基于气象因素和机器学习的流感预警研究*

金丽珠 葛 辉 万 明 王晓风 杜雪杰

(中国疾病预防控制中心 北京 102206)

〔摘要〕 运用机器学习方法挖掘气象因素对流感发病的影响和作用，构建流感预测预警模型。详细阐述模型构建方法及步骤，包括数据采集及预处理、特征构造和选择、具体构建，分析模型预测预警效果，为流感防治工作提供技术支持和参考。

〔关键词〕 气象因素；流感；大数据；机器学习；预警

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2021.07.004

Study on the Early Warning of Influenza Based on Meteorological Factors and Machine Learning *JIN Lizhu, GE Hui, WAN Ming, WANG Xiaofeng, DU Xuejie, Chinese Center for Disease Control and Prevention, Beijing 102206, China*

〔**Abstract**〕 The machine learning method is used to excavate the influence and effect of meteorological factors on the occurrence of influenza, and the influenza prediction and early warning model is built. The paper expounds the methods and steps of model building in detail, including data collection and preprocessing, feature construction and selection and concrete building, analyzes the prediction and early warning effect of the model, so as to provide technical support and references for influenza prevention and control.

〔**Keywords**〕 meteorological factors; influenza; big data; machine learning; early warning

1 引言

流行性感冒是由流感病毒引起的急性呼吸道传染性疾病，具有起病急、传播快、感染性强等特点，对其快速反应与防范依赖于及时有效的监控和准确快速的预测预警，这在全球范围内仍是巨大挑战[1]。近年来随着互联网以及多学科多领域数据、知识的交互发展，大数据技术日趋成熟，与流感相关的海量数据挖掘及分析技术已得到广泛应用，基于大数据的流感预警在提升疫情追踪、响应和预测预警能力方面成效显著[2]。目前气象变量与流感传播相关性方面研究较多，机器学习算法广泛应用于流感预测，如筛选气温、降雨量、相对湿度等重要气象因素，利用神经网络对流感暴发做出预测能够一定程度减少误差。然而采用气象因素预警界值方式针对不同程度气象条件做出不同级别精准预警的研究相对较少。本文利用机器学习算法将多源大数据气象因素和某地区流感发病数据进行整合，基于数据可视化方法，采用建立阈值方式构建预测预警

〔收稿日期〕 2020-12-23

〔作者简介〕 金丽珠，助理研究员，发表论文 9 篇；通讯作者：葛辉，副研究员。

〔基金项目〕 国家科技重大专项“基于海量多元大数据的突发急性传染病时空多尺度预测预警模型与应用示范的构建——基于突发急性传染病多元数据的云服务平台构建”（项目编号：2018ZX10201-002-001）。

模型从而提升预警效果，以期为流感防治工作提供技术支持和参考。

2 数据来源

2.1 主要实验数据

为全国温室数据系统气象数据和国家传染病监测系统流感发病数据。流感患病及传播与季节、地域、气象与环境、人口学因素、人类行为等方面密切相关，其中气象因素是影响区域流感发病的关键因素。全国温室数据系统包含全国气象站点温度、气压、气温、相对湿度、降水、蒸发、风速等气象数据，系统数据更新快、方便获取。

2.2 流感监测网络数据

我国传统流感监测网络于 2000 年建成并在 2009 年实现对全国所有地市级地区覆盖。至今该监测网络已积累大量不同来源、结构的流感监测数据[3]。本研究采用国家传染病监测系统中某市 2012－2016 年流感日发病历史数据。

3 基于气象因素的流感大数据预测预警模型构建方法（图 1）

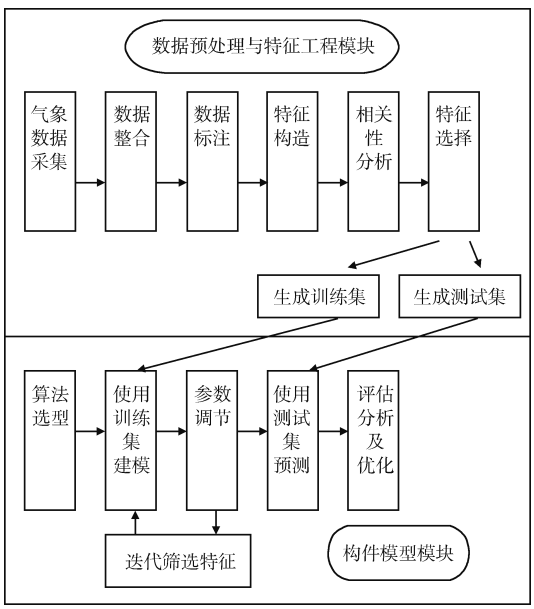

图 1 流感预测模型构建流程

3.1 数据采集及预处理

3.1.1 概述 由于原始数据具有不完整性且含有无关或冗余特征，直接进行机器学习结果不准确且效果较差。数据不完整性主要表现为缺少对实验效果有较大影响的属性值，这部分数据可以通过数据挖掘方式获取以提升实验效果；含有无关特征是指部分数据对于本文算法没有帮助，不能提升算法效果；含有冗余特征是指数据无法为本文算法带来新信息，或者这种特征的信息可以由其他特征推断出来。具体操作流程为：首先通过互联网采集大量气象数据，采用统计学或机器学习等方法检测无关和冗余数据，然后删除掉这部分特征数据，充分提取有效信息。

3.1.2 数据采集 网络爬虫又名网络蜘蛛、网络机器人，是一种按照一定规则自动抓取万维网信息的程序或脚本。其在网页上模拟人的行为，如点击网页和查看网页数据，以抓取数据并存放于存储介质中。本研究采用 Python 语言在全国温室数据系统进行数据采集。采集得到的初始气象因素有 t_ avg（日平均气温）、t_ max（日最高气温）、t_ min（日最低气温）、precip（20 至 20 时累计降水量）、winds_ avg（平均风速）、winds_ max（最大风速）、rh_ avg（平均相对湿度）、rh_ min（最小相对湿度）、QNE_ hPa（平均气压）和 radiation（日累计辐射）10 个特征数据。

3.1.3 数据标注 要找到预警阈值有两个基本逻辑和考虑方向可供选择。一是将每日流感发病数作为解释变量，即因变量，将问题作为机器学习的回归问题看待和解决，训练模型并对未来流感发病人数做预测，当预测发病人数大于某个阈值时发出预警。二是先通过特定数据标注方法将流感发病数这种连续数值转化为 0 和 1 的离散型标签，0 代表不需要预警，1 代表需要预警。数据标注方法是对流感暴发进行定义，衡量当前发病数是否能代表流感已暴发。标注完成后即可将问题作为机器学习中的分类问题进行解决，对数据做预测则将预测为 1 的日期看作需要发出预警的日期。两种思路比较而言，由于流感发病具有典型季节性特征，第 1 种思

路将预测得到的连续值与特定阈值相比较是不合理的，如果按季节划分训练多个模型则可能将问题复杂化；第 2 种思路易于理解和实现。综合考量选择第 2 种思路作为解决问题的核心思路。数据标注属于分类标注，是数据标注的常见类型，即打标签。一般从既定标签中选择数据对应标签，是封闭集合。例如在自然语言处理领域，对于一句话中的文字可以标注主语、谓语、宾语、名词、动词等。将发病数据标记为 label = 0 或 label = 1，这是训练二分类模型的基础，以便进行有监督的分类学习。对于流感暴发这一具体问题，本文提出两种数据标注方法，一是移动百分位数法，二是月度四分位标记法。具体而言，移动百分位数法将当地当前观察周期内病例数与其相应历史基线数据进行实时比较，当前观察周期内发生的病例数达到或超过预警阈值时则认为流感暴发，将数据标签定义为 1。例如回溯历史年数为 3 年，计算周期为 7 天，按天移动，历史同期前后摇摆 2 个参比周期，假设流感爆发的预警阈值为 P80，则只有在当前观察周期（7 天）内病例数不小于历史基线数据中的 80% 时才将 label 设置为 1，否则为 0。月度四分位标记法是将每月发病数超过月度四分位数的日期对应的数据标签定义为 1。

3.2 特征构造和选择

3.2.1 特征构造 重点考虑所设计新特征是否对目标有价值，后续再对特征重要性进行评估。简单来说就是对原特征进行转换，原特征不变，将转换后特征加入到原数据中，通过增加对结果有影响的因素提高模型准确率。如果设计了冗余特征，后续可通过特征选择（相关性分析实验）删掉其中多余特征即可。特征设计需要领域知识、直觉和数学知识。特征设计和选择需要反复迭代验证才能得到更好结果。以获取的基础气象数据为基础，主要考虑气象因素对流感发病时间的滞后性影响，构造得出的新特征有 t_ avg_ day1ago（昨日平均气温）、t_ avg_ day3ago（3 日前平均气温）、rh_ avg_ day1ago（昨日平均相对湿度）、pre_ avg_ day1ago（昨日累计降水量）、wi_ avg_ day1ago（昨日平均风速）、QNE_ day1ago（昨日平均气压）、warning_ if_ day3s（过去 3 天是否有预警）和 warning_ if_ day7s（过去 7 天是否有预警）等 55 个特征数据。

3.2.2 特征选择 对于特定学习算法来说特征是否有效是未知的。因此需要从所有特征中选择对于学习算法有益的相关特征，筛除无关和冗余特征，防止出现维度灾难问题。通过特征选择可以降低学习任务难度、提升模型效率。本研究使用过滤法和嵌入法组合算法进行相关性分析和特征选择。首先使用过滤法进行初步筛选，再使用嵌入法做进一步特征筛选。其中过滤法又分为方差分析法和互信息分析法。对于方差选择法，方差较大的特征较为有用，相反则对算法作用较小。如果某个特征方差为 0，即所有样本该特征取值相同，那么它对模型训练无作用。在实际应用中会指定一个方差阈值，方差小于该阈值的特征将被直接筛掉。除使用方差过滤选择特征外，还可使用互信息等其他统计学指标。互信息从信息熵角度分析各特征和输出值之间的关系评分。互信息值越大说明该特征和输出值之间相关性越大，越需要保留，应删除和输出值之间互信息较小的特征。对于决策树模型以及基于树的集成学习模型，可根据每个特征对应的 coef_ 或 feature_ importance_ 属性值进一步筛选特征。对通过过滤法筛选得到的初步特征集进行再次筛选，将决策树算法作为嵌入法的机器学习完成特征选择，根据特征重要性系数筛选特征，得到最终特征集。使用过滤式与嵌入式组合算法进行特征选择，见图 2。首先，计算出所有特征在方差分析法下的排名 n，方差越大排名越靠前；计算出所有特征在互信息分析法下的排名 m，互信息越大排名越靠前。令方差分析法权重 $\lambda 1 = 0.25$，互信息分析法权重 $\lambda 2 = 0.75$，计算出所有特征加权排名 $\lambda 1 * n + \lambda 2 * m$，加权排名越靠前的特征越需要保留，筛除加权排名最靠后的 10 个特征，完成特征初步筛选。其次，使用嵌入法再次筛选特征：设置阈值为 0.016，筛除所有权值系数 feature_ importance 不大于此阈值特征，得到最终特征子集。经过特征选择筛除 winds_ max（最大风速）、QNE_ hPa（平均气压）、t_ avg_ ins（日平均气温温差）、precip_ avg_

day1ago（昨日累计降水的平均）、winds_ avg_ day1ago（昨日平均风速）、rh_ avg_ lastweek（上周平均相对湿度）、QNE_ lastweek（上周平均气压）和 warning_ if_ week（过去 1 周是否有预警）等 26 个特征，保留 t_ avg（日平均气温）、t_ max（日最高气温）和 precip（日降水量）等 29 个特征。后续使用模型训练时将反复迭代验证，进一步筛除对结果影响较小的特征，只选择少量特征建立模型。

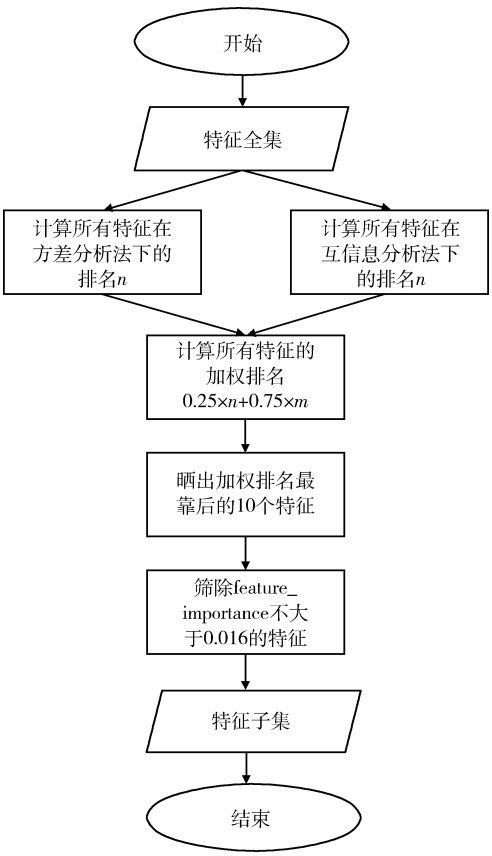

图 2 特征选择流程

3.3 模型构建

3.3.1 *算法选择* 决策树算法既可作为分类算法也可作为回归算法，特别适合集成学习，如梯度提升决策树（Gradient Boosting Decision Tree，GBDT）算法。决策树算法基本不需要预处理及提前进行归一化等运算。分类回归树（Classification And Regression Tree，CART）算法是决策树算法的一种实现，选择 CART 的原因是 scikit - learn 使用优化版 CART 算法作为其决策树算法的实现。CART 算法是研究阈值的主要统计方法，由于可视化生成决策树简单直观，可以从可视化后的树状结构图中看出分类规则，具有速度快、准确性高和能够处理大数据等优点，较适合研究气象条件在流感预警问题上的预警界值问题。XGBoost 算法是对 GBDT 算法优化后在工程上的实现，其效果和模型训练速度一般要优于决策树和 GBDT 算法。本研究选择 CART 和 XGBoost 两种机器学习算法建模，研究基于气象因素的流感大数据预测预警方法。

3.3.2 *基于数据标准划分测试集和训练集并完成模型参数调节* 以 CART 算法为例，参数 criterion 选择默认的 default = "gini"，特征选择标准为基尼系数，即使用决策树算法中的 CART 算法。调节 max_ depth（树的最大深度）、min_ impurity_ decrease（节点划分最小不纯度）、min_ samples_ split（内部节点再划分所需最小样本数）和 min_ samples_ leaf（叶子节点最少样本数）几个关键参数，防止过拟合，提升模型对于未知数据集的准确率，使模型复杂度达到泛化误差最小目标。当样本数量较少、特征较多时，决策树较容易过拟合，一般为建立更健壮的模型，在分类器学习建立模型时需要不断进行迭代实验以进一步筛选特征，使模型泛化能力更强，避免从最终模型中获得的预警条件过于复杂。在尽量保证预测精度的前提下，CART 模型最终保留 7 个特征，见表 1。XGBoost 模型保留 11 个特征，见表 2。

表 1 CART 模型保留的特征数据

变量名	数据基本含义	数据类型	单位/取值
precip	日降水量	连续变量	毫米（mm）
t_ avg_ lastweek	上周平均气温	连续变量	摄氏度（℃）
t_ max_ day1ago	昨日最高气温	连续变量	摄氏度（℃）
inds_ avg_ day2ago	前天平均风速	连续变量	米每秒（m/s）
winds_ max_ day3ago	3 日前平均风速	连续变量	米每秒（m/s）
QNE_ day1ago	昨日平均气压	连续变量	百帕（hPa）
QNE_ day3ago	3 日前平均气压	连续变量	百帕（hPa）

表 2 XGBoost 模型保留的特征数据

变量名	数据基本含义	数据类型	单位/取值
t_ avg	日平均气温	连续变量	摄氏度（℃）
t_ max	日最高气温	连续变量	摄氏度（℃）
winds_ avg	日平均风力	连续变量	米每秒（m/s）
QNE_ hPa	日平均气压	连续变量	百帕（hPa）
radiation	日累计辐射	连续变量	兆焦耳每平方米（MJ/m2）
t_ min_ ins	最低气温温差	连续变量	摄氏度（℃）
precip_ avg_ lastweek	上周每日累计降水的平均	连续变量	毫米（mm）
winds_ max_ day3ago	3 日前最高风速	连续变量	米每秒（m/s）
winds_ max_ lastweek	上周每日最大风速的平均	连续变量	米每秒（m/s）
rh_ min_ lastweek	上周每日最小湿度的平均	连续变量	百分比（%）
radiation_ day2ago	前天的辐射	连续变量	兆焦耳每平方米（MJ/m2）

4 实验与结果

4.1 实验框架

4.1.1 概述 本研究实验数据由气象和流感发病数据整合组成，以日为单位、以气象因素为基本特征，在此基础上构造更为复杂的新特征，以是否预警为因变量进行有监督的分类学习。采用交叉验证方法评估模型效果，即将样本数据重复进行切分，组合为不同训练集和测试集以训练模型、提高模型健壮性。利用 5 折交叉验证法将 2012 – 2016 年数据划分为训练集和测试集。通过计算模型对于测试集的准确率、F1 – score、曲线下面积（Area Under Curve，AUC）等指标综合评估模型效果[4-5]。

4.1.2 准确率 表示模型预测效果的准确程度，是预测值与实际值相同的样本个数占总体样本的比例。其在一定程度上反映模型好坏，模型准确率越高表明模型预测结果越好。当样本中正类和负类占总体比例相差较大时容易导致过拟合现象，此时样本结果偏向比例大的一方导致准确率虚高，所以一般采取 ACC 值和其他评判指标综合评判模型好坏。

4.1.3 F1 – score 即 F1 分数，是分类问题的衡量指标，部分分类问题将 F1 – score 作为最终测评方法。它是精确率和召回率的调和平均数，最大为 1，最小为 0。其中精确率是指针对预测结果而言预测为正（需要预警）的样本中有多少是真正的正样本。召回率是指样本中的正例有多少被预测准确。

4.1.4 AUC ROC 曲线下面积常被用来评价二分类器模型的好坏，其中 ROC 曲线即受试者工作特征曲线。AUC 反映的是概率值，可以直观地对分类器性能进行量化，AUC 值越大则说明分类器性能越好，其值最大不超过 1。

4.2 模型预测预警效果评估

经过综合评估，选择 XGBoost 算法作为模型最优解（如果从简单性方面考虑，可选用 CART 算法），见表 3。

表 3 不同分类算法下的流感预测模型实验结果比较

算法	准确率	F1 – score	AUC
CART	0.82	0.69	0.73
XGBoost	0.87	0.65	0.81

4.3 流感大数据预警阈值设置

根据 XGBoost 模型可视化结果得出流感预警对应的气象条件的阈值如下：一是平均气温小于 14.05℃ 且平均风速小于 2.05m/s & 最高气温小于 13℃；二是平均气温小于 14.05℃ & 平均风速不小于 2.05m/s & 日累计辐射小于 166.1 MJ/m2；三是

平均气温不小于 14.05℃且 3 日前的那天最大风速小于 9.55 m/s & 平均气压小于 868.05 hPa。3 者之间为或的关系，当某一条件满足时就发出预警。

5 讨论

本研究主要包括数据采集、数据标注、特征处理、构建预测模型、模型优化、实验结果对比、可视化数据挖掘和预警阈值等方面。2012 – 2016 年流感样病例数与同期气象因素资料关系分析显示，平均气温、平均风速、最高气温、日累计辐射、最大风速和平均气压均与流感样病例数有关联，此项分析结果与施敏[6]研究杭州市流感样病例与气象因素关系的结果相似。在本研究中低温（平均气温小于 14.05℃和最高气温小于 13℃）时较大可能伴随流感暴发。因为当天气变冷时大众户外活动相对减少，当人群聚集在通风条件较差的室内时流感病毒更容易传播。此外有研究表明流感病毒在空气中的传播依赖于温度和湿度[7]。同时太阳辐射量的变化也会影响流感病例数增减，与已有研究结果相似[8]。本研究尚存在不足。首先，结论中预警阈值有效性有待进一步验证。由于仅研究某地区数据，具有一定局限性，将同样方法应用于全国不同气候条件、不同人口密度的其他地区能否产生同样结论需要进一步验证。其次，本研究采用 CART 与 XGBoost 模型，仅引入关键气象因素，而流感流行及暴发与人口学、社会经济、人类行为、疫苗接种等因素密切相关。最后，发病数据采集方面，因为存在轻症未就诊病例以及漏报病例等情况，造成分析结果偏差较大，数据质量问题导致模型效果无法达到预期。

6 结语

流感流行与暴发受多种因素影响且各因素影响交错复杂。随着计算机、互联网和地理信息系统（Geographic Information System，GIS）技术的迅速发展，传统流感监测已发展为根据流行病传播动力学并利用气象因素、互联网搜索数据加入空间分布等信息等对流感进行监测和评估[9-11]。未来可以开展基于互联网搜索数据的传染病预测研究，综合多种因素全方面、多角度对流感进行预测，从而减少流感对人类社会的危害。

参考文献

1 杜鹏程，于伟文，陈禹保，等．利用系统进化树对 H7N9 大数据预测传播模型的评估［J］．中国生物工程杂志，2014，34（11）：18 – 23.

2 McGowan C J，Biggerstaff M，Johansson M，et al. Collaborative Efforts to Forecast Seasonal Influenza in the United States，2015 – 2016［J］. Sci Rep，2019，9（1）：683.

3 陈翠霞，王小龙，蒋太交，等．基于多源异构大数据挖掘的流感病毒防控预测预警平台构建研究［J］．中国生物工程杂志，2020，40（1）：109 – 115.

4 王书芹，华钢，徐永刚，等．AUC 的不一致性分析［J］．江苏师范大学学报（自然科学版），2013，31（3）：31 – 34.

5 赵存秀．不均衡数据分类器分类性能 AUC 与 Accuracy 的比较［J］．唐山师范学院学报，2019，41（6）：75 – 77，132.

6 施敏．杭州市流感样病例与气象因素关系的研究［D］．杭州：浙江大学，2013.

7 Lowen A C，Mubareka S，Steel J，et al. Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature［J］. Plos Pathogens，2007，3（10）：1470.

8 Tymvios F S，Jacovides C P，Michaelides S C，et al. Comparative Study of Angstrom′s and Artificial Neural Networks′ Methodologies in Estimating Global Solar Radiation［J］. Solar Energy，2005，78（6）：752 – 762.

9 鲁学亮，刘臻．基于 WebGIS 的流感传播模拟与预警系统的设计与实现［J］．测绘与空间地理信息，2014（10）：58 – 60.

10 苏蕊．具有年龄结构的传染病 SIR 流行病模型的研究［J］．数学的实践与认识，2011，41（6）：140 – 143.

11 Mciver D，Brownstein J S. Wikipedia Usage Estimates Prevalence of Influenza – like Illness in Near Real – time［J］. PLoS Comput Biol，2014，10（4）：e1003581.