·冠状病毒肺炎专题·

基于机器学习的新冠肺炎典型药物疗效分析

许娟娟[1]，陈洞天[2]，任宇飞[2]，李金[2]（华中科技大学同济医学院，1. 附属梨园医院药剂科，2. 附属同济医院计算机中心，湖北 武汉 430030）

[摘要] **目的**：新型冠状病毒肺炎具有传播能力强、检测准确率低、治疗难度大等特点，对我国乃至全世界人类健康和社会安全造成了巨大威胁。新冠肺炎目前没有特效药物，因此，需要比较各类药物对新冠肺炎的治疗效果，为重症肺炎患者的治疗提供参考。**方法**：本文选取某院 1 384 名重症出院患者作为数据集，提取最广泛使用的三类药物作为训练特征，构建基于药物使用情况预测患者治愈率的多个机器学习模型，最后基于综合性能最优的模型，分析三类药物对于治愈率的重要性。**结果**：年龄是影响治愈率的最重要因素；激素类药物在所有药物中影响占比最大，中成药的疗效也非常突出。**结论**：相对于传统的针对单个药物药效的统计学研究方法，机器学习可以将所有药物药效进行综合分析，能直观地显示各个药物对治愈率的影响程度。

[关键词] 新冠肺炎；机器学习；药物疗效；随机森林；特征重要性
[中图分类号] R974 [文献标识码] A [文章编号] 1001-5213(2020)11-1177-05 DOI：10. 13286/j. 1001-5213. 2020. 11. 01

Analysis of the efficacy of typical drugs for COVID-19 based on machine learning

XU Juan-juan[1], CHEN Dong-tian[2], REN Yu-fei[2], LI Jin[2] (Tongji Medical College, Huazhong University of Science and Technology, 1. Department of Pharmacy, Liyuan Hospital, 2. Computer Center, Tongji Hospital, Hubei Wuhan 430030, China)

ABSTRACT: OBJECTIVE COVID-19(Corona Virus Disease 2019) has characteristics of strong transmission ability, low detection accuracy, and difficult treatment, which had posed huge threat to mankind health and social security in China and all over the world. There is no specific drug for COVID-19. Therefore, it is necessary to compare the therapeutic effects of various drugs on new-coronary pneumonia and provide a reference for the treatment of patients with severe pneumonia. **METHODS** A total of 1 384 critically ill patients discharged from Tongji Hospital were selected as the data set, and the three most widely used drugs were extracted as the training characteristics to construct multiple machine learning models for predicting the cure rate of patients based on drug use. Finally, based on the model with optimal comprehensive performance, the importance of the three drugs for the cure rate was analyzed. **RESULTS** Age was the most important factor affecting the cure rate; hormone drugs accounted for the largest proportion of all drugs, and the efficacy of Chinese patent medicines was also very significant. **CONCLUSION** Compared with the traditional statistical research method for the efficacy of a single drug, machine learning can comprehensively analyze the efficacy of all drugs and can visually show the degree of influence of each drug on the cure rate.
KEY WORDS: COVID-19; machine learning; drug efficacy; random forest; feature importance

1 引言

新型冠状病毒肺炎（Corona Virus Disease 2019，COVID-19），简称“新冠肺炎”，是指 2019 新型冠状病毒感染导致的肺炎。新冠肺炎疫情自出现以来，已经陆续蔓延到我国及境外多个国家[1]，对人类健康和社会安全造成了巨大威胁。

新冠肺炎目前没有特效药物，截至 2020 年 3 月 3 日国家卫健委已经发布了第七版诊疗方案[1]，推荐药物治疗方案主要包括由抗病毒治疗、免疫治疗、激素治疗组成的西药治疗方案和由中成药、中草药组成的中药治疗方案。一般采用中西医结合治疗的方式，其中西药治疗见效快，但是也存在副作用较大的问题，例如作为免疫抑制剂的糖皮质激素类药物；中药治疗对于普通型患者能明显改善症状，但是对于重症型患者效果不太明显。因此，亟需对各类药物的治疗效果进行研究分析，为新冠肺炎临床治疗提供指导，提高患者治愈率。目前对于药物疗效分析的统计方法主要是针对单个药物疗效进行显著性检验，方法包括 χ^2 四格法、参数统计法、非参数统计法等[2]，只能在受控条件下对单个药

[作者简介] 许娟娟，女，本科，主管药师，研究方向：临床药学，电话：13037176465，E-mail：410913329@qq. com [通讯作者] 陈洞天，男，硕士，中级工程师，研究方向：医疗信息化，电话：18986145582，E-mail：cdt86915998@gmail. com

物的药效进行分析，无法同时对多个药物以及其他因素的影响程度进行分析。

为了能同时在患者生理条件及用药情况比较复杂的情况下，对所有药物疗效进行一个整体的分析，本文选取了机器学习作为研究工具。机器学习是一种基于概率论和统计学等理论知识，并利用具有高计算性能的现代计算机作为工具，模拟人类的学习方式，对未知数据进行预测和推断的应用学科。

本文选取了同济医院的 1 384 名新冠肺炎重症出院患者作为样本，提取使用最多的 3 类药物作为特征数据集，构建了基于药物用量来预测患者治愈率的机器学习模型，对模型的各个特征重要性进行分析，为临床治疗提供指导参考。

2　材料与方法

2.1　数据来源　本研究选择同济医院中法新城院区和光谷院区收治新冠肺炎重症患者以来，截至 2020 年 3 月 5 日的 1 384 名出院患者。(1)入组标准：年龄不限，男女不限；新冠肺炎确诊患者；重症或危重型患者。(2)排除标准：非发热病房患者，非新冠肺炎确诊患者；轻症患者。1 384 名患者满足入组标准，其中 594 名治愈，500 名好转，1 名未愈，273 名死亡，16 名其他情况。组内患者各类药物用量见表 1。

分别选择西药中的盐酸阿比多尔分散片、盐酸阿比多尔颗粒、盐酸莫西沙星片、注射用甲泼尼龙琥珀酸钠，中成药中的连花清瘟胶囊、金叶败毒颗粒、血必净注射液，中草药中的恢复期-颗粒剂、恢复期、进展期-不发热-颗粒剂作为特征数据集，每个特征代表样本患者的药品使用量（西药及中成药为包装单位数，中草药为剂数），同时加入患者的年龄和性别两个基本信息作为补充特征。

2.2　数据探索与预处理　数据预处理能改善数据集的完整性，降低冗余性和相关性，有效提升算法模型质量[3]。本研究采用 Pandas[4] 数据模型工具对数据集进行结构化处理。部分样本见表 2。

各特征解释如下：

(1)连花清瘟、金叶败毒、血必净、恢复期颗粒、恢复期、进展期、阿比多尔 1、阿比多尔 2、莫西沙星、甲泼尼龙分别代表连花清瘟胶囊、金叶败毒颗粒、血必净注射液、恢复期-颗粒剂、恢复期、进展期-不发热-颗粒剂、盐酸阿比多尔分散片、盐酸阿比多尔颗粒、盐酸莫西沙星片、注射用甲泼尼龙琥珀酸钠 10 个药品，特征值的单位为药品的包装单位，如盒、瓶、支，中草药则为剂。

(2)discharge_status 表示患者出院状态，1 代表治愈，2 代表好转，3 代表未愈，4 代表死亡，5 代表其他。

其中，盐酸阿比多尔分散片（arbidol1）和盐酸阿比多尔颗粒（arbidol2）属于同一种药品的不同剂型和规格，盐酸阿比多尔分散片的规格为 0.1g * 12 片/盒，盐酸阿比多尔颗粒的规格为 0.1 g * 6 片/盒，因此我们创建一个新的特征 arbidol，特征值等于 arbidol1 × 2 + arbidol2，然后移除原有的两个特征 arbidol1 和 arbidol2。

处理后的数据集基本情况如表 3 和表 4 所示。

表 1　出院患者药品使用量分类排名

Tab 1　Drug usage ranking of discharged patients

用量排名	西药		中成药		中草药	
	药名	开立次数	药名	开立次数	处方名	开立剂数
1	氯化钠注射液	32 781	莲花清瘟胶囊	6 721	恢复期-颗粒剂	1 613
2	盐酸阿比多尔分散片	5 372	金叶败毒颗粒	2 247	恢复期	551
3	葡萄糖注射液	4 454	血必净注射液	477	进展期-不发热-颗粒剂	377
4	盐酸阿比多尔颗粒	4 138	百令胶囊	353	恢复期 1-乏力气短舌苔偏厚	150
5	胰岛素注射液	3 943	苏黄止咳胶囊	303	气阴两虚方	102
6	盐酸莫西沙星片(薄膜衣)	3 810	便乃通茶	260	进展期-发热-颗粒剂	98
7	注射用甲泼尼龙琥珀酸钠	3 532	双黄连口服液	129	抗病毒协定方 3 号	90
8	重酒石酸去甲肾上腺素注射液	2 390	蓝芩口服液	105	新冠肺炎恢复期-颗粒剂	65

表 2　部分数据集样本

Tab 2　Sample of dataset

性别	年龄/岁	连花清瘟	金叶败毒	血必净	恢复期颗粒	恢复期	进展期	阿比多尔 1	阿比多尔 2	莫西沙星	甲泼尼龙	discharge_status
女	37	0	0	0	0	0	0	5	0	2	2	1
女	30	0	2	0	5	0	0	5	0	2	0	2
女	50	2	3	0	0	0	0	0	0	4	6	2

表 3 数据集基本信息

Tab 3 Basic information of dataset

名称	类型	非空值总数
性别	object	1 384
年龄	int64	1 384
连花清瘟	int64	1 384
金叶败毒	int64	1 384
血必净	int64	1 384
恢复期颗粒	int64	1 384
恢复期	int64	1 384
进展期	int64	1 384
阿比多尔	int64	1 384
莫西沙星	int64	1 384
甲泼尼龙	int64	1 384
discharge_status	int64	1 384
行总数	1 384	
列总数	12	
数据类型统计	int64(11), object(1)	
内存占用	124.4 + kB	

分析信息结论如下：

(1)数据集不存在缺失值，无需进行插值等缺失值处理操作；(2)存在两个离散型特征(categorical feature)：sex 和 discharge_status；(3)不同药品的用量的均值差别较大，原因是不同药品的包装单位和规格存在差异；(4)患者年龄(age)的均值为 59.55，中位数为 63，最大值为 95，可以推测重症患者更加偏向高龄人群。

机器学习模型需要数字型数据才能进行计算，因此需要对离散型特征进行量化处理。两个离散型特征中，discharge_status 是最后的分类器模型输出值，无需处理。sex 特征只有 male 和 female 两个值，可以进行作为独立特征处理，将 female 和 male 转换为两个新的特征，特征值为 0 和 1(分别代表“是”和“否”)，并移除原始特征 sex。

对于不同药品的均值差别较大的问题，主要是不同的药品单位和规格引起，需要消除药品单位和规格的影响。在此我们利用 SKLearn 框架的 StandardScaler 对所有特征进行 Z-Score 标准化处理，消除不同包装单位和规格带来的数值差异。

表 4 数据集统计信息

Tab 4 Statistics information of dataset

名称	年龄	连花清瘟	金叶败毒	血必净	恢复期颗粒	恢复期	进展期	阿比多尔	莫西沙星	甲泼尼龙	discharge_status
总数	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000	1 384.000 000
均值	59.552 023	2.787 572	0.961 705	1.119 942	1.252 168	0.442 197	0.309 249	7.653 902	2.950 867	1.835 260	2.000 723
标准差	15.511 810	3.688 479	2.117 136	6.338 493	3.437 595	1.847 273	1.679 626	6.780 689	4.611 647	4.180 874	1.150 623
最小值	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	1.000 000
25%值	50.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	1.000 000
50%值	63.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	7.000 000	0.000 000	0.000 000	2.000 000
75%值	70.000 000	5.000 000	0.000 000	0.000 000	0.000 000	0.000 000	0.000 000	12.000 000	5.000 000	2.000 000	2.000 000
最大值	95.000 000	18.000 000	21.000 000	100.000 000	40.000 000	15.000 000	24.000 000	46.000 000	29.000 000	81.000 000	5.000 000

所有 1 384 名出院患者中，治愈和好转患者共计 1 094 名，死亡患者共 273 名，剩下 17 名其他情况的患者仅占总人数的 1.23%。为了使研究结果更加直观，我们移除其他情况的 17 名患者样本，将剩下的 1 367 名患者分为治愈好转和死亡两类，并在数据集中增加 cured 特征，代表治愈好转或者死亡，并移除原始特征 discharge_status。

从年龄信息可以看出新冠肺炎在高龄人群中存在更大风险。此外，近期新闻报道及学术研究也表明，老年人是新冠肺炎的易感和高危人群[5]，并且男性比女性的症状更严重[6]。样本的性别和年龄分布统计见图 1。

上图中，左右分别为女性和男性患者的治愈率相对于年龄的密度分布图；横坐标为年龄，纵坐标为患者数量；蓝色和橙色分别代表治愈患者和死亡患者。

由图 1 可见，年龄对治愈率的影响非常大，基本上绝大部分死亡患者样本都分布在年龄为 50 岁以上的区间；对于 50 岁以上的年龄区间，男性患者死亡率大大高于女性患者；对于 50 岁以下的年龄区间，女性患者死亡率高于男性患者。

综上，年龄应该是患者治愈率最关键的影响因素之一；对于老龄患者，性别是影响治愈率的重要因素。

2.3 方法 为了能获取一个相对准确的模型，本研究应用了几个常用的机器学习算法，包括：(1)随机梯度下降(stochastic gradient descent, SGD)：随机梯度下降算法是梯度下降算法中的一种，相对于原始的批量梯度下降(batch gradient descent, BGD)算法收敛速度更快，在目标函数为凸函数的情况下能保证收敛到全局最优解。(2)Logistic 回归(logistic regression)：Logistic 回归是一个广泛使用的多变量的二项分类模型，在医学领域运用尤为常见[8]。医学领域经常对病例群体和非病例群体进行对照研究，研究自变量(常为疾病的危险因素)与因变量(疾病发生)之间的定量关系。(3)随机森

林(random forest):随机森林是一种集成学习方法,被广泛应用于数据分类和非参数回归[9]。随机森林是一个包含多个决策树的分类器,其输出的类别是由所包含决策树输出的类别的众数而定。(4)朴素贝叶斯(Naive Bayesian):朴素贝叶斯算法是应用最为广泛的分类算法之一,具有结构简单、分类精度高和速度快等优点[10]。该算法先通过已给定的训练集,以特征词之间独立作为前提假设,学习从输入到输出的联合概率分布,再基于学习到的模型,求出使得后验概率最大的输出结果。

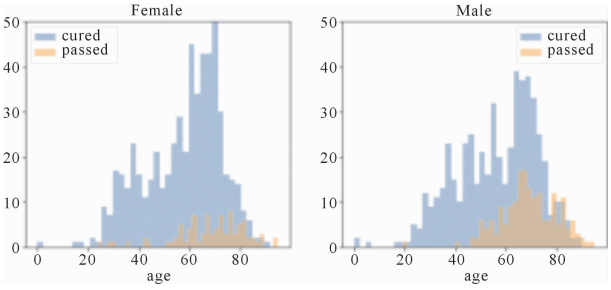

图 1　患者性别、年龄分布图

Fig 1　Distribution of sex and gender

为了能充分有效利用训练数据集,并且基于该数据集获取最优模型[7],本研究采用 10 折交叉验证(10 Fold Cross Validation)的方法,将训练数据集拆分为 10 份,进行 10 次迭代训练,每次迭代训练选择其中 9 份数据作为训练数据,剩下 1 份作为验证数据。完成迭代后,选择其中验证结果最优的模型为最终模型。

3　结果

各模型的参数配置及性能表现见表 5。

综合上述 4 个模型及性能表现,可以得知,除了朴素贝叶斯模型之外,其他 3 个模型的准确率和稳定性均满足需求,并且十分接近。在此我们选取预测准确率最高的随机森林模型作为最终模型。

随机森林作为一种集成学习模型,其特点之一是可以输出特征重要性,即对森林中每一个树的节点进行统计,如果该特征在节点上出现的次数越多,则其特征重要性越高。对于本文中的机器学习模型中的特征重要性,其意义可以直观地描述为,各类药物的用量参与判断患者最终状态是治愈或者死亡的次数,可以作为药物疗效的体现。

本研究中的随机森林模型的特征重要性输出结果如图 2 所示。从图 2 中的特征重要性排名我们可以得出如下结论:年龄确实是新冠肺炎重症患者治愈率的最大影响因素;甲泼尼龙作为一种糖皮质激素类药物,其副作用虽然高,但是对治愈率的影响程度非常高,能有效改善治愈率;中成药对于重症患者治愈率改善的表现比较好,包括连花清瘟胶囊和金叶败毒颗粒;中草药主要用于预防或者恢复,能减少危重症率和病死率[11],对于重症患者起一个支持辅助的作用。

基于上述研究结果,对于新冠肺炎重症患者的治疗,可以给出以下建议:(1)对于重症患者,不应过于限制激素类药物的使用;(2)应普及使用连花清瘟胶囊等治疗效果明显的中成药物;(3)中草药可以用于重症患者的恢复期,辅助患者肺功能的修复。

4　结论

机器学习方法并非新的工具,且在医学领域已经有较多应用,包括疾病诊断、医学影像分析、病历数据挖掘等;但是在药效研究方面,机器学习还鲜有使用。本文将机器学习算法运用于药物药效分析上,是一种新颖的研究方法,相对于传统的药效分析方法,其优点包括:不需要在控制条件下进行(例如控制变量法,设置实验组/对照组),所有影响因素都可以同时纳入分析范围;可以直观地查看所有药物药效的影响效果及排名情况,也可以与患者的生理因素进行对比,获取一个相对感性的认知;能不断学习新的数据,改善自身性能。

但本法也存在不足,例如不能获取精确的药效分析结果,只能说明各类药物对患者治愈率的一个综合影响程度;实验前提是各类药物对患者治愈率的影响是积极的,事实上可能存在服用药物但是加重患者病情的情况,此时实验结果是相反的。

表 5　各模型参数配置及性能表现

Tab 5　Parameters and performance of various models

模型参数配置	随机梯度下降 max_iter(最大迭代数) = 100tol(停止迭代条件) = none(5 次 loss 值不变后停止)	Logistic 回归 max_iter(最大迭代数) = 5 000	随机森林 n_estimators(决策树数量) = 100	朴素贝叶斯 采取默认配置
交叉验证准确率均值 ± 标准差	0. 884 ± 0. 037	0. 888 ± 0. 0356	0. 882 ± 0. 039	0. 548 ± 0. 064
测试集准确率	0. 894	0. 909	0. 912	0. 551

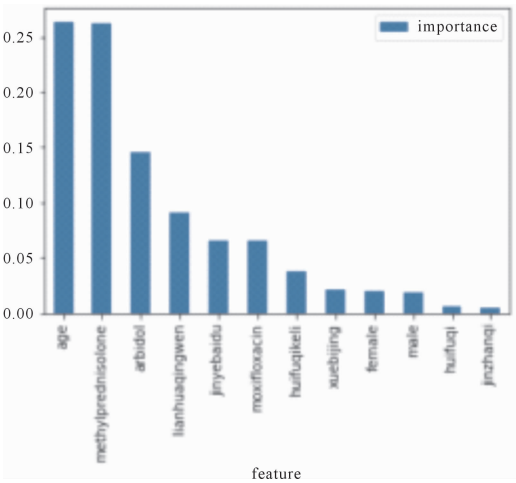

图 2 随机森林模型的特征重要性排名

Fig 2 Feature importance ranking of random forest model

本研究的意义在于在患者生理情况和用药情况比较复杂的情况下，分析各个因素对患者最终治愈率的影响程度。相对于传统方法，机器学习将为药物疗效分析提供一种新的，更加适应多元、复杂数据的研究方法。

参考文献：

[1] The General Office of the National Health and Health Commission and the Office of the State Administration of Traditional Chinese Medicine. Notice on Printing and Distributing Pneumonia Diagnosis and Treatment Program for NewCoronavirus Infection [EB/OL]. (2020-03-03). http://www.gov.cn/zhengce/zhengceku/ 2020-03/04/content_5486705.htm

[2] Lu GA. Basic methods for statistical analysis of medicinal effects in medical scientific research [J]. Chin J Basic Med Tradit Chin Med (中国中医基础医学杂志), 1997(S2): 36-44.

[3] Liu MJ, Wang XF. Data preprocessing in data mining [J]. Computer Sci (计算机科学), 2000, 27(4): 54-57.

[4] Pandas - Python Data Analysis Library [EB/OL]. https://pandas.pydata.org/

[5] Tang DZ, Wang J, Liang QQ, *et al*. Discussion on the prevention and treatment of new coronavirus pneumonia in the elderly by regulating the state of "nephrine" [J]. Tianjin J Tradit Chin Med(天津中医药), 2020, 37(2): 125-131.

[6] Sina Technology. Gender differences in COVID-19: Women have longer incubation period and men more susceptible to infection? [EB/OL]. (2020-03-07). https://tech.sina.com.cn/d/f/2020-03-07/doc-iimxxstf7083798.shtml

[7] Fan YD. Overview of cross-validation methods in model selection [D]. Shanxi University, 2013.

[8] Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain [J]. J Clin Epidemiol, 2001, 54(10): 979-985.

[9] Dong SS, Huang ZX. Analysis of random forest theory [J]. Integr Tech (集成技术), 2013, 2(1): 1-7.

[10] Zhu XD. Improved Naive Bayesian classification model [D]. Xiamen University, 2014.

[11] Wang YG, Qi WS, Ma JJ, *et al*. Traditional Chinese medicine clinical features and syndrome differentiation for COVID-19 [J]. J Tradit Chin Med (中医杂志), 2020, 61(4): 281-285.

[收稿日期]2020-03-14