基于随机森林模型与 SHAP 算法的流感样病例影响因素分析研究

李 进[1] 魏艳龙[1] 薛红新[2]
LI Jin WEI Yanlong XUE Hongxin

摘 要 通过机器学习算法和 SHAP（shapley additive explanations）方法分析影响流感样病例（influenza-like illness，ILI）的主要因素，挖掘和流感相关的关键词汇，进行初步筛选，再通过分析这些关键词与 ILI 序列的时滞相关性，对筛选出的关键词进行二次过滤。首先使用关键词变量分别构建支持向量回归、XGBoost 模型、随机森林回归；然后应用 SHAP 方法进行模型可解释性分析，探讨影响 ILI 的主要关键词；最后利用随机森林回归方法构建的预测模型具有最高预测性能，其平均绝对百分比误差 M_{APE} 是 0.141。模型显示“流感疫苗副作用、流感疫苗、感康、流行性感冒、康泰克、感冒”是预测 ILI 值的重要影响因素。随机森林回归方法所构建的预测模型能够更准确地预测流感样病例值，结合 SHAP 方法可以对 ILI 值预测提供详细解释，有助于医疗机构制定有效的应急措施。

关键词 流感预测；百度指数；SHAP；支持向量回归；随机森林回归；XGBoost

doi：10.3969/j.issn.1672-9528.2024.02.001

0 引言

流感疫情是一个严重的公共卫生问题，每年在全球范围内都引起重大疾病，导致巨大的经济损失和人员死亡。提前预测流感疫情对于预防 ILI 和合理分配医疗资源至关重要。根据中国国家流感中心的常规监测数据报告的流感活动通常滞后 1 ～ 2 周。因此，作为及时估计流感活动和检测疫情的便捷来源，网络搜索数据有助于改善传统流感监测的结果[1]。

利用互联网搜索数据进行公共卫生事件监测能迅速预测流行病传播态势，为医疗部门提供疾病传播相关情报[2]。Ning 等人[3]提出了 ARGO-C（augmented regression with clustered google data），这是一种综合的、符合统计学原理的方法，它结合了互联网搜索数据的聚类结构，以提高流感预测的准确性。De 等人[4]提出一种基于个性化页面排名（PageRank）和循环排名（CycleRank）的语言识别方法，自动选择与 ILI 最相关的 Wikipedia 页面进行监测，并利用页面信息来准确估计欧洲国家流感样疾病发病率的可行性。国外相关研究主要采用 Google、Twitter 和 Wikipedia 的搜索引擎数据[5-7]，而在国内主要互联网搜索工具有百度指数和微指数[8-9]。

机器学习方法如线性回归、人工神经网络、决策回归树等[10-12]被广泛用于预测流感。但部分机器学习算法存在“黑箱”机制，这对于研究人员来说并不清楚样本特征如何影响最终预测值，而 SHAP 算法属于事后解释方法，对具有“黑盒”机制的模型有较强的解释性。Cao 等人[13]采用相关性分析和可解释的机器学习算法来评估疫情流行期间各因素对新增病例和 COVID-19 增长率的定量贡献。Luo 等人[14]提出一种可解释空间标识（ISID）神经网络来预测区域周级的传染病数量，使用 SHAP 方法来解释 ISID 模型，结果表明 ISID 模型和 SHAP 算法为流行病预测提供可解释性分析。SHAP 方法被广泛用于机器学习模型解释传染病影响因素。但是，使用 SHAP 算法进行关键词因素分析的流感预测研究较少。

综上，本研究以百度指数提供的网络搜索数据和流感样病例数据，结合随机森林回归模型和 SHAP 算法探索关键词和流感的关系，以期成为流感疫情防控提供参考。

1 资料与方法

1.1 资料来源

1.1.1 中国国家流感中心流感周报

本文使用的 ILI 数据是中国国家流感中心发布的流感样病例，每年共发布 52 周的流感数据。这些数据主要依靠监测哨点医院，整合分布在全国各地的哨点医院报告的流感数据。样本时间段为 2017 年 4 周（2017 年 1 月 29 日） 至 2020 年 35 周（2020 年 8 月 30 日），共 188 周的 ILI 数。

1.1.2 百度指数

百度指数是一种衡量关键词搜索热度的统计指标，通过分析互联网用户在百度上的搜索量，以特定关键词为统计对象，计算其在百度网页搜索中的搜索频率加权。指数越高，表示用户对特定关键词的关注程度越高。

1. 太原师范学院 山西晋中 030619
2. 中北大学 山西太原 030051
[基金项目] 国家自然科学基金（62106238）；省高等学校科技创新项目（2020L0283）；山西省基础研究计划（202203021212185）

1.2 关键词选择

根据流感患病过程，本研究首先从流感预防阶段、治疗手段和常用药物等方面选择初始搜索词，从已有的相关研究中初步选择与 ILI 有关的 37 个检索词。

1.2.1 关键词时滞相关性分析及筛选

关键词序列与 ILI 序列在时间上有相关关系。为了预测未来几周的 ILI，采用时差相关分析法可识别出 ILI 与关键词的时序特征，其中具有领先特征的关键词可以预测未来。此外，各个关键词与 ILI 滞后时间不同，因此研究选取各关键词中相关系数最高的滞后指标来构建各自模型。

研究使用 Python3.6 进行 Pearson 相关性分析，初筛关键词的滞后时间序列与 ILI 序列的 Pearson 相关系数需不低于 0.5。对每个关键词搜索指数与 ILI 之间的相关性排序分析发现，37 个关键词中有 22 个关键词滞后序列与 ILI 的相关系数小于 0.5，15 个关键词滞后序列与 ILI 相关系数大于 0.5。

1.3 方法

1.3.1 支持向量回归（support vector regression，SVR）

SVR 是一种基于支持向量机的回归方法。与传统的回归方法不同，SVR 的目标是找到一个在高维特征空间中最好的拟合数据的超平面 [15]。它通过最小化预测值与实际值之间的误差平方和，同时考虑到支持向量与超平面之间的最大间隔，从而实现回归任务。SVR 可以应用于线性和非线性回归问题，并且在处理小样本和高维度数据时具有较好的泛化性能。本研究使用网格搜索算法进行 SVR 参数寻优。

1.3.2 极端梯度提升算法 (extreme gradient boosting，XGBoost)

XGBoost 是一种采用 Boosting 思想的决策树集成方法。它通过学习外部变量之间的关系，生成决策树预测一个值，得到预测值与真实值之间的误差，再添加一棵树学习该误差，最终累计多棵树的预测值作为预测结果。XGBoost 模型通过二阶泰勒展开损失函数，能更高效地求得模型最优解，具有一定可解释性。

1.3.3 随机森林回归（random forests regression，RFR）

RFR 是采用 Bagging 方式的决策树集成方法。它由多棵决策树形成的组合预测模型，用随机的方式对样本进行训练和预测。当输入待预测的样本时，最终的预测结果为多个决策树输出结果的平均数。

1.3.4 基于 SHAP 的模型可解释性分析

SHAP 是 Lundberg 等人 [16] 提出的一种基于 Shapley value 的计算，用于解释黑盒模型是如何衡量特征对最终结果值的影响。计算公式为：

$$\varphi_i(v)=\sum_{s\subseteq\{x_1,\ldots,x_p\}/\{x_j\}}\frac{|s|!(p-|s|-1)!}{p!}(v(s\cup\{x_j\})-v(s)) \quad (1)$$

式中：$\varphi_i(v)$ 是特征 i 的 shapley 值，即对预测值的贡献度；s 是模型中使用的特征的子集；x 是要解释的样本特征值的向量；p 为特征的数量；（$|s|!(p-|s|-1)!$）/$P!$ 表示权重，$v(s)$ 指在特征组合 s 下的模型输出值。

假设 g 是解释模型，M 是外部变量的数目，z 表示该特征是否存在（取值 0 或 1），φ 为每个特征的 Shapley 值，φ_0 是指所有样本的预测均值，则有以下公式：

$$g=\varphi_0+\sum\varphi_i Z \quad (2)$$

采用 SHAP 值对模型进行解释，能更好地了解训练过程中特征的贡献度，从而为预防流感提供一定的参考价值。本文使用 Python3.6 中的 SHAP 工具包对流感预测模型各关键词的影响力进行分析。基于 SHAP 值，对训练集中关键词的样本数据进行计算，得到所有周中关键词的 SHAP 值，为预测流感提供相关的评估和解释。

1.3.5 模型效果评价指标

为了评估各模型的预测效果，采用均方误差（mean absolute error，M_{AE}）、均方根误差（root mean squarederror，R_{MSE}）和平均绝对百分比误差（mean absolute percentage error，M_{APE}）三个指标对各模型预测结果进行衡量，相应的计算公式为：

$$M_{AE}=\frac{1}{n}\sum_{i=1}^{n}|y_i^{'}-y_i| \quad (3)$$

$$R_{MSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{'}-y_i)^2} \quad (4)$$

$$M_{APE}=\frac{100\%}{n}\sum_{i=1}^{n}|\frac{y_i^{'}-y_i}{n}| \quad (5)$$

式中：n 表示样本个数，y_i 表示 ILI 的真实值，$y_i^{'}$ 表示 ILI 的预测值。

2 结果

2.1 关键词分类结果

通过滞后性相关分析结果，关键词在时间上有三种性质类别。若关键词与流感序列时间同步，则称为“同步”关键词，该关键词可实时对流感做出预测。若关键词在时间上提前于流感序列，则称为“先行”关键词，说明使用该词可提前预测流感。本研究中使用先行关键词作为预测变量。在初筛关键词的时滞相关性分析中，5 个关键词在时间同步时相关系数达到最大值，被归类为“同步”关键词；另外，10 个关键词在时间提前时相关系数达到最大值，被归类为“先行”关键词，具体的关键词分类结果见表 1。

表 1 关键词分类结果

类别	关键词
“同步”关键词	肠胃感冒、流感疫苗有必要打吗、高烧、发烧、流感吃什么药
“先行”关键词	流感疫苗副作用、流感疫苗、流行性感冒、感冒、流感治疗、鼻塞、康泰克、感康、泰诺、白加黑

2.2 三种模型的评价效果

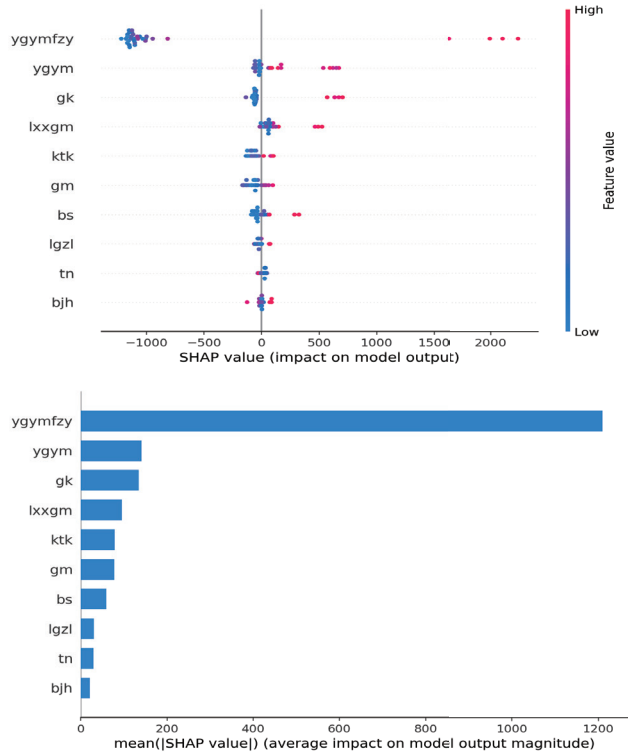
SVR、RFR、XGBoost 模型均使用筛选后的关键词变量进行建模，本研究选用 2017 年 4 周 至 2020 年 3 周共计 156 周的数据作为训练样本，选择 2020 年 4 周至 2020 年 35 周共 32 周的数据作为测试样本。采用 M_{AE}、R_{MSE}、M_{APE} 模型效能比较。结果显示 RFR 预测效果最好，M_{APE} 为 0.141，认为模型具有较好的预测性能。模型评价指标的比较结果，如表 2 所示。

表 2 SVR、XGBoost、RFR 模型的评价指标结果

模型	M_{AE}	R_{MSE}	M_{APE}
SVR	1.418e+03	1.687e+03	2.504e-01
XGBoost	9.098e+02	1.272e+03	1.561e-01
RFR	8.248e+02	1.339e+03	1.413e-01

2.3 基于 SHAP 的模型解释性分析

基于 SHAP 模型，对流感预测模型的结果进行解释性分析。图 1 为基于 RFR 的流感预测模型的特征重要性排序图和 SHAP 概括图。



注：ygymfzy 表示“流感疫苗副作用”；ygym 表示“流感疫苗”；lxxgm 表示“流行性感冒”；lg 表示“感冒”；lgzl 表示“流感治疗”；bs 表示“鼻塞”；ktk 表示“康泰克”；gk 表示“感康”；tn 表示“泰诺”；bjh 表示“白加黑”

图 1 特征重要性排序图和 SHAP 摘要图

如图 1 所示，图中每一个点代表一个样本，以 SHAP 值取零为中间分界线，对于处在左侧的样本点，对流感预测值表现为负向贡献，处在右侧的样本点表现为正向贡献。颜色越接近红色表示特征本身数值越大，反之接近蓝色表示特征本身数值越小。“流感疫苗副作用、流感疫苗、感康、流行性感冒、康泰克、感冒”等特征对模型影响较大。其中“流感疫苗副作用”是影响流感的最重要特征，随着“流感疫苗副作用”的增加，流感人数变化绝对值不断增大。除此之外，SHAP 还可对某一周流感人数的影响因素进行分析。

图 2 是 2020 年 14 周流感人数的 SHAP 特征贡献图，蓝色部分表示为该周流感人数的负向影响因素，主要因素有“流感疫苗副作用、康泰克、感康、流感疫苗”等。

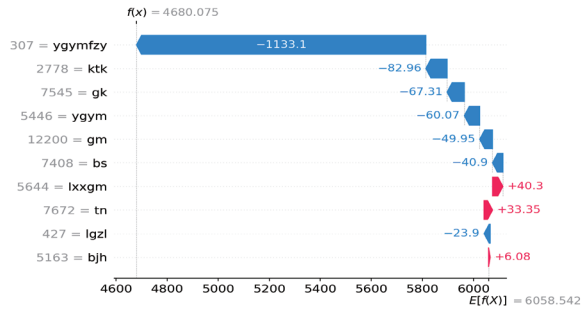


图 2 2020 年第 14 周 ILI 的 SHAP 解释示例

图 3 是 2020 年第 8 周流感人数的 SHAP 特征贡献图，红色部分表示该周流感人数的正向影响因素，其影响因素是“鼻塞、流感疫苗、流行性感冒”。

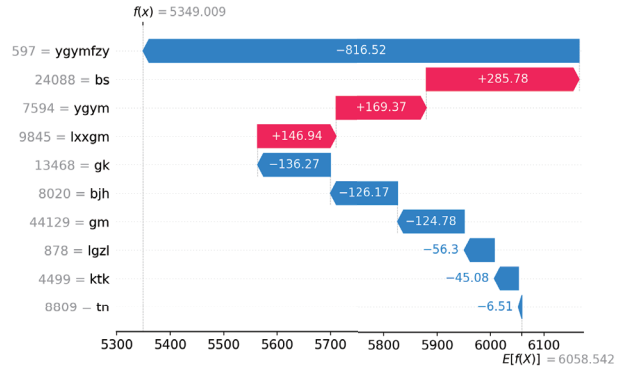


图 3 2020 年第 8 周 ILI 的 SHAP 解释示例

3 讨论

在本文中，构建基于 RFR 的流感预测模型来评估流感发病人数。与 SVR、XGBoost 相比，利用关键词构建 RFR 的预测模型具有最好预测效果。

具体来说，根据关键词序列与 ILI 序列的滞后相关性，筛选出 10 个影响流感人数的相关因素，其中“流感疫苗副作用”在全局训练过程中表现的特征重要度对流感的影响最大。研究[17]表明“泰诺”“康泰克”“流感治疗”“感冒”“流行性感冒”“高烧”“流感疫苗副作用”“流感疫苗有必要打吗”和“流感疫苗”等是预测流感的重要因素。本文也表明“流感疫苗副作用、康泰克、流感疫苗、感冒、流行性感冒”对流感预测的重要性程度。本文分析关键词与流感序列之

间的相关性，并选用最佳特征变量提高模型性能，最后用SHAP 算法解决传统机器学习的“黑箱”缺点，从而直观地展示模型中各关键词对结果的影响。另外本文具有两个主要的局限性，首先是流感数据及关键词信息为全国性范围数据，数据粒度细化后的结果需要进一步研究来证实。例如，文献 [18-19] 使用省市级的流感数据源，增强数据细化程度，为省市区域范围的预测流感提供更加准确的预测效果。其次，数据类型单一仅有关键词类型，流感序列结合天气信息可进一步提高模型预测性能。文献 [20-21] 使用 PM2.5、SO_2 等空气污染物，探讨了流感样病例与空气污染物之间的潜在关系。

总之，本文利用关键词信息，构建基于 RFR 的流感预测模型具有较好的预测性能，从全局训练过程和局部分析过程帮助医院判断与流感相关的主要关键词，有利于开展针对性干预措施，对实现精准医疗和卫生资源优化配置具有重要意义。

参考文献：

[1] GUO P, ZHANG J, WANG L, et al. Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model[J]. Scientific reports, 2017, 7:1-11.

[2] 薛红新 , 刘绕星 , 况立群 , 等 . 基于百度指数与机器学习方法的流感样病例预测 [J]. 中国预防医学杂志 , 2023, 24(8):788-794.

[3] NING S, HUSSAIN A, WANG Q. Incorporating connectivity among internet search data for enhanced influenza-like i-llness tracking[EB/OL].(2023-08-20)[2023-09-06]. https://arxiv.org/abs/2308.10091.

[4] DE TONI G, CONSONNI C, MONTRESOR A. A general method for estimating the prevalence of influenza-like-symptoms with wikipedia data[J].Plos one,2021,16(8):1-20.

[5] NSOESIE E O, OLADEJI O, ABAH A S A, et al. Forecast-ing influenza-like illness trends in Cameroon using Google Search Data[EB/OL].(2021-03-24)[2023-09-07].https://www.nature.com/articles/s41598-021-85987-9.

[6] SHARPE D, HOPKINS R, COOK R L, et al. Using a bay-esian method to assess google, twitter, and wikipedia for ili surveillance[J].Online journal of public health informatics, 2017, 9(1): 25-29.

[7] HICKMANN K S, FAIRCHILD G, PRIEDHORSKY R, et al. Fo-recasting the 2013–2014 influenza season using wikipedia[J].Plos computational biology, 2015, 11(5):194-209.

[8] 鲁力 , 邹远强 , 彭友松 , 等 . 百度指数和微指数在中国流感监测中的比较分析 [J]. 计算机应用研究 , 2016, 33(2):392-395.

[9] 张晋毓 . 微博舆情数据分析在我国疫病监控中的应用 [D]. 湘潭 : 湘潭大学 , 2019.

[10] KARA A. Multi-step influenza outbreak forecasting using deep LSTM network and genetic algorithm[J]. Expert systems with applications, 2021, 180: 115153.

[11] TSAN Y T, CHEN D Y, LIU P Y, et al. The prediction of in-fluenza-like illness and respiratory disease using LSTM and ARIMA[J]. International journal of environmental research and public health, 2022, 19(3):1-17.

[12] WU H, CAI Y, WU Y, et al. Time series analysis of w-eekly influenza-like illness rate using a one-year period of factors in random forest regression[J]. Bioscience trends, 2017, 11(3): 292-296.

[13] CAO Z, TANG F, CHEN C, et al. Impact of systematic f-actors on the outbreak outcomes of the novel COVID-19 disease in China: factor analysis study[J]. Journal of medical Internet research, 2020, 22(11): e23853.

[14] LUO L, LI B, WANG X, et al. Interpretable spatial ide-ntity neural network-based epidemic prediction[J]. Scientific reports, 2023, 13(1): 18159.

[15] 陈昱吉 , 成贵学 . 基于 ICOA 和 SVR 的短期负荷预测 [J]. 计算机仿真 , 2022, 39(11):65-69.

[16] LUNDBERG S , LEE S I .A unified approach to interpr-eting model predictions[C]//Conference and Workshop on Neural Information Processing Systems. California:NIPS Pre-ss, 2017: 4765-4774.

[17] 王若佳 . 融合百度指数的流感预测机理与实证研究 [J]. 情报学报 ,2018,37(2):206-219.

[18] 庄雅丽 , 卢捷 , 吴树凯 , 等 . 广东省 2015-2022 年流感暴发疫情特征分析 [J]. 中华流行病学杂志 ,2023,44(6):942-948.

[19] 邓源 , 任翔 , 郭玉清 , 等 .2008-2020 年我国北方 15 个城市流感与气象因素的关联性研究 [J]. 中华流行病学杂志 , 2023, 44(5):765-771.

[20] LIU X X, LI Y, QIN G, et al. Effects of air pollutants on oc-currences of influenza-like illness and laboratory-confirmed influenza in Hefei, China[J]. International journal of biometeorology, 2019, 63: 51-60.

[21] SU W, WU X, GENG X, et al. The short-term effects of air pollutants on influenza-like illness in Jinan, China[J]. BMC public health, 2019, 19(1): 1-12.

【作者简介】

李进（1998—），男，山东临沂人，硕士研究生，研究方向：机器学习。

魏艳龙（1982—），男，山西太原人，副教授，硕士生导师，主要研究方向：超高温环境下瞬态温度参数测试。

薛红新（1991—），女，山西吕梁人，讲师，博士，主要研究方向：人工智能、计算机视觉。

（收稿日期：2023-10-22）