· 急救大数据 ·

基于机器学习的院内消化道致命性再出血预测与指标筛选

魏子健[1] 李静[1] 李雪岩[2] 赵宇卓[3] 贾立静[3] 黎檀实[3]

[1]北京交通大学经济管理学院 100044;[2]北京联合大学管理学院 100101;[3]解放军总医院急诊科,北京 100853

通信作者:黎檀实,Email:lts301@163.com

【摘要】 目的 提出基于机器学习的院内消化道致命性再出血预测和指标筛选方法。**方法** 从解放军总医院急救数据库中提取确诊为消化道出血样本 728 例次,其中确定发生院内消化道致命性再出血患者 343 例次。提取、筛选得到相关生理或化验指标共计 64 项。在十折交叉验证的基础上,分别使用 Logistic 回归、以决策树为弱分类器的自适应增强(AdaBoost)算法、以决策树为弱分类器的 XGBoost 算法进行分类预测并对比;利用 XGBoost 算法进行序列特征前向搜索,以训练时迭代出的指标重要性进行筛选,并得到预测院内消化道致命性再出血的关键指标。**结果** Logistic 回归和基于决策树的 AdaBoost 算法、XGBoost 算法在各特征输入维度下均得到了较好的 F1.5 分数,其中 XGBoost 算法效果最好、评分最高,即能够尽可能找出更多的可能发生院内消化道致命性再出血的患者。通过 XGBoost 算法迭代结果得到了预测院内消化道致命性再出血的前 30 个重要性较高的指标,其中前 12 个关键指标迭代时 F1.5 分数达到峰值(0.893),分别为血红蛋白测定(Hb)、钙(CA)、红细胞计数(RBC)、平均血小板体积测定(MPV)、平均红细胞血红蛋白浓度(MCH)、收缩压(SBP)、血小板计数(PLT)、镁(MG)、淋巴细胞(LYM)、葡萄糖(GLU,血气分析)、葡萄糖(GLU,血生化)、舒张压(DBP)。**结论** Logistic 回归及基于决策树的 AdaBoost 算法和 XGBoost 算法都能达到预测院内消化道致命性再出血的预警目的,其中 XGBboost 算法更佳,并能得到 12 个关键指标。

【关键词】 消化道出血; 机器学习; 自适应增强算法; XGBoost 算法

基金项目:国家自然科学基金(81272060);国家自然科学基金青年科学基金(81701961, 71103014);北京市科技新星计划项目(XX2018019);解放军总医院医疗大数据科研项目(2017MBD-30)

DOI:10.3760/cma.j.issn.2095-4352.2019.03.020

Prediction and feature selection for fatal gastrointestinal bleeding recurrence in hospital via machine learning

Wei Zijian[1], Li Jing[1], Li Xueyan[2], Zhao Yuzhuo[3], Jia Lijing[3], Li Tanshi[3]

[1]Department of School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China; [2]Management School, Beijing Union University, Beijing 100101, China; [3]Department of Emergency, Chinese PLA General Hospital, Beijing 100853, China

Corresponding author: Li Tanshi, Email: lts301@163.com

【Abstract】 Objective To propose a method of prediction for fatal gastrointestinal bleeding recurrence in hospital and a method of feature selection via machine learning models. **Methods** 728 digestive tract hemorrhage samples were extracted from the first aid database of PLA General Hospital, and 343 patients among them were diagnosed as fatal gastrointestinal bleeding recurrence in hospital. A total of 64 physiological or laboratory indicators were extracted and screened. Based on the ten-fold cross-validation, Logistic regression, AdaBoost and XGBoost were used for classification prediction and comparison. XGBoost was used to search sequence features, and the key indicators for predicting fatal gastrointestinal bleeding recurrence in hospital were screened out according to the importance of the indicators during training. **Results** Logistic regression, AdaBoost and XGBoost all get better F1.5 score under each feature input dimension, among which XGBoost had the best effect and the highest score, which was able to identify as many patients as possible who might have fatal gastrointestinal bleeding recurrence in hospital. Through XGBoost iteration results, the Top 30 indicators with high importance for predicting fatal gastrointestinal bleeding recurrence in hospital were ranked. The F1.5 scores of the first 12 key indicators peaked at iteration (0.893), including hemoglobin (Hb), calcium (CA), red blood cell count (RBC), mean platelet volume (MPV), mean erythrocyte hemoglobin concentration (MCH), systolic blood pressure (SBP), platelet count (PLT), magnesium (MG), lymphocyte (LYM), glucose (GLU, blood gas analysis), glucose (GLU, blood biochemistry) and diastolic blood pressure (DBP). **Conclusions** Logistic regression, AdaBoost and XGBoost could achieve the purpose of early warning for predicting fatal gastrointestinal bleeding recurrence in hospital, and XGBoost is the most suitable. The 12 most important indicators were screened out by sequential forward selection.

【Key words】 Gastrointestinal hemorrhage; Machine learning; AdaBoost; XGBoost

Fund program: National Natural Science Foundation of China (81272060); National Natural Science Youth Foundation of China (81701961, 71103014); Beijing Municipal Science and Technology New Star Program (XX2018019); Big-date Research and Development Project of Chinese PLA General Hospital (2017MBD-30)

DOI: 10.3760/cma.j.issn.2095-4352.2019.03.020

因各种原因导致的消化道出血是急诊常见症，大多数情况下患者不仅发病时病情危重，在治疗后还极易发生再出血，据统计其发生率高达 50%[1]。医护人员通常对院内消化道致命性再出血的认知具有滞后性，当医护人员通过生命体征变化或血常规化验指标判断患者存在致命性再出血时，其可能已经处于低血容量休克状态。因此，通过数据分析方法确定患者院内消化道致命性再出血的关键核心指标，并提出一种可靠的院内消化道致命性再出血预测模型，对该类疾病的救治具有重大意义。

除应用传统的评分指标外，国内外学者应用机器学习方法进行疾病预测与指标挖掘得到了较好的结果。Chu 等[2]通过神经网络（ANN）、支持向量机（SVM）、K 近邻、线性判别分析（LDA）、随机森林、Logistic 回归等机器学习方法对消化道出血患者进行了风险预测；Cismondi 等[3]基于重症监护医学信息数据库Ⅱ（MIMIC Ⅱ）重症加强治疗病房（ICU）数据中 746 例消化道出血患者进行了特征选择与预测；Ayaru 等[4]提出了一种基于梯度提升 Gradient Boosting 的急性下消化道出血预测方法，并将其与多元 Logistic 回归进行了比较。现在种种新的机器学习算法被提出，如使用新算法与结构的深度神经网络，其非线性拟合能力比传统模型更强，将其应用在预测与指标筛选上能够得到更好的结果。

2018 年 3 月，解放军总医院主持发布了解放军总医院急救数据库，并成功举办了第一届急救大数据与 Datathon 活动[5]，进一步验证了机器学习与医疗行业的“跨学科合作”大有可为。在活动中，解放军总医院急救数据库展现出了数据体量大、种类多、规范化程度高等特点[6]，收录了较多消化道出血患者的就诊数据，记录了较多可用于预测的生命体征及化验指标，与本研究的院内消化道致命性再出血预测问题契合性较高。本研究基于解放军总医院急救数据库，应用 Logistic 回归和基于决策树的自适应增强（AdaBoost）算法、XGBoost 算法对院内消化道致命性再出血进行预测比较，旨在提出一种更具准确性的预测方法，并筛选出关键指标，以便提前采取医疗手段并调配救治资源。

1　资料与方法

1.1　数据来源：数据来源于解放军总医院急救数据库，以其中全部明确诊断消化道出血的患者为研究对象，样本总数共 728 例次。提取患者入科 12 h 内医嘱记录，对该时间阶段数据进行回顾性分析，根据急诊消化道出血诊疗常规，将 343 例次确定发生消化道致命性再出血患者（需给予输血治疗）纳入试验组；将 385 例次未发生消化道致命性再出血患者纳入对照组。

1.1.1　纳入标准：明确诊断为消化道再出血的患者；年龄≥18 岁；来诊血常规示血红蛋白（Hb）≥90 g/L。

1.1.2　排除标准：年龄＜18 岁；入院后首次化验结果 Hb＜90 g/L 或未进行该化验；临床资料缺失。

1.2　机器学习分类预测方法：分别构建 Logistic 回归和以决策树为弱分类器的 AdaBoost 算法、XGBoost 算法。Logistic 回归是计量与机器学习的常用模型之一，因其训练简单、速度快，在医学领域已被广泛应用。AdaBoost 和 XGBoost 属于提升算法，根据所有弱分类器组合成强分类器，因此二者往往比单独的分类算法表现更好。二者的区别在于：前者在迭代中不断提升前一个弱分类器错分样本的权重，后者则不断加入弱分类器以降低分类损失。本研究以决策树作为弱分类器，因此能够返回每个指标的特征重要性：一个指标越多地被用来构建决策树，其重要性也就越高。

1.3　研究与数据处理：在数据库中收集提取患者常规诊疗信息，包括性别、年龄、生命体征以及血气分析、血生化、凝血功能、血常规等化验指标，共获得原始指标 499 项，筛掉缺失值较多或临床意义较小的指标，得到 64 项指标；对患者是否发生消化道致命性再出血进行标记；考虑到临床工作中对患者发生再出血预测的实际需要，将确定发生消化道致命性再出血患者的获取数据时间跨度控制在输血前。不同患者的测量时间节点与周期并不相同，数据并非规整的时间序列。对于数据集，将患者指标以时间跨度三等分，在等分内对数据进行平均处理，得到病例个数、时间步长、指标个数 3 个具有代表性的时间截面，并处理成各模型可以输入的数据结构。缺失值用所有患者的同指标平均值替换，并进行标准化后得到最终数据集。

1.4　预测精度判断标准：一些分类问题中会使用正确率作为评价指标。

$$\text{正确率}=\frac{\text{被正确预测的患者（真阳性+真阴性）}}{\text{所有患者（阳性+阴性）}}$$

为了更加全面地比较同一模型内部选取不同参数所导致的不同预测结果，应当从精确率、召回率两个角度综合衡量模型分类效果。精确率主要考察正样本的分类是否准确，召回率则考察是否尽可能地找出了正样本。

$$\text{精确率}=\frac{\text{被正确预测为真的患者（真阳性）}}{\text{所有被预测为真的患者（真阳性+假阳性）}}$$

$$\text{召回率}=\frac{\text{被正确预测为真的患者（真阳性）}}{\text{所有实际为真的患者（真阳性+假阴性）}}$$

实际上，精确率和召回率这两个指标往往会相互冲突制约，为了综合权衡两者，本研究使用 F 分数作为考察模型分类的指标。

$$\text{F分数}=(1+\beta^2)\times\frac{\text{精确率}\times\text{召回率}}{\beta^2\times\text{精确率}+\text{召回率}}$$

在本研究中应尽可能找出潜在的再出血患者，故召回率更加重要，因此 F 分数中 β 取 1.5，即 F1.5 分数。

1.5　交叉验证方法：使用十折交叉验证法，将初始数据分割成 10 个子样本，每次取 1 个子样本作测试集，其余子样本作训练集，重复 10 次。该方法能够充分利用数据，且较大程度地验证模型的泛化能力。

1.6　关键指标筛选方法：应用序列前向搜索算法作特征选择，即在开始时将特征集合分别减去 1 个特征作为子集，每

个子集有 $N-1$ 个特征，对每个子集进行评价，然后在最优子集上逐步减少特征，使模型性能最大提升或损失最小。

2 结 果

2.1 未经筛选的预测情况：Logistic 回归、AdaBoost 算法和 XGBoost 算法根据预测结果绘制混淆矩阵（图 1），计算出正确率、精确率、召回率和 F1.5 分数（表 1）显示，XGBoost 算法的 F1.5 分数最高，较高的召回率意味着其能够尽可能找出更多的可能发生院内消化道致命性再出血的患者；AdaBoost 算法表现其次，Logistic 回归表现较差。

表 1 全部指标输入时 Logistic 回归、AdaBoost 算法、XGBoost 算法预测院内消化道致命性再出血的正确率、精确率、召回率和 F1.5 分数

模型	正确率	精确率	召回率	F1.5 分数
Logistic 回归	0.795	0.776	0.794	0.787
AdaBoost 算法	0.815	0.785	0.841	0.822
XGBoost 算法	0.857	0.852	0.847	0.846

当决策树充分生长后，在树上运行剪枝算法，各分支处的特征即为选出来的特征子集，但通过决策树选择的特征并不稳定，其选择的特征个数与样本量密切相关[7]。提升算法可以解决这一问题，Assareh 等[8]通过以决策树为弱分类器的 AdaBoost 算法进行了特征选择。充分训练 XGBoost 算法模型后可以获取决策树中所有特征的重要性，得到前 30 个重要性较大的指标（表 2）。

2.2 基于 XGBoost 算法的指标筛选：序列前向选择是一种启发式搜索算法。在初始状态时，关键指标集为空集，每次迭代从未选择的指标中选择 1 个加入到关键指标集中，使关键指标集的 F1.5 分数达到最优。为使每次加入新指标均能达到最优 F1.5 分数，在每次迭代的选择指标过程中，会将每个未选择指标分别加入进行运算，并取其中得到最高 F1.5 分数的新加入指标。

经过 30 次迭代，记录逐渐扩大的关键指标集的输入下 XGBoost 算法的评分，表 3 中“新加入指标”表示，在每次迭代中选择其加入到关键指标集时，在特定指标个数下 F1.5 分数达到最高，并且每个指标越早被加入，说明其越重要。因此表 3 也表示了 30 个指标的重要性排序。

图 2 显示，当输入 12 个指标时 F1.5 分数达到了峰值 0.893，高于全指标时的 0.857。因此，表 3 中前 12 个指标即为院内消化道致命性再出血的关键指标，依次为血红蛋白测定（Hb）、钙（CA）、红细胞计数（RBC）、平均血小板体积测定（MPV）、平均红细胞血红蛋白浓度（MCH）、收缩压（SBP）、血小板计数（PLT）、镁（MG）、淋巴细胞（LYM）、葡萄糖（GLU，血气分析）、葡萄糖（GLU，血生化）、舒张压（DBP）。

表 2 全部指标输入时 XGBoost 算法预测院内消化道致命性再出血重要性较高的指标

序号	指标	重要性	序号	指标	重要性
1	血红蛋白测定（Hb）	0.0805	17	乳酸脱氢酶（LDH）	0.0145
2	舒张压（DBP）	0.0562	18	镁（MG）	0.0141
3	红细胞计数（RBC）	0.0509	19	碱性磷酸酶（ALP）	0.0137
4	体温（T）	0.0502	20	淋巴细胞（LYM）	0.0136
5	收缩压（SBP）	0.0493	21	γ-谷氨酰转移酶（GGT）	0.0132
6	心率（HR）	0.0419	22	钾（K）	0.0124
7	呼吸频率（RR）	0.0411	23	平均血小板体积测定（MPV）	0.0123
8	血细胞比容测定（Hct）	0.0335	24	葡萄糖（GLU，血生化）	0.0122
9	脑利钠肽前体（proBNP）	0.0321	25	嗜酸粒细胞（EGC）	0.0119
10	葡萄糖（GLU，血气分析）	0.0242	26	氧分压（PO_2）	0.0119
11	钙（CA）	0.0240	27	平均红细胞血红蛋白浓度（MCH）	0.0117
12	乳酸（Lac）	0.0198	28	二氧化碳分压（PCO_2）	0.0115
13	血小板计数（PLT）	0.0187	29	肌红蛋白定量（Mb）	0.0115
14	肌酐（Cr）	0.0166	30	总胆红素（TBil）	0.0113
15	白细胞计数（WBC）	0.0149			
16	细胞外液剩余碱（BEecf）	0.0147			

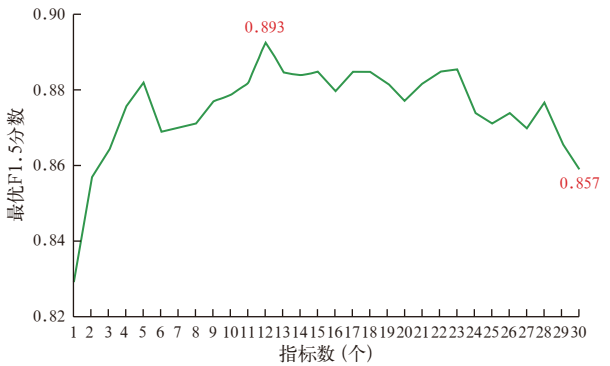
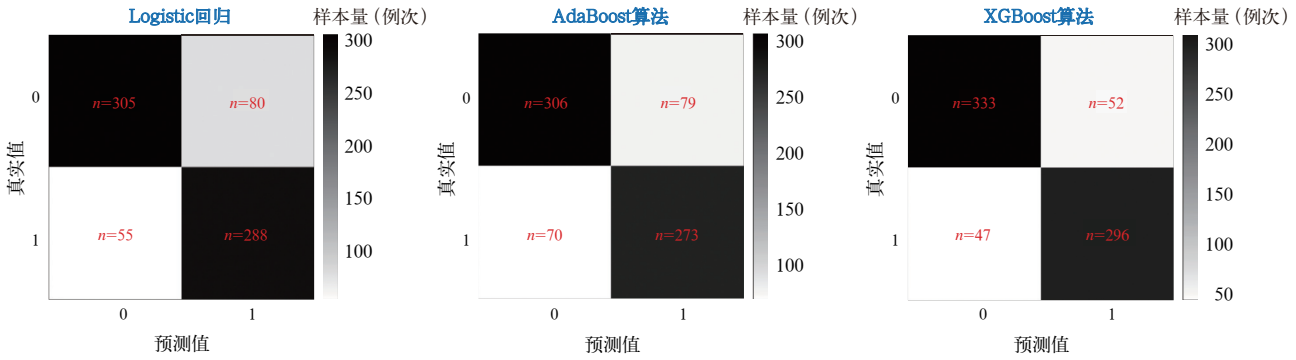


图 2 不同输入指标个数时 XGBoost 算法预测院内消化道致命性再出血的 F1.5 分数变化



注：0 为未发生再出血，1 为发生再出血

图 1 全部指标输入时 Logistic 回归、AdaBoost 算法和 XGBoost 算法预测院内消化道致命性再出血的混淆矩阵

表 3　不同输入指标个数时 XGBoost 算法预测院内消化道致命性再出血的正确率、精确率、召回率和 F1.5 分数

指标个数	新加入指标	正确率	精确率	召回率	F1.5 分数	指标个数	新加入指标	正确率	精确率	召回率	F1.5 分数
1	血红蛋白测定（Hb）	0.826	0.809	0.829	0.821	16	白细胞计数（WBC）	0.876	0.867	0.880	0.875
2	钙（CA）	0.850	0.834	0.857	0.847	17	总胆红素（TBil）	0.886	0.880	0.885	0.881
3	红细胞计数（RBC）	0.856	0.833	0.864	0.852	18	脑利钠肽前体（proBNP）	0.885	0.877	0.885	0.880
4	平均血小板体积测定（MPV）	0.874	0.862	0.876	0.868	19	体温（T）	0.885	0.88	0.882	0.878
5	平均红细胞血红蛋白浓度（MCH）	0.869	0.849	0.882	0.868	20	乳酸脱氢酶（LDH）	0.880	0.875	0.877	0.874
6	收缩压（SBP）	0.876	0.871	0.869	0.868	21	细胞外液剩余碱（BEecf）	0.876	0.864	0.882	0.874
7	血小板计数（PLT）	0.879	0.875	0.870	0.871	22	氧分压（PO_2）	0.879	0.868	0.885	0.877
8	镁（MG）	0.882	0.876	0.871	0.871	23	血细胞比容测定（Hct）	0.876	0.865	0.886	0.876
9	淋巴细胞（LYM）	0.883	0.879	0.877	0.876	24	嗜酸粒细胞（EGC）	0.875	0.871	0.874	0.870
10	葡萄糖（GLU，血气分析）	0.882	0.873	0.879	0.876	25	γ-谷氨酰转移酶（GGT）	0.871	0.864	0.871	0.866
11	葡萄糖（GLU，血生化）	0.883	0.874	0.882	0.878	26	二氧化碳分压（PCO_2）	0.875	0.870	0.874	0.869
12	舒张压（DBP）	0.880	0.864	0.893	0.881	27	呼吸频率（RR）	0.868	0.858	0.870	0.863
13	乳酸（Lac）	0.879	0.867	0.885	0.877	28	碱性磷酸酶（ALP）	0.865	0.850	0.877	0.865
14	肌红蛋白定量（Mb）	0.883	0.874	0.884	0.879	29	钾（K）	0.867	0.863	0.866	0.862
15	肌酐（Cr）	0.881	0.867	0.885	0.878	30	心率（HR）	0.864	0.859	0.859	0.856
						64	全指标	0.857	0.852	0.847	0.846

3　讨　论

本研究基于 Logistic 回归和基于决策树的 AdaBoost 算法、XGBoost 算法对患者院内消化道致命性再出血进行预测，3 种方法都能达到对此类事件进行预测预警的目的，且优势各不相同。Logistic 回归是医学领域中常用的方法，构造简单、易于计算、可解释性强，输出结果具有概率意义；决策树较 Logistic 回归能够得到更好的非线性决策边界[9]，应用 AdaBoost 算法和 XGBoost 算法的集成学习方法精度会进一步提升。有研究表明，XGBoost 算法的计算效果甚至可以媲美需要耗费大量计算资源的深度神经网络[10]。本研究结果也显示，以决策树为弱分类器的 XGBoost 算法在输入指标数较少时仍能保持较好的预测结果，且召回率较大，这意味着其在舍弃一定精确率时找出潜在院内消化道致命性再出血患者的能力较强，在临床上的预测预警意义更大。

在临床工作中，提早预测院内致命性消化道再出血的发生是挽救患者生命、避免患者死亡恶性事件发生的重要手段，本研究在理论层面能达到该目的，但仍存在诸多局限：① 病例数仍比较有限，欲提高预测准确性还需扩大样本量；② 目前数据库难以录入患者在连续一段时间内各项指标的变化情况，难以形成规范标准的时间序列，在对数据进行清洗时难免丢失部分信息，可能对预测准确性造成影响；③ 对比的数据分析模型与算法仍有限，是否有更好的模型可以达到更好的预测目的仍未可知；④ 研究采取病例数据回顾性分析方法，虽然已建立形成一整套完整的数据分析技术体系，但距离形成临床可应用模型仍有较大差距。以上所涉及关键问题将是本研究下一步工作重点。

综上所述，对于院内消化道致命性再出血，以决策树为弱分类器的 XGBoost 算法是一种较好的预测方法。根据其序列前向搜索结果，血红蛋白测定（Hb）、钙（CA）、红细胞计数（RBC）、平均血小板体积测定（MPV）、平均红细胞血红蛋白浓度（MCH）、收缩压（SBP）、血小板计数（PLT）、镁（MG）、淋巴细胞（LYM）、葡萄糖（GLU，血气分析）、葡萄糖（GLU，血生化）、舒张压（DBP）共 12 个指标为院内消化道致命性再出血的关键指标。

利益冲突　所有作者均声明不存在利益冲突

参考文献

[1] 周彪，牛旭平. 上消化道出血的诊断与治疗新进展[J]. 中国现代医生，2009, 47 (19): 31-34. DOI: 10.3969/j.issn.1673-9701.2009.19.015.
Zhou B, Niu XP. New progress of the diagnosis and treatment of upper gastrointestinal bleeding [J]. Chin Mod Doctor, 2009, 47 (19): 31-34. DOI: 10.3969/j.issn.1673-9701.2009.19.015.

[2] Chu A, Ahn H, Halwan B, et al. A decision support system to facilitate management of patients with acute gastrointestinal bleeding [J]. Artif Intell Med, 2008, 42 (3): 247-259. DOI: 10.1016/j.artmed.2007.10.003.

[3] Cismondi F, Celi LA, Fialho AS, et al. Reducing unnecessary lab testing in the ICU with artificial intelligence [J]. Int J Med Inform, 2013, 82 (5): 345-358. DOI: 10.1016/j.ijmedinf.2012.11.017.

[4] Ayaru L, Ypsilantis PP, Nanapragasam A, et al. Prediction of outcome in acute lower gastrointestinal bleeding using gradient boosting [J]. PLoS One, 2015, 10 (7): e0132485. DOI: 10.1371/journal.pone.0132485.

[5] 张政波，薛万国，曹德森，等. 急救大数据与 Datathon 活动[J]. 中华危重病急救医学，2018, 30 (6): 603-605. DOI: 10.3760/cma.j.issn.2095-4352.2018.06.020.
Zhang ZB, Xue WG, Cao DS, et al. Big data in emergency medcine and Datathon event [J]. Chin Crit Care Med, 2018, 30 (6): 603-605. DOI: 10.3760/cma.j.issn.2095-4352.2018.06.020.

[6] 赵宇卓，王俊梅，潘菲，等. 急救数据库建设初探[J]. 中华危重病急救医学，2018, 30 (6): 609-612. DOI: 10.3760/cma.j.issn.2095-4352.2018.06.022.
Zhao YZ, Wang JM, Pan F, et al. Pilot research: construction of emergency rescue database [J]. Chin Crit Care Med, 2018, 30 (6): 609-612. DOI: 10.3760/cma.j.issn.2095-4352.2018.06.022.

[7] Wang YY, Li J. Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data [J]. Intern J Rem Sen, 2008, 29 (10): 2993-3010. DOI: 10.1080/01431160701442070.

[8] Assareh A, Volkert LG, Li J. Feature selections using AdaBoost: application in gene-gene interaction detection [C]// IEEE. 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, Philadelphia, 2012. Philadelphia: IEEE, 2012: 831-837. DOI: 10.1109/BIBMW.2012.6470248.

[9] De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis [J]. Ecology, 2000, 81 (11): 3178-3192. DOI: 10.2307/177409.

[10] Sheridan RP, Wang WM, Liaw A, et al. Extreme gradient boosting as a method for quantitative structure-activity relationships [J]. J Chem Inf Model, 2016, 56 (12): 2353-2360. DOI: 10.1021/acs.jcim.6b00591.

（收稿日期：2018-09-21）