基于环境气象因素影响的异常就诊量预测

于广军[1,2]，熊赟[3,4]，彭思佳[4,5]，阮璐[3,4]
1. 上海市儿童医院，上海 200040；2. 上海交通大学医学院，上海 200025；
3. 复旦大学计算机科学技术学院，上海 200433；4. 上海市数据科学重点实验室，上海 200433；
5. 复旦大学化学系，上海 200433

摘要
通过分析某儿童医院传染科就医人数异常（突增、突减）情况，建立就医人数与气象特征间的分类模型，实现对传染科就医突变情况的高准确率预测，以便院方合理调配科室、安排医生出诊人数。建立的模型对就医人数突增情况的预测准确率达到92.8%，召回率达到83.5%；对就医人数突减情况的预测准确率达到87.4%，召回率达到92.4%，并与多种分类器进行比较，实验表明该方法在预警传染科就诊人数的突变方面综合表现更佳。

关键词
环境气象因素；随机森林；异常预测

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018030

Abstract detection of hospital admissions based on meteorological factors

YU Guangjun[1,2], XIONG Yun[3,4], PENG Sijia[4,5], RUAN Lu[3,4]
1. Children's Hospital of Shanghai, Shanghai 200040, China
2. Shanghai Jiaotong University School of Medicine, Shanghai 200025, China
3. School of Computer Science, Fudan University, Shanghai 200433, China
4. Shanghai Key Laboratory of Data Science, Shanghai 200433, China
5. Department of Chemistry, Fudan University, Shanghai 200433, China

Abstract
The hospital admission data from medicine department and infectious disease department of a hospital was analyzed and a classify model between the number of patients and meteorological factors was built. High accuracy of prediction in abnormal number of patients by utilizing random forest classifier was achieved, and decision support to Public Health Department was provided so that the hospital can make a reasonable allocation of doctors. All experiments were conducted on real data from the hospital and the results show that the final trained model achieve relatively high accuracy and recall.

Key words
meteorological factor, random forest, abnormal detection

1 引言

环境气象被认为是影响人类健康的因素之一，某些疾病的发生与恶化通常具有明显的周期性气候特征[1,2]。如春季气温回升，细菌滋生，小儿麻疹、风疹、水痘、手足口病等病高发；秋冬季气温下降，肺结核、哮喘、肺炎、流行性感冒等疾病较为严重。另外，空气污染物颗粒，如NOx、NO_2、CO、O_3、SO_2、PM2.5、PM10等，都有可能导致相关疾病发生率升高[3,4]。

就医人数作为疾病发生率的一种表现，易于统计，分析不同科室就医人数与气候变化的关系，为就医人数建立预测模型和公共卫生部门做出决策提供支持[5]，同时可以为人们选择就医时段提供参考。当前，反映气候状况的气温、空气环境的相应指标数据可以被准确全面地记录和整理。因为涉及隐私，对特定疾病发病情况的收集相对困难，而特定科室的就医人数也可以在一定程度上反映疾病的发生情况。

利用大数据技术从医疗相关数据中发现潜在的关系与模式，帮助医院和公共卫生部门做出决策已经受到关注。例如，2009年，Google公司借助大数据技术从用户的搜索行为中预测了甲型H1N1流感的爆发，比美国疾病控制与预防中心先一步发出预警[6]；一些研究根据用户在Twitter上的文章构建了流感样病例率的预测模型，取得了较准确的结果[7]。这表明大数据可以帮助预测疾病（尤其是流行病）的爆发趋势，然而，这一领域的研究存在以下问题。

（1）数据的隐私保护

不管是用户的搜索行为、社交网络的言论，还是医院或实验室的医疗数据，都涉及用户隐私，如何在保护隐私的前提下对数据进行分析与挖掘是一大挑战。

（2）数据的规模

数据是大数据分析技术的基础，在需要用户授权的情况下，目前往往只能收集到少量的用户行为数据，无法得到具有普遍性的结论。

（3）就诊量的波动性

特定科室的就医人数聚合了多种疾病的发病率信息，就医人数与气候指标间未必存在直接相关性，如何给出合理的预测预警是需要考虑的问题。

不同于现有的研究，本文利用反映气候状况的气温、空气环境的相应指标数据对就诊量进行预测，数据准确且规模大；通过对特定科室的就医人数进行预测，间接预测了特定类型疾病的发生发展情况，避免了对涉及隐私的病人发病情况的收集；并且侧重于预测就医人数的突发，建立就医人数突发的预警模型。

2 基于环境气象因素的就诊量预测模型

环境气象因素与一些特定疾病的发生息息相关，尤其是流行病和小儿疾病。当前，反映气候状况的气温、空气环境的相应指标数据可以被准确全面地记录和整理。因此，通过环境气象因素来对就诊量进行预测，是一个合理的选择。本文用到的环境因素包括两类：气温和大气污染物。气温因素包括3个指标：最高气温、最低气温和平均气温；大气污染指标包括PM2.5、SO_2、NO_2、CO。

由于环境因素并不是就诊量变化的唯一因素，因此直接对就诊量的数值进行预测是不合适的。本文试图对就医人数的异常情况建模，即预测就诊量的环比变化情况。比如，预测当天的就医人数相对前几天是平稳的还是突变的。因此，预测模型

是一个分类模型。笔者选择随机森林[8]作为分类器。

将气温因素和污染物因素及医院传染科平均就医人数作为模型的特征，建立就医人数与环境特征间的随机森林分类器，实现对就医突变情况预测。具体如下。

选择温度、PM2.5、SO_2、NO_2、CO指标和平均就医人数作为模型的特征，并假定各特征之间相互独立。变量T_t=<$T_{\min,t}$，$T_{\max,t}$，$T_{\mathrm{mean},t}$>表示日期t当天的最高气温、最低气温和平均气温。考虑就医人数与温度的时滞效应，选择预测日期前N天（不含当日）的温度变化作为特征，分别计算N天平均温度$\hat{T}_N$、（N+1）天内最大温差Dev(T_N)，其中：

$$\hat{T}_N=\sum_{i=1}^{N}T_{\mathrm{mean},t-i}/N \quad (1)$$

$$\mathrm{Dev}(T_N)=\max_{0\le i\le N}T_{\max,t-i}-\min_{0\le i\le N}T_{\min,t-i} \quad (2)$$

对PM2.5、SO_2、NO_2、CO指标做同样的处理，形成污染物的特征$\hat{P}_N$、$\hat{S}_N$、$\hat{N}_N$、$\hat{C}_N$，表示N天内PM2.5的平均值。$\hat{V}_N$表示经过标准化后的前N天的平均就医人数（不含当日）。因此模型的特征集表示为：

$$F=(\hat{T}_N,\mathrm{Dev}(T_N),\hat{P}_N,\hat{S}_N,\hat{N}_N,\hat{C}_N,\hat{V}_N) \quad (3)$$

对就医人数的异常情况建模，异常情况是指环比变化情况。模型的目标变量应该反映预测当天的就医人数的突变情况，假定阈值α>0，β<0，G=($V_t-\hat{V}$)/std(V_t)，当$G>\alpha$时，判定就医人数突增，当$G<\beta$时，判定就医人数突减，即当标准化后的就医人数偏离超过均值的α或β倍标准差时判定为就医人数突增或突减，见**表1**。

表1 判定就医异常情况

就医人数描述	$G(Y)$	$L(Y)$
突增	$G\ge\alpha$	1
突减	$G<\beta$	−1
平稳	$\beta\le G<\alpha$	0

其中，将$L(Y)$作为标签，特征集F中，$\hat{T}_N$,Dev(T_N),$\hat{P}_N$,$\hat{S}_N$,$\hat{N}_N$,$\hat{C}_N$,$\hat{V}_M$的取值为连续变量，L为类别标签。

本文在训练样本集D={$\hat{T}_N$,Dev(T_N),$\hat{P}_N$,$\hat{S}_N$,$\hat{N}_N$,$\hat{C}_N$,$\hat{V}_N$,L}上构造了10棵决策树组合的随机森林分类器。

3 实验分析

（1）数据集

模型涉及多源数据集，包括大气污染物监测数据、气温数据和就医人数的数据，下面以某市为例进行说明。

- 大气污染物监测数据来源于国家气象中心，包括该市2013年1月1日—2014年11月30日的大气SO_2、NO_2、CO、PM2.5污染物浓度。
- 气温数据为国家气象信息中心提供的2013年1月1日—2014年11月30日该市气象站每日常规连续监测数据，包括最高气温、最低气温和平均气温。
- 就医数据是2013年1月1日—2014年11月30日该市某儿童医院传染科科室就医人数每日变化数据。

上述3个数据集描述性统计结果见**表2**。

（2）环境气象与就诊人数趋势相关性分析

图1为该市PM2.5浓度的日平均值，为每日连续监测数据，可见浓度在冬季（2013年12月—2014年2月）达到当年最高值，春季次之，夏季（2014年9月—2014年10月）最低。污染物取值大于零，对污染物水平取对数发现，其对数取值满足正态分布。**图2**、**图3**、**图4**是该市大气污染因子SO_2、NO_2、CO的连续分布图，类似的，污染物在冬季（2013年12月—2014年2月）达

表 2 描述性统计结果

名称	平均	标准误差	中位数	众数	标准差	方差	峰度	偏度	最小值	最大值
最高温度/℃	21.7	0.3	23.0	26.0	9.0	80.2	−0.8	−0.1	2.0	40.0
最低温度/℃	15.2	0.3	16.0	23.0	8.7	75.7	−1.1	−0.2	−2.0	31.0
平均温度/℃	18.5	0.3	19.5	24.0	8.7	76.4	−1.0	−0.2	0.0	35.5
就诊人数/人	145.8	2.9	130.0	70.0	76.5	5853.0	−0.1	0.8	11.0	412.0
PM2.5/	76.2	1.9	62.0	30.0	51.2	2619.4	6.9	2.1	12.0	465.0
SO_2/(μg·m^{-3})	19.54	0.479	15.0	13.0	12.68	160.7	3.56	1.92	6.0	77.0
NO_2/(μg·m^{-3})	56.57	0.889	52	47	23.49	551.83	−0.01	0.71	10	128
CO/(μg·m^{-3})	20.39	0.3	18	16	7.94	63.1	6.42	2.04	10	76

注：大气 PM2.5 平均浓度为 76.2 μg/m^3，超过国家空气质量二级标准（35 ~ 75 μg/m^3），属轻度污染。

到当年最高值，春季次之，夏季（2014年9月—2014年10月）最低。

如**图5**所示，该市日气温变化具有明显的时间规律性，高峰值出现在夏季的8月，低峰值出现在冬季的1月。

传染科就诊人数变化趋势如**图6**所示，具有明显的周期性，其高峰值出现在初夏（6月—7月），8月—9月为次低谷，低峰值出现在深冬（1月—2月）。初夏之际，气温骤升，适合细菌繁殖，因此传染科的就诊量最高。而在1月—2月，气温是一年中最低的，不利于细菌的繁殖，就诊量也就相对较低。

（3）实验结果

实验以该市2013年1月1日—2014年11月30日气温数据，PM2.5、SO_2、NO_2、CO大气污染浓度和该市儿童医院传染科日就诊人数为基础，构造上述特征和标签，对就医人数进行了前文所述的标准化处理，并且根据标准化后的就医人数最多及最少的20%界定。实验采用随机森林作为分类器，以十折交叉验证的方式分别计算分类器在突增、突减以及正常情况下的预测准确率。为评估就诊人数随环境变化的时滞效应，研究延迟天数N从1到7变化时，各情况下的准确率和召回率，具体情况见**表3**。

实验结果表明，在N较小（N=1或N=2）时，分类的准确率都不高，表明环境的变化无法在短时间内立刻影响到就医人数。随着N的增大，分类效果逐渐提升，并在延迟4日时效果最好，此后随时间的增加，分类效果递减。这表明环境因素对传染科就诊人数的影响时滞在4日左右，延迟如果过大，则环境因素的影响变弱。实验结果表明分类模型可取N=4来获取最好的分类效果。

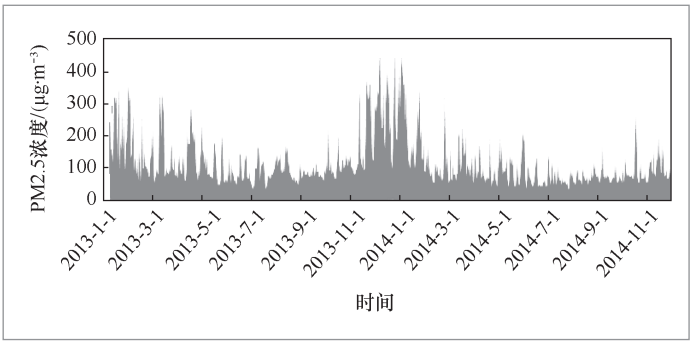

图 1 2013 年 1 月 1 日—2014 年 11 月 30 日某市 PM2.5 浓度随时间变化的趋势

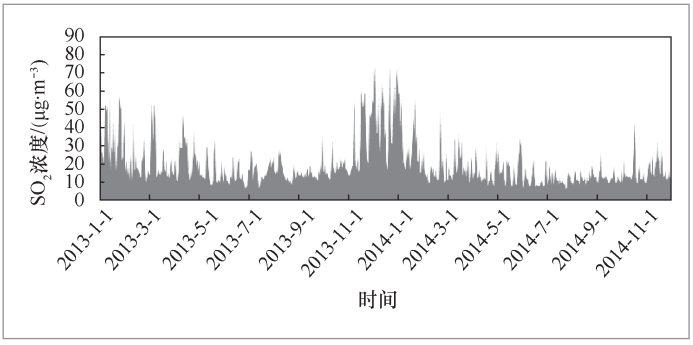

图 2 2013 年 1 月 1 日—2014 年 11 月 30 日某市 SO_2 浓度随时间变化的趋势

表 3 随机森林分类器预测准确率及召回率随延迟变化

*N*值	突增情况		突减情况		正常情况	
	准确率	召回率	准确率	召回率	准确率	召回率
N=1	0.858	0.785	0.828	0.898	0.722	0.690
N=2	0.866	0.825	0.840	0.897	0.746	0.712
N=3	0.891	0.848	0.846	0.893	0.762	0.742
N=4	0.928	0.835	0.874	0.924	0.805	0.781
N=5	0.908	0.815	0.855	0.910	0.778	0.762
N=6	0.913	0.825	0.860	0.885	0.743	0.762
N=7	0.907	0.839	0.841	0.897	0.768	0.740

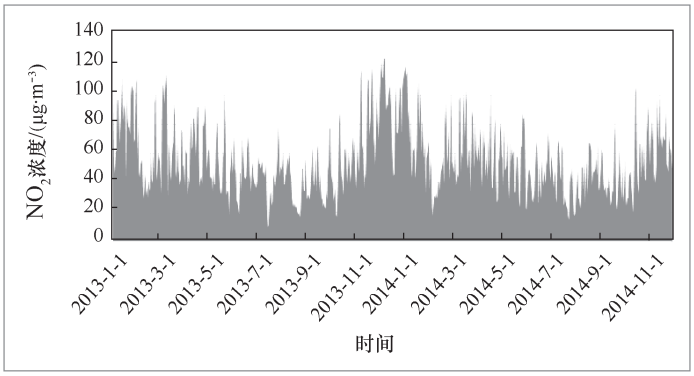

图 3 2013 年 1 月 1 日—2014 年 11 月 30 日某市 NO_2 浓度随时间变化的趋势

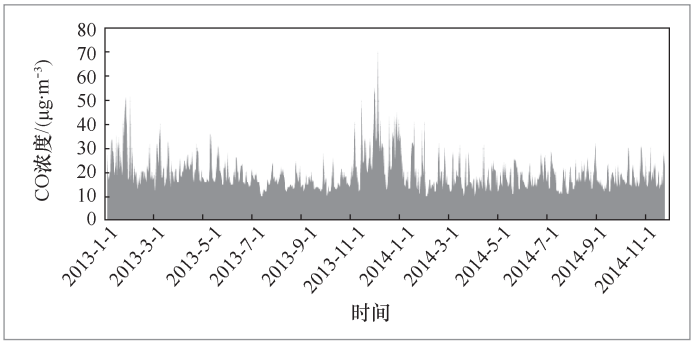

图 4 2013 年 1 月 1 日—2014 年 11 月 30 日某市 CO 浓度随时间变化的趋势

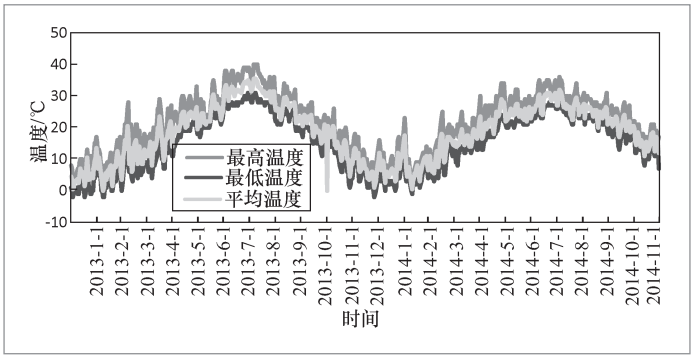

图 5 2013 年 1 月 1 日—2014 年 11 月 30 日某市气温随时间变化的趋势

在延迟设定为4日（*N*=4）时，就医人数突增预测的准确率为92.8%，召回率为83.5%；就医人数突减预测的准确率为87.4%，召回率为92.4%；就医人数没有明确波动的情况预测准确率为80.5%，召回率为78.1%。总体上看，分类器对正常情况的预测表现一般，这是因为就医人数受多种因素影响，其他变量的变化也会导致就医人数发生异常变化，使得结果不属于正常情况。可以将其他因素加入分类器中，提高对正常情况的预测效果。分类器对突增和突减情况的预测要明显好于对正常情况的预测，具有较高的准确率和召回率，说明从环境因素预测就诊量的异常波动是有效的。

随机森林的分类结果通过*K*个决策树结果的投票来决定，提升了单个决策树的分类精度，防止了过拟合的出现，是一种比较可靠的分类方法。为了评估随机森林分类效果，本文挑选了6种常用的分类算法（高斯朴素贝叶斯、SVM、*K*近邻、决策树、XGBoost、逻辑回归）进行分类试验，并与随机森林分算法进行对比。结果见**表4**。

实验表明各种分类算法在不同情况下各有优劣。高斯朴素贝叶斯方法在突增召回率上表现不错，但是准确率过低，且对正常情况的预测效果很差，说明过多地将正常情况判定为了异常情况；SVM和XGBoost在突增上的表现和随机森林接近，但在突减上表现略差。*K*近邻在各项指标上的表现都较差。决策树的召回率比较

表 4 各种分类算法进行分类试验的结果

算法	突增预测准确率	突增预测召回率	突减预测准确率	突减预测召回率	正常情况预测准确率	正常情况预测召回率
随机森林	0.928	0.835	0.874	0.924	0.805	0.781
高斯朴素贝叶斯	0.723	0.894	0.835	0.896	0.756	0.546
SVM	0.941	0.873	0.856	0.909	0.810	0.773
*K*近邻	0.898	0.846	0.858	0.858	0.741	0.756
决策树	0.858	0.852	0.849	0.819	0.707	0.736
XGBoost	0.925	0.882	0.849	0.895	0.794	0.749
逻辑回归	0.828	0.901	0.857	0.914	0.813	0.677

好，但准确率低。逻辑回归算法在正常情况下的召回率太低。总体来说，随机森林的表现要优于其他算法。

4 结束语

本文研究分析了分类模型在某儿童医院传染科就诊人数突变的应用。抽取待预测日期前一段时间内气温、污染物浓度以及就医人数整体水平作为分类模型的特征，利用随机森林模型预测就医人数的异常情况，具有较高的准确率和召回率。对就医人数的预测可以辅助医院合理安排医疗人员，亦可为公众合理安排就医时间提供帮助。由于科室就医人数聚合了不同疾病患者的就医信息，而不同疾病与外界环境的关系不尽相同，科室就医人数与环境因素间的相关性在一定程度上被弱化了。为判明特定疾病与环境因素间的关系需要收集更精准的数据，这是未来要完成的工作。

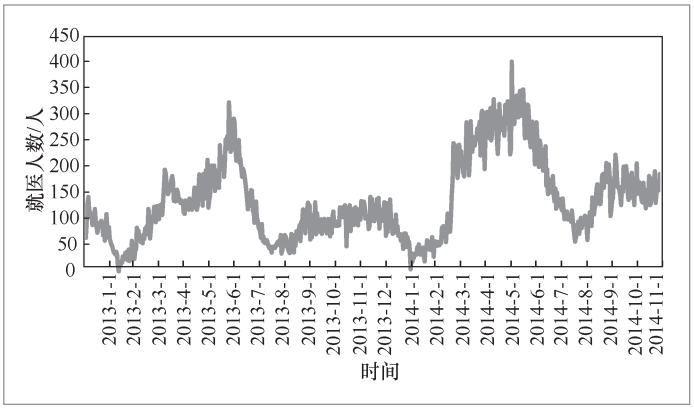

图 6 2013 年 1 月 1 日—2014 年 11 月 30 日某市儿童医院传染科就诊人数变化趋势

参考文献:

[1] KOUTRAS A, SAKELLAKIS M, MAKATSORIS T, et al. Seasonal variability in the incidence of carcinomatous meningitis[J]. Journal of Neurosurgery, 2015, 122(3): 543-546.

[2] HONG J S, KANG H C. Seasonal variation in case fatality rate in Korean patients with acute myocardial infarction using the 1997-2006 Korean National Health Insurance Claims Database[J]. Acta Cardiologica, 2014, 69(5): 513-521.

[3] LEE S L, WONG W H S, LAU Y L. Association between air pollution and asthma admission among children in Hong Kong[J]. Clinical & Experimental Allergy, 2006, 36(9): 1138-1146.

[4] LE T G, NGO L, MEHTA S, et al. Effects of short-term exposure to air pollution on hospital admissions of young children for acute lower respiratory infections in Ho Chi Minh City, Vietnam[J]. Research Report (Health Effects Institute), 2012, 12(4): 174-181.

[5] SOYIRI I N, REIDPATH D D, SARRAN C. Forecasting asthma-related hospital admissions in London using negative

binomial models[J]. Chronic Respiratory Disease, 2013, 10(2): 85-94.
[6] DAVIDSON M W, HAIM D A, RADIN J M. Using networks to combine "big data" and traditional surveillance to improve influenza predictions[J]. Scientific Reports, 2015(5): 8154.
[7] LAMPOS V, BIE T D, CRISTIANINI N. Flu detector: tracking epidemics on twitter[C]// European Conference on Machine Learning and Knowledge Discovery in Databases, September 20-24, 2010, Barcelona, Spain. Heidelberg: Springer Press, 2010: 599-602.
[8] BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

作者简介

于广军(1970-),男,博士,上海市儿童医院研究员、院长,国家卫生信息工程技术研究中心副主任、中国医院协会信息管理专业委员会常委、上海副主委。2006年开始负责具体组织实施上海"医联工程"。2011年、2013年分别获得上海市科技进步奖一等奖、中国医院协会科技创新奖一等奖和国家科技进步奖二等奖。

熊贇(1980-),女,博士,复旦大学计算机科学技术学院教授。2004年起从事数据领域方面的研究工作,作为项目负责人主持国家自然科学基金、上海市科学技术工作委员会发展基金以及企业合作项目。相关研究成果在本领域国际权威期刊或会议发表论文40余篇、出版著作3本。目前主要研究方向为数据科学和大数据。

彭思佳(1995-),女,复旦大学化学系本科生。2015年起进行能源相关催化材料的研究,相关研究成果在SCI期刊发表论文3篇。目前主要研究方向为数据科学。

阮璐(1992-),女,复旦大学计算机科学技术学院硕士生,主要研究方向为异质网络、网络表示学习。

收稿日期: 2018-03-16
基金项目: 国家高技术研究发展计划("863"计划)基金资助项目(No.2015AA020105);上海市科技发展基金资助项目(No.16JC1400801, No.17511105502)
Foundation Items: The National High Technology Research and Development Program of China (No.2015AA020105), Shanghai Science and Technology Development Fund (No. 16JC1400801, No. 17511105502)