94.08%)	98 (90.74%)	28960 (94.10%)	0.035
	是	97 (0.31%)	2 (1.85%)	95 (0.31%)	
	不详	1730 (5.60%)	8 (7.41%)	1722 (5.60%)	
过敏性疾病史	否	30620 (99.14%)	102 (94.44%)	30518 (99.16%)	<0.001
	是	265 (0.86%)	6 (5.56%)	259 (0.84%)	
是否中医辨证	否	17971 (59.32%)	44 (40.74%)	17927 (59.39%)	<0.001
	是	12323 (40.68%)	64 (59.26%)	12259 (40.61%)	
是否首次使用丹红注射液	否	4578 (14.86%)	16 (14.81%)	4562 (14.86%)	0.002
	是	15890 (51.57%)	72 (66.67%)	15818 (51.52%)	
	不详	10342 (33.57%)	20 (18.52%)	10322 (33.62%)	
用药次数	1 次	367 (1.19%)	27 (25.23%)	340 (1.11%)	<0.001
	2-5 次	4680 (15.16%)	53 (49.53%)	4627 (15.04%)	
	6-10 次	9698 (31.41%)	12 (11.21%)	9686 (31.48%)	
	11-20 次	15157 (49.09%)	14 (13.08%)	15143 (49.22%)	
	21-30 次	809 (2.62%)	1 (0.93%)	808 (2.63%)	
	31-40 次	122 (0.40%)	0 (0.00%)	122 (0.40%)	
	40 次以上	40 (0.13%)	0 (0.00%)	40 (0.13%)	
最后一次用药时间		90.00 [75.00,112.00]	85.00 [71.00,100.00]	90.00 [75.00,112.00]	0.006
最后一次静滴速度		51.00 [40.00,64.00]	60.00 [50.00,70.50]	51.00 [40.00,64.00]	<0.001
病情变化情况	痊愈	2010 (6.67%)	1 (0.93%)	2009 (6.69%)	<0.001
	好转	27027 (89.64%)	62 (57.94%)	26965 (89.75%)	
	无效	1052 (3.49%)	43 (40.19%)	1009 (3.36%)	
	加重	15 (0.05%)	1 (0.93%)	14 (0.05%)	
	死亡	46 (0.15%)	0 (0.00%)	46 (0.15%)	
症状改善情况	显著改善	3678 (12.06%)	3 (2.83%)	3675 (12.09%)	<0.001
	改善	25744 (84.39%)	61 (57.55%)	25683 (84.48%)	
	无效	1069 (3.50%)	42 (39.62%)	1027 (3.38%)	
	加重	15 (0.05%)	0 (0.00%)	15 (0.05%)	
中医证候疗效判定	无	18244 (61.44%)	53 (50.48%)	18191 (61.47%)	0.027
	有	11452 (38.56%)	52 (49.52%)	11400 (38.53%)	
体征变化情况	显著改善	3506 (11.96%)	3 (2.86%)	3503 (11.99%)	<0.001
	改善	24791 (84.57%)	60 (57.14%)	24731 (84.67%)	
	无效	999 (3.41%)	40 (38.10%)	959 (3.28%)	
	加重	18 (0.06%)	2 (1.90%)	16 (0.05%)	

表3 数据集样本量

	不良反应组	正常组
原始训练集	75	18196
测试集	18	4549

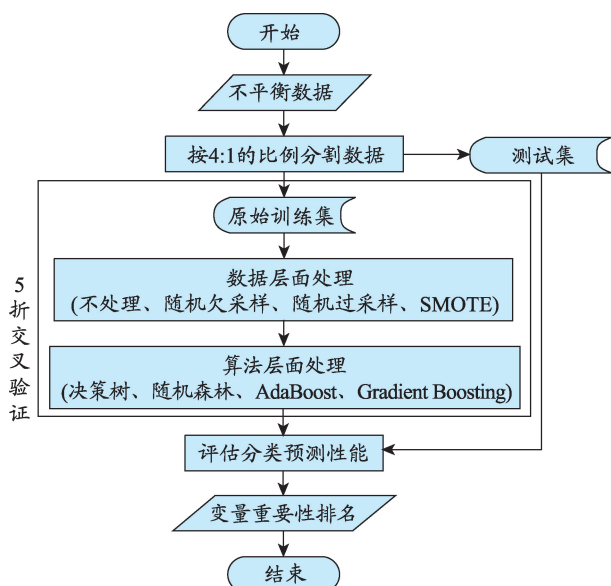


图1 建模流程图

出一些分类效果最明显的特征来降低数据集的维度，这也是提高算法的分类性能的一个重要方法。

分别从患者基本信息、给药情况、综合情况三个方面进行特征选择，为选出最佳的特征子集，使用随机森林进行测试，设置20次迭代，结果为：(1)患者基本信息：体重、年龄、性别、有无过敏疾病史的各变量；(2)给药情况：是否首次使用丹红注射液、用药次数、最后一次用药时间、最后一次静滴速度、单次给药量；(3)综合情况：病情变化情况、症状改善情况、中医证候疗效判定、体征变化情况、是否存在不合理用药。

2.2.3 临床经验判断

除原始数据集包含的变量以外，在数据预处理的过程中还将怀疑过敏物、适应病症、合并化药、合并中成药等变量进行哑变量处理(取值0和1)，得到四类数据集，基于相关临床经验判断，由于更关注不良反应组的过敏、病证及用药情况，故保留这四类数据集中不良反应组的发生频数大于0的变量，其中有19种怀疑过敏物，42种适应病症，138种化药名称，35种中成药名称。

2.2.4 初筛结果

本研究对象为所有使用丹红注射液的患者，是否发生不良反应作为结局变量，取值0和1，0表示未发生

不良反应，1表示发生不良反应；根据初筛特征变量的结果，最终放入预测模型进行训练的解变量共有252个，包括如下四大方面的变量：(1)患者基本信息：年龄、性别、体重指数、有无个人过敏史、有无家族过敏史、有无过敏疾病史、不良反应组发生频数大于0的怀疑过敏物(共19个)；(2)病证情况：不良反应组发生频数大于0的适应病症(共42个)、是否中医辨证；(3)给药情况：是否首次使用丹红注射液、用药次数、最后一次用药时间、最后一次静滴速度、单次给药量、合并用药种类数、不良反应组发生频数大于0的合并化药(共138个)、不良反应组发生频数大于0的合并中成药(共35个)；(4)综合情况：病情变化情况、症状改善情况、中医证候疗效判定、体征变化情况、是否存在不合理用药。

2.3 训练集与测试集的划分

经过特征初筛后，还需对数据进行剔除缺失值的操作，若一个样本中有变量存在缺失值，就将这个样本直接剔除。最终整理得到的新数据集共有22838个样本，其中不良反应组有93个样本，正常组有22745个样本。在不良反应组和正常组中分别随机抽取80%的样本组成训练集，余下20%的样本组成测试集。得到的数据集样本量如表3。

2.4 模型构建

在建模过程中，首先对原始训练集进行数据层面的处理，即采用不处理、随机欠采样、随机过采样和SMOTE算法这四种处理方法进行处理。不处理的训练集即为原始训练集(共有18271个样本)；随机欠采样，即是在原始训练集中的18196例正常样本中随机抽取75个，与不良反应样本数持平，将其记为训练集A(共150个样本)；随机过采样，即是对原始训练集中的75例不良反应样本进行重复随机抽样，不断抽取直到达到和正常样本数相同的18196个不良反应样本，将其记为训练集B(共36392个样本)；SMOTE算法，则是取最近邻数 $k=70$ ，反复生成新的模拟的不良反应样本，直至其样本数与正常样本数接近1:1，将其记为训练集C(共31670个样本)。

接着，分别对原始训练集、训练集1、训练集2和训练集3采用5折交叉验证法进行算法层面的处理，建立决策树、随机森林、以决策树为弱分类器的AdaBoost算法和以决策树为弱分类器的Gradient Boosting算法这四种模型。

最后，利用测试集对总共16种算法组合进行分类

预测,查看 *recall*、*precision*、*F-value*、*G-mean*、ROC 曲线以及 AUC 值,综合这些评价准则,比较各算法组合的分类预测效果。

3 结果

本研究利用 R 3.3.2 中的 DMwR 包、rpart 包和 caret 包等对丹红注射液不良反应监测数据进行分类预测,各个算法的分类结果比较如表 4 所示。

从 *recall* 来看,在数据层面对不平衡数据的处理方法中,不处理的分类效果不出所料是最差的,但随机欠采样的效果反而比随机过采样和 SMOTE 算法好。这可能是因为数据集样本量多,欠采样不会丢失太多重要信息,反而比容易产生过拟合的过采样方法效果要好。

从 *precision* 来看,没有特别出众的算法搭配,不过查准率展现的是在被预测为正类的样本中实际为正类的比率,从临床医学上来说,人们更关注的是发生不良反应的患者能否被准确分类预测,因此在两者不可兼得的情况下,该指标的重要性不及查全率。

F-value 作为查全率和查准率的调和均值,取值会偏向于两者较低的一端,因此效果同查准率一样有好有坏,当分类效果中查全率与查准率不可兼得时,在临床医学领域的参考意义同样不大。

G-mean 是反映少数类识别率和多数类识别率的综合指标,可以发现随机欠采样与两种 boosting 算法的组合,以及决策树与随机欠采样、随机过采样、SMOTE 算法的组合效果都比较好,能够达到 80% 以上。

根据表 3 中所列的评价准则,主要利用查全率和 *G-mean* 对分类性能进行评估判断,得出初步结论:随机欠采样和 AdaBoost 算法组合的分类效果最好,查全率高达 88.89%, *G-mean* 值高达 85.74%;其次是随机过采样和决策树组合、随机欠采样和 Gradient Boosting 组合效果较好,查全率和 *G-mean* 值也都能达到 80% 以上;再次一些的是 SMOTE 算法和决策树组合,这一组合虽然 *G-mean* 值有 82.19%,但查全率仅维持在 72.22%,分类性能还算可以;其余算法组合的分类结果都不是特别理想。

下面结合各算法组合的 ROC 曲线图(如图 2)与 AUC 指标(如表 5)进行进一步综合判断。

结合图 2 和表 5 可以发现,单从 ROC 曲线和 AUC 指标来说,决策树与数据层面的各种处理方法相结合的分类效果均不太理想;随机森林与随机过采样、

表 4 各算法分类结果比较

不平衡 数据处理	学习分类器	<i>recall</i>	<i>precision</i>	<i>F-value</i>	<i>G-mean</i>
不处理	决策树	0	NA	NA	0
	随机森林	0	NA	NA	0
	AdaBoost	16.67	100	28.58	40.83
	Gradient Boosting	11.11	100	20	33.33
随机 欠采样	决策树	83.33	1.52	2.99	80.94
	随机森林	94.44	0.38	0.76	8.91
	AdaBoost	88.89	1.99	3.89	85.74
	Gradient Boosting	83.33	2	3.91	83.59
随机 过采样	决策树	83.33	2.58	5.01	85.41
	随机森林	11.11	66.67	19.05	33.33
	AdaBoost	38.89	12.73	19.18	62.03
	Gradient Boosting	38.89	25.93	31.11	62.22
SMOTE 算法	决策树	72.22	4.23	7.99	82.19
	随机森林	33.33	6.82	11.32	57.21
	AdaBoost	16.67	17.65	17.15	40.77
	Gradient Boosting	16.67	33.33	22.22	40.8

SMOTE 算法这两种过采样处理方法相结合的分类效果较好;AdaBoost 算法分别与随机欠采样、随机过采样、SMOTE 算法组合的分类效果较好;Gradient Boosting 算法与随机欠采样和 SMOTE 算法的组合效果较为理想。

综上所述,随机欠采样和 AdaBoost 算法相结合、随机欠采样和 Gradient Boosting 算法相结合这两种算法组合的分类性能相对最好。

4 结论

通过比对 16 种不同的方法与算法组合,发现随机欠采样和 AdaBoost 算法、随机欠采样和 Gradient Boosting 算法这两种组合针对本研究使用的数据集分类性能相对较为理想,也由此得出了一些结论。

4.1 不平衡数据分类问题的总体性趋势

根据各算法组合的分类预测效果,可以在数据层面发现一些处理不平衡数据分类问题的总体性趋势:当数据不平衡时,对数据进行采样处理,使得不平衡数据趋于平衡,其分类预测效果整体要优于不进行不平衡数据处理的情况;另外,采用 SMOTE 算法处理不平衡数据,在与随机欠采样和随机过采样的处理方法的对比上,并不具有明显优势。除此之外,虽然得到两种算法组合的分类预测效果较好,但这样的算法组合运用在别的领域,甚至只是相同领域的不同数据集上时,

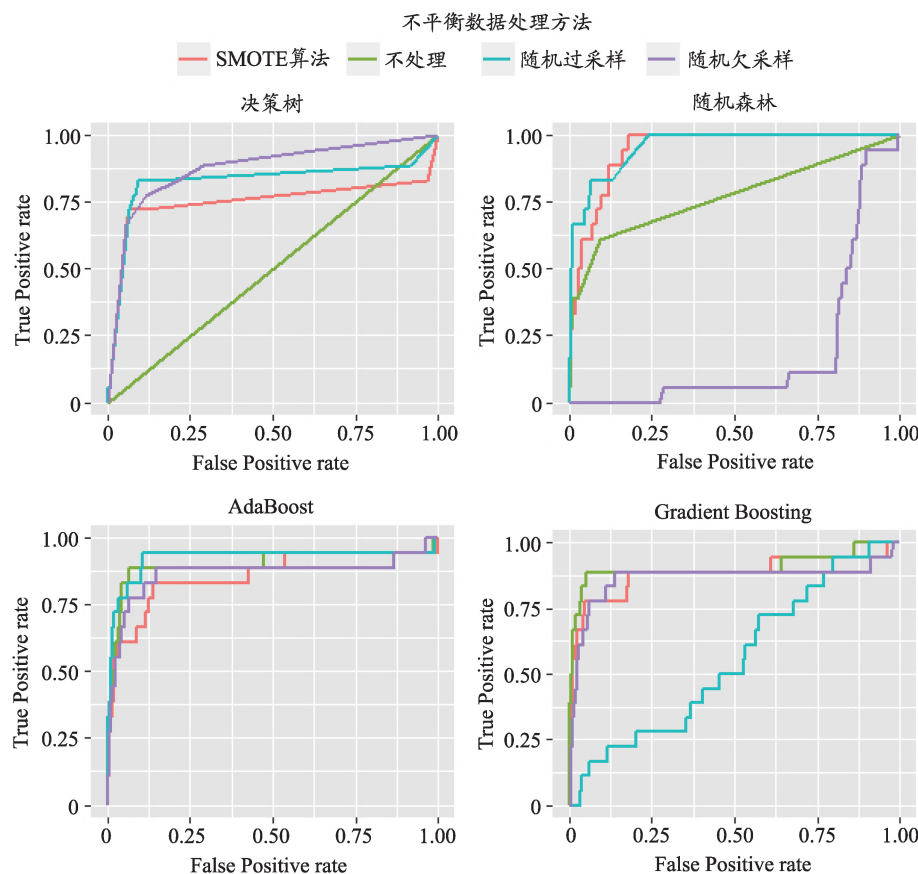


图2 各算法的ROC曲线图比较

表5 各算法的AUC指标比较

	不处理	随机欠采样	随机过采样	SMOTE算法
决策树	0.5	0.877	0.831	0.755
随机森林	0.772	0.187	0.956	0.944
AdaBoost	0.901	0.866	0.922	0.858
Gradient Boosting	0.906	0.863	0.55	0.882

分类效果还需要实际操作来证实。

4.2 临床上使用丹红注射液需注意的可能影响因素

本研究在对比算法组合预测效果的同时,也输出了最佳算法组合的变量重要性排名,以此从侧面挖掘一些丹红注射液不良反应的可能影响因素,能够在临床上提供一些用药的科学根据,医生结合实际经验,能更科学合理地使用丹红注射液为患者进行治疗,并最大程度避免用药后不良反应的发生。总得来说,为了

预先防范使用丹红注射液可能发生的不良反应,在用药之前,需要从患者的年龄、性别、体重指数、过敏物历史以及是否首次使用丹红注射液这些方面来考虑该患者是否适合使用丹红注射液,是否有很大可能发生不良反应;在用药过程中,医护人员也要把关闭用药次数、用药剂量、注射时间、静滴速度、与丹红注射液合并使用的药物名称以及患者的病情、体征变化情况,特别是要格外注意用药规范,不可出现不合理的用药情况。

综上所述,本研究虽采用了一些现今较为流行的不平衡数据处理方法,但没有结合其它更复杂的处理方法进行应用和比较,比如特征选择算法、代价敏感算法等,这还需更加深入的研究。另外,本研究只对不平衡数据处理方法进行了应用,并未对方法的适用性或改进条件进行深入探讨,这都将是今后研究的重点。

参考文献

1 张冰,吴嘉瑞. 关于中药注射剂安全性问题的思考. 临床药物治疗杂志, 2006, 4(6): 14-18.

2 周超凡. 中药注射剂不良反应的警示. 中国药物警戒, 2005, 2(2): 65-68.

- 3 Sun Y, Kamel M S, Wong A K C, *et al.* Cost-sensitive boosting for classification of imbalanced data. *Pattern Recogn*, 2007, 40(12): 3358–3378.
- 4 Yanmin S, Andrew K C, Mohamed S K. Classification Of Imbalanced Data: A Review. *Int J Pattern Recog*, 2009, 23(4): 687–719.
- 5 王保中, 曹茂荣, 邹薇红, 等. 丹红注射液的毒理学研究. 中国实用医药, 2008, 3(11): 27–29.
- 6 柳青, 雷招宝. 丹红注射液的不良反应及合理应用. 中成药, 2010, 32(11): 1964–1966.
- 7 陈倩, 易丹辉, 杨伟, 等. 基于 HIS“真实世界”的丹红注射液上市后临床应用分析. 中国中药杂志, 2011, 36(20): 2817–2820.
- 8 刘玉平. 丹红注射液的临床不良反应监测. 山西医药杂志, 2014(7): 820–822.
- 9 李春晓, 唐进法, 孟菲, 等. 基于巢式病例对照方法的丹红注射液安全性再评价研究. 中国中药杂志, 2012, 37(18): 2735–2738.
- 10 Chawla N V, Bowyer K W, Hall L O, *et al.* Smote: Synthetic Minority over-sampling Technique. *J Artif Intell Res*, 2002, 16(1): 321–357.
- 11 Breiman L. Random Forests. *Mach Learn*, 2001, 45(1): 5–32.
- 12 Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. European Conference on Computational Learning Theory. *Springer-Verlag*, 1997: 119–139.
- 13 Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat*, 2001, 29(5): 1189–1232.
- 14 Xue J H, Hall P. Why Does Rebalancing Class-Unbalanced Data Improve AUC for Linear Discriminant Analysis? *Ieee T Pattern Anal*, 2015, 37(5): 1109.

Application Research of Imbalanced Data Processing Methods on Prediction of Adverse Reactions of Traditional Chinese Medicine

Wu Dongyuan¹, Yang Wei^{1,2}, Tang Jinfa³, Li Xuelin³, Wang Xiaoyan³, Liu Hongmei⁴, Yi Danhui⁵

(1. College of Science, Minzu University of China, Beijing 100081, China;

2. Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China;

3. The First Affiliated Hospital of Henan University of Traditional Chinese Medicine, Zhengzhou 450000, China;

4. Rehabilitation Hospital, National Research Center for Rehabilitation Technical Aids, Beijing 100176, China;

5. School of Statistics, Renmin University of China, Beijing 100872, China)

Abstract: In view of the imbalance of the adverse reaction data of traditional Chinese medicine (TCM), this paper explored and applied the processing method of imbalanced data to predict adverse reactions of TCM. This paper took patients who used *Dan-Hong* (DH) injection as the research object, excavated centralized monitoring data from 37 hospitals, and predicted adverse reactions from patients who used DH injection. This paper combined four data-level approaches, including non-processing, random undersampling, random oversampling and SMOTE algorithm, with four algorithm-level approaches, including decision tree, random forest, AdaBoost and Gradient Boosting, to process the imbalanced data, and then to compare their prediction performance. Finally we found that two algorithms, combining random undersampling with AdaBoost, and combining random undersampling with Gradient Boosting, had better prediction performance than other algorithms. Their *recall* and *G-mean* both reached 80%; and AUC was more than 0.86. It was concluded that the imbalanced data processing methods were preliminary explored. This method is applicable to the prediction of TCM adverse reactions in combination with practical experiences. It can accurately predict whether adverse reactions occurred in patients who used DH injection. It can play a certain warning role in clinical practice. At the same time, according to the importance of the output variable ranking, we can minimize the occurrence of adverse reactions after treatment. It provided some scientific references for the safety reassessment of DH injection.

Keywords: Imbalanced data, adverse reaction, centralized monitoring data, sampling, boosting

(责任编辑:张娜娜,责任译审:王 晶)