ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.16, No.27, September . 2020
E-mail:kfyj@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

空气PM2.5等级预测系统

胡书明,张明丽

(西北工业大学,陕西 西安 710129)

摘要:如今随着现代工业的不断发展,人类的生产生活活动每天都在产生大量废气,严重影响了这里生活环境中的空气质量。在浩繁空气污染物中,PM2.5是对人体健康危害非常大的一种污染物。现有的测量PM2.5测量手段有着无法实时测量、精确度不高、适用性不广等缺点,所以可以利用机器学习的方法通过空气中其他物质的浓度来对PM2.5的浓度等级进行预测。该项目通过对北京市数年来的空气成分数据进行数据挖掘,从而实现对于PM2.5等级的预测。在项目中,主要使用了三种算法,分别是决策树,支持矢量机(SVM)和K临邻算法(KNN),并且对比这三种算法的优劣性。实验结果表明,该系统能够有效地预测空气质量,对人们的日常生活具有重要意义。

关键词:决策树;支持矢量机;K临邻;PM2.5

中图分类号:TP311 文献标识码:A

文章编号:1009-3044(2020)27-0209-02

DOI:10.14004/j.cnki.ckt.2020.2938

开放科学(资源服务)标识码(OSID):

1 引言

随着现代工业的不断发展,人类的生产生活活动每天都会产生大量的废气,这些气体排放到空气当中,严重影响了这里生活环境中的空气质量。人们的环保意识和可持续发展意识正在不停加强,对PM2.5等空气中的有害成分也越来越重视。

1.1 空气质量概述

颗粒物质(PM)是大气中所有颗粒物质的总称,其中空气动力学直径≤ 2.5μm(PM2.5)的类型是影响人类健康的最重要的因素。与由一种或两种物质组成的大多数污染物不同,PM包括多种粒径的物质,为了简化PM水平的评估并促进PM污染控制政策的实施,通常将空气PM水平按照每立方米空气中的总颗粒质量分类,其中几个颗粒尺寸范围由最大颗粒的空气动力学定义。

1.2 对于空气质量的预测

在如今,PM2.5的测量主要有3种方法:重量法、微量振荡天平法和β射线法。在这三种方法中,重量法测量PM2.5最为精确,但是无法做到实时监测,另外两种方法只适用于部分情况的测量,并且成本高昂。这里希望能够做到对PM2.5在空气中浓度的实时精确获取,以现有的测量方法并不能直接做到,但是这里可以利用机器学习的方法,通过空气中的其他成分来对PM2.5的浓度进行预测。

1.3 项目目标以及意义

该项目的目的是通过对于空气中其他影响空气质量的成分,来对于空气中PM2.5的浓度等级进行预测。本次项目中,这里选取了北京市从2013年12月至2018年12月之间的每日空气质量报告,通过对于空气中其他成分的数据挖掘,来对于第二天的PM2.5等级进行预测。

2 实现过程

2.1 数据集

数据集来源为中国空气质量在线监测分析平台,网址为https://www.aqistudy.cn/historydata。参考巫升平[3]的数据集组成,这里选取了7个属性。下面列出了数据集的格式,每个属性值及其单位。

表1 数据集的格式、属性值及其单位

序号	名字	描述	单位
1	PM2.5	细颗粒物,指环境空气中空气动力学当量直径小于等于2.5微米的颗粒物[4]。	AQI单位
2	PM10	可吸入颗粒物,指空气动力学当量直径≤10微米的颗粒物称为可吸入颗粒物	AQI单位
3	SO2	二氧化硫	AQI单位
4	CO	一氧化碳	AQI单位
5	NO2	二氧化氮	AQI单位
6	O3_8	臭氧	AQI单位
7	PM2.5等级	共7级(0:极优,1:优,2:良,3:轻度污染,4:中度污染,5:重度污染,6:严重污染)	无

2.2 系统结构

该空气PM2.5等级预测系统的结构如下:

收稿日期:2020-05-16

作者简介:胡书明(1996—),男,河南商水人,硕士研究生,助教,研究方向为软件工程。

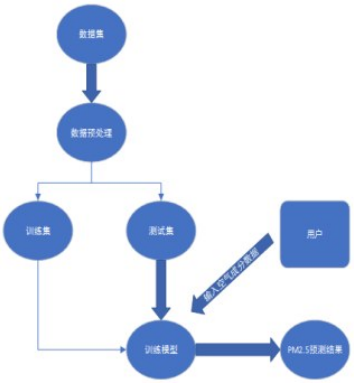

图1　系统结构

3预测结果

这里使用了3种预测算法：1)决策树；2)朴素叶斯；3)KNN。分类精度最低的是88.09%，分类精度最高的是90.87%。对于该大气模型而言，效果已经算是良好。

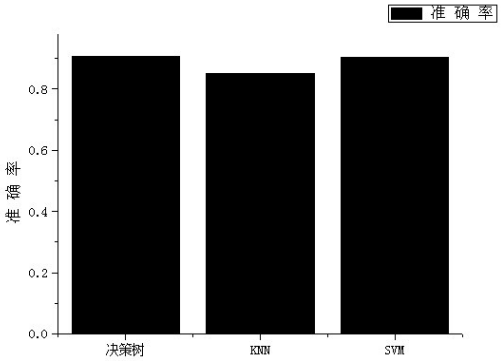

图2　预测系统使用的不同算法准确率

3.1决策树

决策树(Decision Tree)是在已知各种情况发生概率的基础上，直观运用概率分析的一种图解法[5]。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。Entropy定义为系统的凌乱程度，使用算法ID3，C4.5和C5.0生成树算法使用熵。这一度量是基于信息学理论中熵的概念。

决策树的学习过程如下：

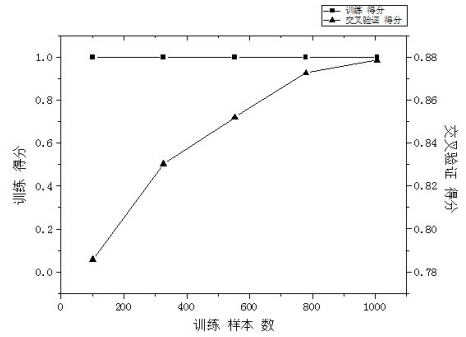

图3 决策树训练过程

3.2 SVM

SVM是新兴发展的一种以统计学习理论为基础的机器学习方法，能有效地避免经典机器学习方法中的(包括神经网络)的过学习、维数灾难、局部极小等传统分类存在的问题，在小样本条件下仍具有很好的泛化能力，因此受到极大的关注。

SVM的学习过程如下：

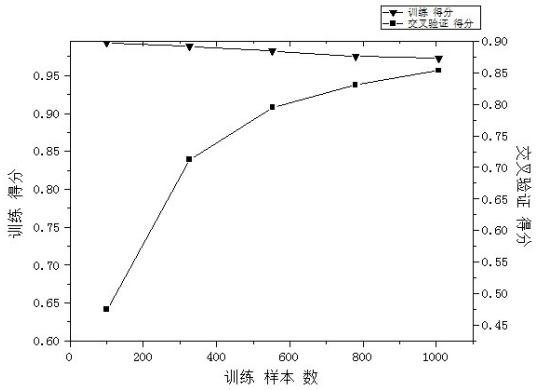

图4　高斯核SVM的学习曲线

3.3 KNN

K最近邻(k-Nearest Neighbor，KNN)分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。由于KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN方法较其他方法更为适。

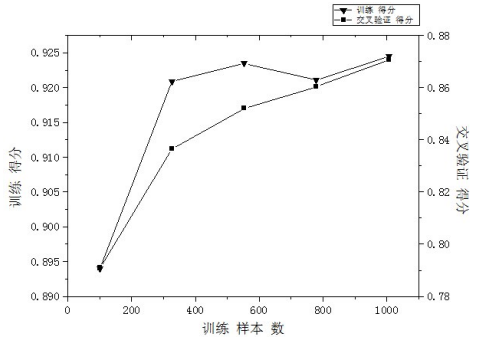

图5　KNN的学习曲线

训练过程的查准率(Precision)，查全率(recall)，以及f1测度值如下：

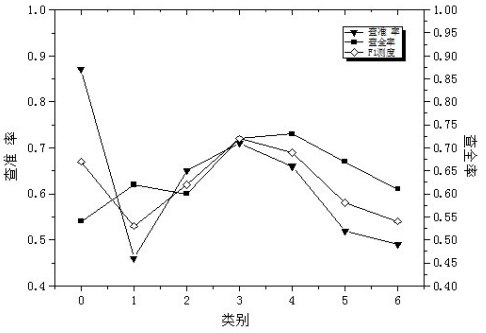

图6 查准率，查全率，以及f1测度值

3.4 结果对比

对于KNN来说，准确率高，对异常值和噪声有比较高的容忍度。和朴素贝叶斯之类的算法比，对数据没有假定，准确度高，对异常点不敏感。可以用于非线性分类，计算量大，对于存储器的需求也大。对于SVM，最终决策函数只由少数的支持矢量所确定，计算的复杂性取决于支持矢量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾难”。在高维空间

(下转第226页)

种模块与STM32F103VET6的连接方式可以分析出通过调用ST公司提供的SPI、GPIO及串口相关的库函数即可完成4种模块驱动程序的设计。结合具体应用可以设计出图5所示的系统程序流程图。

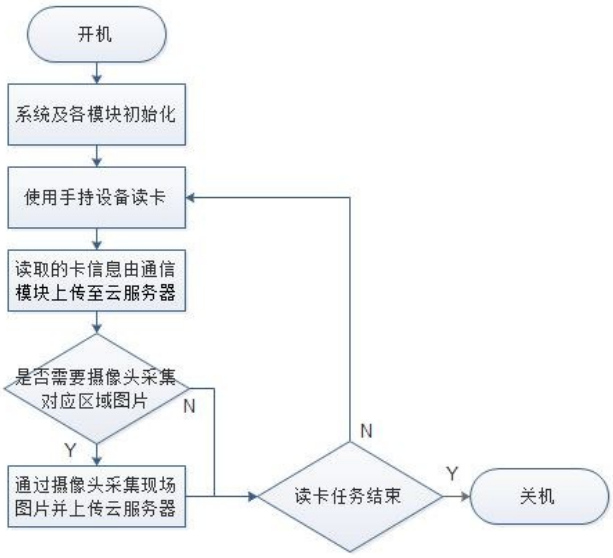

图5 系统程序流程图

手持设备采集的现场MIFARE1卡的数据及摄像头采集的相应区域图片通过SIM7600C通信模块传输至云服务器，购买的云服务器上一般运行的Linux操作系统，因此需要开发基于Linux操作系统的服务器程序。服务器程序负责对socket服务器进行封装，并在请求到来时，对请求的各种数据进行整理。应用程序则负责具体的逻辑处理。为了方便应用程序的开发，可选用Django、Flask、web.py等Web框架。WSGI（Web Server Gateway Interface）是一种规范，它定义了使用Python编写的web应用程序与web服务器程序之间的接口格式，实现web应用程序与web服务器程序间的解耦。web应用程序除需要完成接收现场采集的数据外，还需借助MySQL数据库保存上述现场数据供后续的分析及PC或手机的客户端访问查询。

4结束语

此系统的设计能够满足试验田中经常进行的各种对比实验的需求，减少实验人员的工作量，提高了实验的准确性。将RFID技术、云服务及4G通信技术等应用于农业领域，为物联网在农业的普遍推广进行了有益的尝试。本系统在便捷性、可维护性、信息化程度和成本控制上都有了很大提升，在已有的平台上还有更多地可以加入其中的新思路和构想，未来将做更多尝试以满足不断发展的农业现代化需求。

参考文献：

[1] 陈孝赏,陈伟强,刘守坎.台州市11个鲜食马铃薯品种比较试验[J].浙江农业科学,2019,60(12):2226-2227,2230.
[2] 原变青,贾岚,杨婷.基于云的RFID系统架构与安全性分析[J].电脑知识与技术,2020,16(1):27-28.
[3] 王亓剑,赵馀,章华,等.浅析WSN、RFID技术在我国农业中的应用[J].产业与科技论坛,2020,19(1):52-53.
[4] 邢玉广,张彦军.基于RFID的智能IC卡管理机的设计与研究[J].电子测量技术,2018,41(24):40-45.
[5] 王艳,王树磊,孙浩洋,等.SIM7600和千寻位置差分数据的高精度定位研究[J].单片机与嵌入式系统应用,2020,20(4):2-5.
[6] 秦钰林,周若麟,张珂欣,等.基于NB-IoT窄带通信和多传感器组网技术的森林火灾监测预警系统[J].物联网技术,2020,10(6):14-16,19.

【通联编辑：梁书】

（上接第210页）

有效，在维度数量大于样本数量的情况下仍然有效。Decision trees易于理解，乃至比线性回归更直观；模型可以通过树的形式进行可视化展示，与人类做决策思考的思维习惯契合。可以直接处理非数值型数据。

4结束语

这里已经完成了这个项目的以下步骤：对数据进行了预处理，然后对数据进行了分析。理解特征之间的关系，基于特征之间的依赖关系选取特征，然后进行数据分析。采用多种的算法，采用对于本软件最有效的三种算法（这里就是Decision Tree，KNN，SVM）。此中Accuracy是根据测试集样本正确率计算的。

通过分析可以得出结论，可以利用机器学习算法进行空气质量预测分析，从而预测下一天的空气质量。该空气质量软件预测系统是有效的，有助于满足预测的要求。使用该空气质量预测系统可以有效地预测空气质量，对人们的日常生活具有重要意义。

参考文献：

[1] 施晓娟,张会然,阎锡新.大气悬浮颗粒物所致气道黏液高分泌的研究进展[J].广东医学,2017,38(S1):279-281.
[2] 巫升平.成都市空气污染物季节性变化规律[J].科技风,2017(23):140-141.
[3] 杜飞燕.PM2.5暴露对大鼠清除肺炎克雷白杆菌的影响及其机制[D].石家庄:河北医科大学,2012.
[4] 莫洪武,万荣泽.分类算法在煤矿勘探数据分析系统中的比较[J].煤炭技术,2013,32(12):135-136.
[5] 杨伟光.面向大数据分析的决策树算法研究[J].电子技术与软件工程,2018(23):175.
[6] 杨铁建.基于支持向量机的数据挖掘技术研究[D].西安:西安电子科技大学,2005.
[7] 周明飞,熊伟,刘还珠.KNN方法在贵州晴雨预报中的试验[J].贵州气象,2010,34(6):3-5.
[8] 赵宇.基于支持向量机的多用户检测算法、功率控制算法和波达方向估计算法[D].合肥:中国科学技术大学,2006.

【通联编辑：闻翔军】