基于机器学习的新冠肺炎与流感快速鉴别方法的研究

葛晓伟① 梁盼② 马晓旭③ 程铭*

摘　要　新冠肺炎作为一种新发性传染疾病，与流感均含有发热、咳嗽等临床特征，快速准确地将新冠肺炎与流感进行鉴别，有助于对患者进行救治。采用独立样本*t*检验的方法对患者检验数据中多个指标进行差异性分析，选择差异性较大的指标，评估机器学习中线性与非线性算法、集成算法等多种算法在流感与新冠肺炎快速鉴别中的应用效果。结果显示，SVM算法在新冠肺炎与流感的鉴别问题上效果更好。

关键词　新冠肺炎　机器学习　快速鉴别

Doi:10.3969/j.issn.1673-7571.2020.09.006

[中图分类号] R319;TP391　　**[文献标识码] A**

Research on the Rapid Identification Method of COVID-19 and Influenza Based on Machine Learning / GE Xiao-wei, LIANG Pan, MA Xiao-xu, et al.

Abstract　As a new infectious disease, COVID-19 and influenza contain clinical features such as fever and cough. Quickly and accurately distinguish the COVID-19 from influenza, which is helpful for the treatment of patients. In this paper, the independent sample *t* test method is used to analyze the differences of multiple indicators in the patient test data, select the indicators with greater differences, the application effect in rapid identification of COVID-19 and influenza by linear and nonlinear algorithms and integrated algorithms in machine learning are evaluated. The results show that the SVM algorithm is better at distinguishing COVID-19 from influenza.

Keywords　COVID-19, machine learning, rapid identification

Fund project　Medical Science and Technology Research Project of He'nan Province (No. 2018020087)

Corresponding author　Information Department, the First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, He'nan Province, P.R.C.

1 引言

新冠肺炎（COVID-19）是一种由新型冠状病毒感染引起的以肺部病变为主的新型传染病[1]，其常见的临床症状为发热、干咳、气促、外周血白细胞一般不高或降低、胸片有炎症性改变等[2-3]。流感作为一种古老的疾病，中世纪末就有较为详细的历史记载，人类流感病毒根据其核蛋白的抗原性可以分为甲、乙、丙三种类型，其中，丙型相对较少[4]。流感病毒作为呼吸道传染病，常发于秋冬季，患病者多会出现畏寒、发热、头痛、乏力、全身酸痛等症状，进而引发高热，恶心、便秘等，重症者可危及生命[5]。

由于新冠肺炎与流感均含有发热、咳嗽、鼻塞、流涕、咽痛等症状[6-7]。面对突如其来的疫情，患者出现发热、咳嗽、鼻塞等症状后，既担心是新冠肺炎，又怕在医院就诊时发生交叉感染。此外，新冠肺炎是新发传染病，属于乙类传染病，按甲类传染病来管理；而流感则属于丙类传染病，两者规定的管理地位有所不同，如不将两者区分开来，尤其是将流感视为新冠肺炎进行集中收治，就会大大增加交叉感染的几率。因此，如何把流感和新冠肺炎快速鉴别开来成为当务之急。

目前，区分流感和新冠肺炎两种疾病最好的标准，仍是进行病原学检测或免疫学检测，即通常说的核酸检测。长

基金项目：河南省医学科技攻关计划项目（编号：2018020087）

*通信作者：郑州大学第一附属医院信息处，450052，河南省郑州市二七区建设东路1号
①郑州大学第一附属医院信息处，450052，河南省郑州市二七区建设东路1号
②郑州大学第一附属医院放射科，450052，河南省郑州市二七区建设东路1号
③郑州大学第一附属医院呼吸内科，450052，河南省郑州市二七区建设东路1号

期专注于医学信息学相关研究[8-11]，考虑如何通过更为普遍的血常规、PCT（降钙素原）、CRP（C反应蛋白）等检验指标快速区分新冠肺炎与流感，探讨机器学习算法在进行新冠肺炎与流感鉴别上的可行性，并评估选择最优模型为本文研究的目的。

2 数据与方法

2.1 数据来源 数据来源于某三甲医院2018—2019年门急诊、住院等确诊的甲型流感与乙型流感患者1 326例；新冠肺炎已确诊患者73例，共1 399个患者的全部血常规、PCT、CRP、核酸检测等检验指标计39 975条数据用于本研究。

2.2 数据预处理 抽取医院LIS系统中患者住院号/门诊号、检验时间、诊断、血常规、PCT、CRP、咽拭子核酸检测等检验指标数据建立原始特征库，并根据数据采集标准初步筛选（表1）；由于新冠肺炎患者与流感患者相比数据量较小，病情复杂，在进行数据采集时，选择其阳性确诊前最近的两次与入院第一次检验结果。对采集并初步筛选之后的数据进行异常值与噪声数据处理。

表1 患者特征标量数据采集

特征变量	参数	采集次数	筛选方式
血常规	20 个	1 或 2 次	阳性时间最近
PCT	1 个	1 或 2 次	阳性时间最近
CRP	1 个	1 或 2 次	阳性时间最近
核酸检测	1 个	1 次	阳性

由于特征变量的检验结果受临床治疗（尤其抗生素治疗）干扰，造成正负样本数据差异性较小，不能反映患者实际检验情况，应当以核酸检测结果的检验时间为判断基准，其他特征的检验时间需要不超过核酸检测时间之后24小时。

由于咽拭子核酸检测结果存在假阴性，核酸检测结果与患者临床表现不一致，此种结果是不确定的，属于摇摆数据，因此，删除做过多次核酸检测结果均为阴性的此类数据。

数据异常值通常由于人为误操作或机器故障而引起，如检验指标远大于实际值等，因此设置规则，删除与特征均值之比大于60和小于0.01的数据。

3 实验

采用经过数据预处理的，实际有效数为13 123条患者样本数据，以新冠肺炎患者数据数量为基准，流感患者随机取相同数据作为样本集，其中，20%的数据作为评估数据集，80%作为训练数据集。采用独立样本*t*检验的方法对患者检验数据中23个指标进行分析，选择差异性较大的指标用于分类，并通过设计算法评估框架，从机器学习多种分类算法中选择最优方法。

3.1 特征选择 采用独立样本*t*检验[12]的方法对流感（甲型、乙型流感）与新冠肺炎各项指标进行差异性分析，结果由表2可知，甲型流感与新冠肺炎分类选取15个指标（$P<0.05$），乙型流感与新冠肺炎分类选取17个指标（$P<0.05$），其中，中性粒细胞绝对值（Neut#）、淋巴细胞绝对值（Lymph#）、单核细胞绝对值（Mono#）等多个指标在新冠肺炎与流感的区分上差异性较大。

表2 流感与新冠肺炎各项指标*P*值结果

序号	指标	甲流 & 新冠肺炎	乙流 & 新冠肺炎	序号	指标	甲流 & 新冠肺炎	乙流 & 新冠肺炎
1	WBC	0.000616	0.128	13	Eos#	0.073228	0.005
2	RBC	2.02E-07	0.066	14	Baso#	0.067426	0.371
3	Hb	8.06E-05	0.11	15	Hct	2.03E-07	0.032
4	PLT	6.69E-13	3.1E-07	16	MCV	0.666192	0.269
5	Neut%	1.44E-39	7.77E-22	17	MCH	1.10E-06	0.073
6	Lymph%	7.02E-36	1.41E-19	18	MCHC	2.19E-15	3.78E-07
7	Mono%	4.12E-19	6.31E-13	19	MPV	2.35E-33	1.37E-12
8	Eos%	0.840582	0.513	20	Pct	5.11E-06	5.89E-05
9	Baso%	1.82E-24	4.20E-14	21	PDW	8.04E-05	0.2
10	Neut#	3.68E-07	0.008	22	hsCRP	0.003651	0
11	Lymph#	0.058221	9.83E-13	23	PCT	0.059937	5.66E-07
12	Mono#	4.61E-05	1.33E-05				

3.2 基于机器学习的鉴别算法 采用数据特征直方图、密集分布图对数据进行分析，大部分数据存在一定程度的偏态分布，需要对数据进行正态化处理，通过Box-Cox转换提高模型的准确度，最终对六种线性与非线性方法、四种集成算法的分类效果评估，参数调整，选择最优模型。

3.2.1 线性与非线性算法的评估 通过设计一个评估框架来选择合适的鉴别算法，针对逻辑回归算法（LR）、线性判别分析（LDA）、分类与回归树算法（CART）、支持向量机（SVM）、贝叶斯分类器（NB）和K近邻算法（KNN）六种算法，采样10折交叉验证来分离数据，并通过准确度比较算法，以找到最优算法。

通过对六种线性与非线性算法进行评估，通过图1能够看到每次执行的分布情况，SVM算法执行结果分布相对紧凑，可能具有更好的效果，选择SVM算法调参，在对新冠肺炎与甲型流感鉴别分类时，惩罚系数C=2.0，径向基函数kernel为rbf时，具有最优的准确率为0.953；对新冠肺炎与乙型流感鉴别分类时，惩罚系数C=1.7，径向基函数kernel为rbf时，具有最优的准确率为0.944。

3.2.2 集成算法的评估 为提高算法准确度，选择集成算法中两种装袋算法随机森林（RF）与极端随机树算法（ET），两种提升算法AdaBoost（AB）与随机梯度上升（GBM）进行比较，仍然采用10折交叉验证方法来验证集成算法准确度。

通过对四种集成算法进行评估，算法结果的离散情况如图2所示，选择随机梯度上升（GBM）算法进行进一步优化调参，在对新冠肺炎与甲型流感鉴别分类时，estimators=50时，具有最优的准确率为0. 926；对新冠肺炎与乙型流感鉴别分类时，estimators=300时，具有最优的准确率为0. 883。

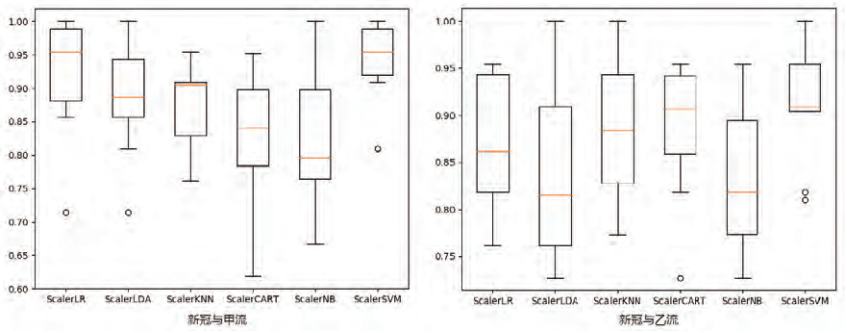

图1 线性与非线性算法评估结果

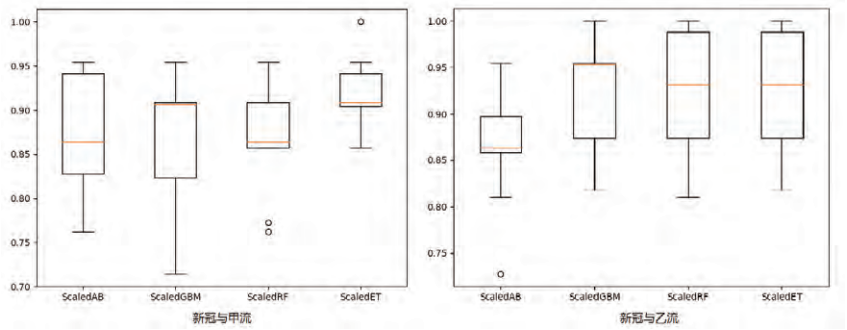

图2 集成算法评估结果

表3 鉴别分类预测结果

项目	精确率	召回率	F1-score
新冠 & 甲流	0.97	0.96	0.96
新冠 & 乙流	0.94	0.95	0.95

3.2.3 确定预测模型 根据对算法的评估结果，支持向量机SVM具有更好的准确性，选择SVM算法对新冠肺炎与甲型流感、乙型流感进行鉴别分类。从表3可以看出，SVM对正态化的数据具有较高的准确度，新冠肺炎与甲型流感预测精确度达到0.97，新冠肺炎与乙型流感的预测精确度达到0.94。

4 结论

根据流感与新冠肺炎检验指标数据集特征、维度、数据类型、分布状态等特性，对数据集采取多种处理措施，有效降低数据的噪声与冗余；并利用独立样本t检验的方法对数据集进行特征差异性分析，最终采用多种线性与非线性、集成算法等机器学习方法对数据进行训练，评估能够对流感与新冠肺炎的最优分类算法，以验证利用检验指标及机器学习方法在流感与新冠肺炎快速鉴别问题上的可行性，并根据选择的最优模型对数据进行分类预测。

参考文献

[1] Huang C,Wang Y,Li X,et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan,China[J]. Lancet,2020,395:497-506.

[2] 中华人民共和国国家卫生健康委员会.新型冠状病毒肺炎诊疗方案(试行第七版)[A/OL]. 国卫办医函〔2020〕184号.

[3] 中国疾病预防控制中心新型冠状病毒肺炎应急响应机制流行病学组.新型冠状病毒肺炎流行病学特征分析[J].中华流行病学杂志,2020,41(2):145-151.

[4] Fred Sun Lu,Suqin Hou,Kristin Baltrusaitis,et al.Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams:A Case Study in the Boston Metropolis[J].Jmir Public Health

（下转第52页）

表2 不同职称用户使用效果（$\bar{x}\pm s$）

项目	主治及以下（n=11）	副主任及以上（n=7）	F值	P值
是否减轻工作量	1.27±0.47	1.29±0.49	0.003	0.956
是否规范临床诊疗行为	1.09±0.30	1.14±0.38	0.105	0.751
是否提高病历质量	1.18±0.40	1.43±0.53	1.244	0.281
是否防漏防误	1.09±0.30	1.57±0.53	6.024	0.026
静态知识是否有用	1.18±0.40	1.29±0.49	0.241	0.630

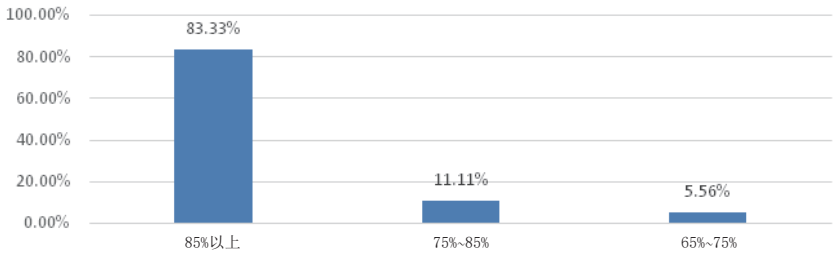

图4 临床辅助决策系统推理准确性

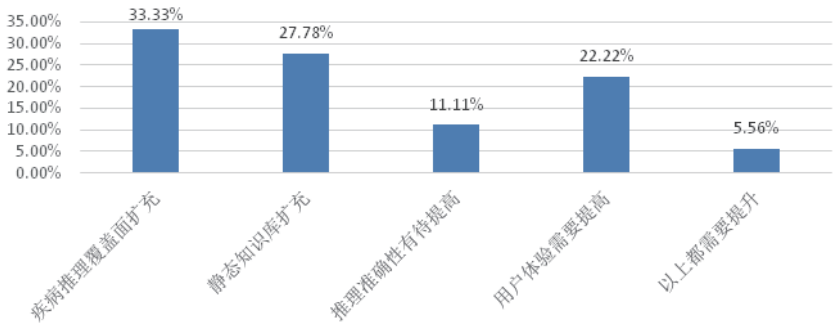

图5 临床辅助决策系统待优化提升的模块

准度；从人工智能辅助决策系统使用和准确性来看，该系统的应用普遍获得了临床医生的肯定，可以实现为临床医生提供更加智能、精细、多样的决策支持服务，在医疗领域里有着广阔的应用前景，对于提升医疗服务水平有着极为重要的意义[5]。值得注意的是系统的使用效果和体验按职称是分层次的，副主任及以上的医师对系统的效果普遍低于主治及以下的医师；基于大数据的临床辅助决策系统的构建和应用还需要在疾病推理覆盖面、用户体验、静态知识库等方面进行优化和提升，进一步完善系统应用范围和应用场景，从使用者的角度提高用户体验，将真正的需求和痛点整合到辅助决策系统内。

基于大数据构建的临床人工智能辅助决策系统，可以多维度地帮助临床医生作出临床诊断，为患者提供便利化和智能化的医疗服务。在看到应用前景的同时，也要认识到临床辅助决策系统仍处于“弱人工智能”阶段，必须要加强要树立用户至上、安全为本的设计理念，加大技术研发力度，推动人工智能技术不断走向成熟和完善。

参考文献

[1] 孙扬,李迪,舒琴,等.深度学习在医疗辅助决策领域的应用[J].中国病案,2019,20(11):28-31.
[2] 杨晓华,是俊风.流程化进阶管理在病案质量监控中的应用[J].中国卫生质量管理,2018,15(4):26-28.
[3] 连其平.新形势下病案信息的价值与管理[J].中华全科医学,2018,7(6):757-758.
[4] 谢锦艳,冯月明.增强法律意识重视病案管理[J].实用全科医学,2016,4(3):318.
[5] 陈绮钿,刘琛玺,李富强,等.病历质控系统在电子病历中的应用[J].中国数字医学,2016,11(6):108-110.

【收稿日期：2020-07-13】
（责任编辑：肖婧婧）

（上接第23页）

& Surveillance,2018,4(1):e4.
[5] 刘思齐,蒋龙元.流感样症状患者临床特征分析[J].中华卫生应急电子杂志,2019,5(3):133-135.
[6] Ko J,Baldassano SN,Loh PL,et al. Machine learning to detect signatures of disease in liquid biopsies - a user's guide[J].Lab on A Chip,2018,18(3):395
[7] Schneider WF,Guo Hua. Machine Learning[J].Journal of Physical Chemistry A,2018,122(4):879.
[8] Cheng M,Li LM,Ren YF,et al.A Hybrid Method to Extract Clinical Information From Chinese Electronic Medical Records[J].IEEE Access,2019(7):70624-70633.
[9] Ren YF,Fei H,Liang XH,et al.A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records[J].BMC Medical Informatics and Decision Making,2019,19(Suppl 2):51.
[10] Pan L,Jian-bo G,Javier PT.CT findings and clinical features of pancreatic hemolymphangioma:a case report and review of the literature[J]. Medicine,2015,94(3):e437.
[11] 葛晓伟,李凯霞,程铭.基于CNN-SVM的护理不良事件文本分类研究[J].计算机工程与科学,2020,42(1):161-166.
[12] 毕京峰,魏振满,王鹏.临床研究中应用T检验存在的典型错误辨析[C]//中医药现代化国际科技大会,2010.

【收稿日期：2020-03-09】
（责任编辑：张倩）