doi:10.11959/j.issn.1000-436x.2017266

基于大数据和物联网的空气质量预测监测研究

刘燕[1,2]，张永平[2]，朱成[2]，皋军[2]，刘其明[2]
（1. 南京理工大学电子工程与光学工程学院，江苏 南京 210094；2. 盐城工学院信息工程学院，江苏 盐城 224051）

摘 要：空气质量预监测已成为一种迫切需求，而这是一个复杂的系统工程。从基于大数据的智能决策角度研究智能空气指数预测，引入流行的分类算法，挖掘历史数据隐含的信息，实现空气质量预测；构建了基于物联网的空气质量监测系统，利用分类算法实现实时采集数据的智能处理。针对空气指数历史数据和实时采集数据规模较大的问题，为提高数据处理速度、增强空气质量预测的实时性，引入云计算技术加速数据处理；为使用户随时随地了解空气指数，还设计了基于 Android 平台开发空气指数预报客户端。

关键词：空气质量；智能预测；空气监测；Android
中图分类号：TP391 **文献标识码**：A

Intelligent forecasting and monitoring of air index based on big data and internet of things

LIU Yan[1,2], ZHANG Yong-ping[2], ZHU Cheng[2], GAO Jun[2], LIU Qi-ming[2]
(1. School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China;
2. School of Information Engineering, Yancheng Institute of Technology, Yancheng 224051, China)

Abstract: Air quality forecast has become an urgent need. However, numerical forecast of air quality is a complex systems engineering. The intelligent forecasting of air index was studied from the perspective of big data and intelligent decision making. For the index prediction of air quality, the popular classification algorithm was introduced to realize the intelligent analysis of historical data. To obtain air quality information in real time, a monitoring system based on Internet of Things was established, and intelligent processing of real-time data collected by the classification algorithm was achieved. Due to the large amount of historical air index data and real-time data collected, the technology of cloud computing and big data was introduced to speed up the data processing and improve the storage of data. In addition, the client based on Android was developed to allow users to query the air quality anytime, anywhere.

Key words: air quality, intelligent forecast, air quality monitoring, Android

1 引言

由于气候变化、工业生产和人口聚集等原因，我国很多地区的空气质量状况不容乐观。当前我国还存在着空气质量监测和控制水平不高、人们主动参与环保的意识还有待加强等问题，在大气污染的治理方面还没有取得显著的成绩。然而，随着经济的发展和生活水平的提高，人们对空气质量的关注

收稿日期：2017-09-24
通信作者：张永平，njustzhyp@sina.com.cn
基金项目：国家自然科学基金资助项目（No.61502411）；江苏省科技型企业创新基金资助项目（No.BC2015178）；江苏省高校自然科学基金资助项目（No.16KJB520042）；盐城工学院骨干教师科研启动基金资助项目（No.XJ2015017）；江苏省生态建材与环保装备协同创新中心基金资助项目（No.GX2015206）；毫米波国家重点实验室基金资助项目（No.K201731）
Foundation Items: The National Natural Science Foundation of China (No.61502411), The Innovation Foundation of Science and Technology Enterprises of Jiangsu Province (No.BC2015178), The Natural Science Fund of Colleges in Jiangsu (No.16KJB520042), Research Fund by Yancheng Institute of Technology (No.XJ2015017), Project of Collaborative Innovation Center of Ecological Building Materials and Environment Protection Equipment in Jiangsu Province (No.GX2015206), Project of Sate Key Laboratory of Millimeter Waves (No.K201731)

越来越高，及时的空气质量预测预报已成为一种迫切的需求。

为了使人们能够随时随地了解自己所处环境的空气质量状况，本文跳出了一般空气质量数值预测预报“建立监测点→采集数据→建立模型→分析数据→得出结论”的模式，研究基于大数据和物联网的空气质量预测监测，利用智能决策算法分析空气质量历史数据和实时采集数据，以预测未来的空气质量状况或监测当前的空气质量状况。为了加快海量数据的处理速度，本文研究并引入了热门的大数据处理技术；此外，还开发了基于 Android 的空气质量预报客户端，以方便人们随时随地了解空气质量状况。

2 相关工作

智能空气指数预测监测研究主要涉及如下一些领域和工作。

2.1 空气质量预测

近年来，我国接连发生罕见的大范围、长时间的雾霾天气过程，受其影响，全国多个城市出现了连续数日的重度污染天气，而人口的逐渐聚集加重了这一影响。雾霾天气和重度污染的空气给人们的身体健康和生产生活带来诸多不便，引起政府和公众的广泛关注，成为各大门户网站、微博、新闻媒体等报道和讨论的热点[1~4]，空气质量的预测预报成为迫切需要。

一般地，空气质量数值预测预报是通过各类预报方法与手段相结合，气象、物理、化学、地理等多学科耦合研究，建立空气质量模型，对多种大气污染物在内的不同尺度下不同类型污染过程进行模拟预测研究，这是一个非常复杂的系统工程[5~10]。它需要设置监测点、采集数据、分析数据，根据空气中各种成分的含量对照空气质量标准得出当前或未来一段时间某地的空气质量状况。

目前，我国国家环境空气监测网还不健全，监控点较少（国控点仅 1 436 个）、覆盖较低、精确性不够（我国空气质量指数预报准确率范围为 23.43％～81.15％，总体准确率仅为 48.37％）、溯源能力较弱、预测能力较低。这就需要研究补充技术，在不增加或少增加投入的前提下获得更精准的空气质量数据。热门的大数据技术和重现蓬勃生机智能决策算法，可以通过分析空气质量状况的历史信息来预测当前情况下的空气质量，不必布置更多的监测点就可以对未来一段时间内的空气质量进行预测，这为空气质量数值预测预报带来了新的思路。

2.2 贝叶斯分类

分类是一种分析数据的形式，主要用来从待分析的数据中提取需要的信息和知识，或是对未来的信息进行预测[12~15]。本文在验证基于大数据的智能空气质量预测和实时监测数据处理时，初步选用了贝叶斯分类方法，主要是其针对每个项目通常只会有相对较少的特征数，并且对项目的训练和分类也仅仅是针对特征概率的数学运算，算法简单、可靠。

贝叶斯分类的处理过程如图 1 所示。图 1 中 $P(\omega_i)$ 为先验概率、$P(x|\omega_i)$ 为条件概率、$P(\omega_i|x)$ 为后验概率，有如下关系

$$P(\omega_i|x)=\frac{P(x|\omega_i)P(\omega_i)}{P(x)} \tag{1}$$

后验概率决策方法为

$$P(\omega_i|x)=\max P(\omega_j|x), x\in\omega_i \tag{2}$$

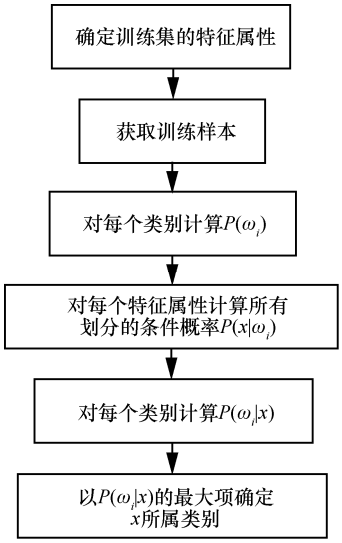

图 1 贝叶斯算法分类/决策过程

2.3 大数据和云计算

智能空气质量预测需要对大量的历史数据进行分析，处理速度受数据量规模的影响较大，引入云计算技术加速数据的数据分析的速度是一个可行的思路。事实上，可以认为大数据和云计算是“一体两面”的：云计算提供作为计算资源的底层，支撑着上层的大数据处理；而大数据为云计算提供了应用平台，驱动着云计算的发展[16~19]。

Hadoop 是一个开源云计算平台，能够实现大数据的分布处理[19]。Hadoop 的分布式文件系统

（HDFS）可以对海量数据进行存储，使用 MapReduce 作为编程模型。MapReduce 将一个大的任务拆分成多个任务碎片，分别由集群中的一个计算资源进行处理，然后这些碎片化的处理结果又由多个计算资源进行合并, 形成最终的处理结果。Hadoop 可靠性和效率都很高,支持多个计算资源并行处理，是当前标准的大数据处理工具。在智能空气质量预测研究中，需要对已存在的历史数据进行分析，这是 Hadoop 的优势所在。

2.4　物联网

简单地说，物联网就是物物相连的网络，通过把 RFID、传感器、二维码等智能感知系统嵌入“物理实体”以随时获取该物体的信息，从而将各种物体连接起来。物联网具备了信息感知、信息传输和信息处理等能力。鉴于当前空气质量指数预测的低准确率，为了获取实时空气质量信息，本文研究了基于物联网技术的简易空气质量监测系统，可以采集传感器节点附近空气成分及分析数据。

3　系统架构

空气质量指数预测监测研究主要包含前端系统和后台系统两大部分，如图 2 所示。

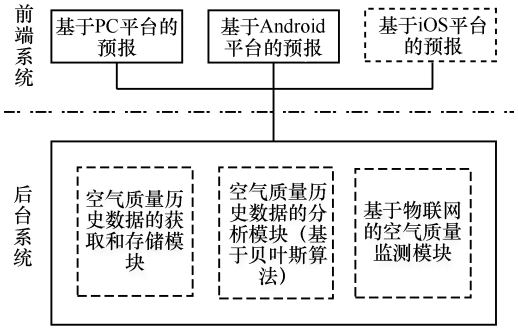

图 2　智能空气质量预测监测架构

3.1　前端系统

智能空气质量指数预测监测的前端系统就是向用户预报空气质量的客户端。

当前计算机在我国已经普及到了千家万户，88%的家庭拥有个人计算机，所以开发基于个人计算机的空气质量预报平台可以把预测的空气质量信息传递给众多家庭，这是前端系统开发的一个方面。但是人们并不总是在家里、办公室等室内进行活动，事实上外出时人们更需要及时了解附件的空气质量状况，这是我们关注的另一个重点。统计数据显示，近年来我国智能移动终端（尤其是智能手机和平板电脑）的普及率越来越高，至 2016 年平均每百人拥有 95 部手机，而 2016 年全国平板电脑拥有量也已达到了 2.8 亿台，智能终端已经成为人们生活中的必备品。在移动终端用户中，运行 Android 操作系统的终端数量在用户总数中占到了 8 成以上，基于 Android 的空气质量预报将是预报研究的突破点。

本课题的前端系统包含了基于 Web 的空气质量预报平台和基于 Android 的空气质量预报客户端，这保证了绝大多数人能享受到我们的研究成果提供的服务，相关结果可见本文第 6 节。在移动终端用户中，还有相当一部分人使用了 iOS 平台，而且其用户群体比较稳定甚至有所增长。为了扩大覆盖用户的范围，开发基于 iOS 平台空气质量预报客户端已经被列入“日程”成为系统前端开发接下来的工作。

3.2　系统后台

智能空气质量预测监测的后台系统，包含空气质量历史数据的获取和存储、空气质量历史数据的分析和基于物联网的空气质量实时监测 3 个功能模块。这 3 个模块相互协作共同完成了空气质量历史数据的获取、存储和分析，并形成预测；同时又为有条件的地区提供了实时空气质量监测系统，为用户提供所关心区域的实时、准确空气质量状况。

后台系统 3 个功能模块的实现需要大数据/云计算技术的支持，这可以提高空气质量历史数据存取、分析和预测速度，并为空气质量监测系统提供安全的数据存储和快速的空气指数分析。后台系统架构如图 3 所示。

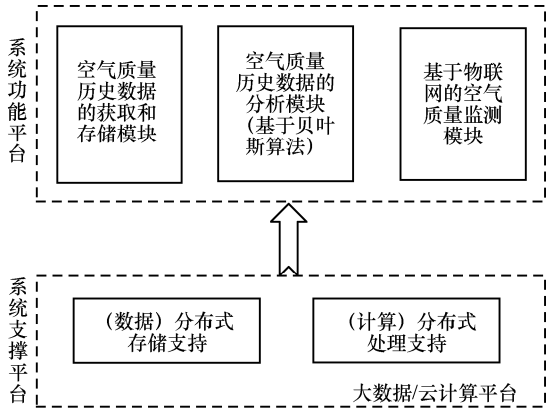

图 3　后台系统架构

1) 数据的获取和存储

空气质量指数数据的获取有 2 个途径：一是从环境保护监测和发布中心获取公共数据，二是实地

安装监测站点获得私有数据。目前，通过安装监测站点获得数据需要的投入过大，仅布置了少量的站点；研究的数据主要是获得的公共信息数据。海量数据的存储分为本地存储和云端存储，本地存储主要使用工具软件 MySQL 保存数据，而云端存储就是把数据上传到 HDFS 系统中。

2) 数据的分析

智能空气质量指数预测预报研究通过分析海量空气质量指数的历史数据，探究不同空气质量状况及变化趋势对未来一段时间内的空气质量的影响，这是大数据技术的天生“能力”，也是本文的研究思路，不必布置更多的监测点就可以对未来一段时间内的空气质量进行预测预报。

智能空气质量指数预测是基于贝叶斯分类实现的，它首先对获取的大量空气质量及其变化样本进行训练，得到各种空气质量状况下，下一时间段（通常为 1 天）空气质量好和空气质量差的先验概率和条件概率；然后利用贝叶斯分类进行决策。这样就可以得到根据历史数据预测的未来一段时间的空气质量。这里需要注意的是，要预测某地的空气质量状况，一定选择该地区的历史数据。鉴于需要分析的空气质量数据量比较大，引入 Hadoop 分布式数据处理框架，即先把海量空气质量数据上传的 HDFS 文件系统中，然后利用 MapReduce 框架对数据快速分析并汇总决策结果。

3) 基于物联网的空气质量监测

我国当前空气质量监测技术相对滞后、监测区域小。为了补充国家监控体系不足，本课题拟将基于物联网技术在有条件的地区设置部分监控节点，利用高精度传感器模组构建空气质量指标的采集和数据传输模块，然后对获取的空气成分指数进行分析，得到当前空气质量状况。实验中选择了盐城市区的某些区域设置节点，对空气质量状况实时监测，获取 PM2.5、PMl0、SO_2 等空气成分信息，并分析实时空气质量状况。

4 基于贝叶斯算法的空气指数智能预测

为了简化工作，在智能空气质量指数预测预报研究中按照空气质量标准（如表 1 所示）把空气质量状况简单地按空气良好和空气污染严重进行分类：当空气质量指数 $AQI>150$ 时认为空气污染严重，反之认为空气良好。

基于贝叶斯决策的智能空气质量指数预测就是在已知一部分原始数据的情况下，根据表 1 给出的标准判断空气质量情况，并分析产生该空气质量状况的概率及影响；然后用主观概率对未知事件进行概率估计，并对发生概率用贝叶斯公式进行修正，即用贝叶斯公式计算出相应的后验概率；最后再结合期望值和修正概率，做出最优决策。

在这里，通过贝叶斯公式进行决策，即比较假设空气良好的后验概率和假设空气污染严重的后验概率，如果空气良好的后验概率大于空气污染严重的后验概率

$$P(\omega_1 \mid x) > P(\omega_2 \mid x) \tag{3}$$

则可以得出预测“今天空气质量较好”；如果空气良好的后验概率小于空气污染严重的后验概率

$$P(\omega_1 \mid x) < P(\omega_2 \mid x) \tag{4}$$

则可以得出预测“今天空气严重污染”。其中，$P(\omega_1 \mid x)$ 代表空气良好的后验概率，$P(\omega_2 \mid x)$ 代表空气污染严重的后验概率。

基于贝叶斯的智能空气质量预测应用示例如下。

假设对空气质量检测数据的可靠度为 95%（这个数据可以从空气质量历史数据中统计获得），也就是说，当空气严重污染时，每次检测污染的概率为 95%；当空气良好时，每次检测没有污染的概率为 95%。假设对某地的历史空气质量情况进行检测，已知空气严重污染的概率是 10%。想知道，每

表 1 AQI 指数标准参考

空气质量	AQI 指数	PM2.5 日均浓度	PM10 日均浓度	活动建议
优	0~50	0~35	0~50	各类人群正常活动
良	51~100	35~75	51~150	极少数敏感人群应减少户外活动
轻度污染	101~150	75~115	151~250	儿童、老年人应减少户外活动时间
中度污染	151~200	115~150	251~350	一般人群应适量减少户外活动时间
重度污染	201~250	150~250	351~420	一般人群应尽量减少户外活动时间
严重污染	250~500	250~500	421~500	一般人群应避免户外活动

次检测污染严重的概率有多高？如果令“A”为空气污染严重事件、“B”为空气良好事件、“C”为检测结果为空气污染严重事件，可以得到如下结证。

1) P(A) 为空气污染严重的概率，不考虑其他情况该值为 0.1，这个值就是事件 A 的先验概率。

2) P(B) 为空气良好的概率，值为 0.9。

3) P(C|A) 为空气严重污染时检出空气污染严重的概率，根据假设它的值为 0.95。

4) P(C|B) 为空气良好时检测为空气污染严重的概率，也就是检测出错的概率，其值为 1−0.95= 0.05。

5) P(C) 为不考虑其他因素影响的空气污染严重检出概率，就是检测出污染严重的先验概率。其值可以通过贝叶斯公式计算得到

$$P(\mathrm{C}) = P(\mathrm{C},\mathrm{A}) + P(\mathrm{C},\mathrm{B}) \tag{5}$$

$$P(\mathrm{C}) = P(\mathrm{C}|\mathrm{A})P(\mathrm{A}) + P(\mathrm{C}|\mathrm{B})P(\mathrm{B}) \tag{6}$$

从而，该概率的值为

$$(10\% \times 95\%) + (90\% \times 5\%) = 14\% \tag{7}$$

综上所述，能计算出某日空气污染严重并且被正确检测出的概率为

$$P(\mathrm{A}|\mathrm{C}) = \frac{P(\mathrm{C}\,|\,\mathrm{A})P(\mathrm{A})}{P(\mathrm{C})} = \frac{P(\mathrm{C}\,|\,\mathrm{A})P(\mathrm{A})}{P(\mathrm{C}\,|\,\mathrm{A})P(\mathrm{A}) + P(\mathrm{C}\,|\,\mathrm{B})P(\mathrm{B})} \tag{8}$$

其值为 0.68 。

同样地，可以再计算出 $P(\mathrm{B}|\mathrm{C}) = 0.036$，然后对比 $P(\mathrm{A}|\mathrm{C})$ 和 $P(\mathrm{B}|\mathrm{C})$，由于 $P(\mathrm{A}|\mathrm{C})$ 大于 $P(\mathrm{B}|\mathrm{C})$，所以结果认为该日空气污染严重。

尽管天气污染检测的准确率很高，但只可以得出以下结论：如果某日空气检测为污染严重，那么当天的污染概率大约为 68%，也就是说当天污染的可能性比较大。事实上，对比我国空气质量预测的“总体准确率 48.37%”，这里的准确性已经非常可观了。根据以上原理可以得知，空气质量预测过程中首先利用从海量历史数据中挖掘规律，获得先验概率和空气污染状况概率；然后利用贝叶斯分类算法得出后验概率；最后通过对比后验概率获取预测的未来空气污染情况。

5　基于物联网的空气质量监测

5.1　空气质量监测架构

空气质量监测系统应能够进行空气成分数据的获取、传输和管理，并通过网络把数据传送到数据存储和处理平台（在本课题中使用大数据平台），其基本架构如图 4 所示。基于物联网的空气质量监测从功能上来说可分为空气质量数据获取终端、服务器和数据处理平台。

空气质量数据获取终端是一个 ZigBee 无线传感器网络，其设施主要包含一个协调器节点和若干一个终端节点。终端节点包括用于检测各项空气成分指标的传感器、核心控制单元、无线通信模块以及嵌入式软件系统等，负责采集数据；它采集到的数据直接通过无线网络上传到服务器上。

服务器上搭载有 J2EE 应用服务器和 MySQL 数据库，用于提供数据访问的接口和保存监测点位置信息及空气质量数据。服务器能够把数据上传到后台的数据存储和处理平台。在后台的空气质量数据处理平台，分析各种空气成分及其对空气质量指数的影响，对照空气质量指数标准（参见表 1），把空气质量分为空气质量良好和空气污染严重 2 种情况。

5.2　基于贝叶斯分类的空气成分数据处理

在空气质量监测系统中，采集到空气中各主要成分（这里主要关注的是 PM2.5、PM10、SO_2、NO_2 等）的信息之后，并没有使用传统方法分析成分，确定空气质量，而是利用朴素贝叶斯算法分析各成分对最终空气质量指数的影响，确定空气质量是良好还是污染。

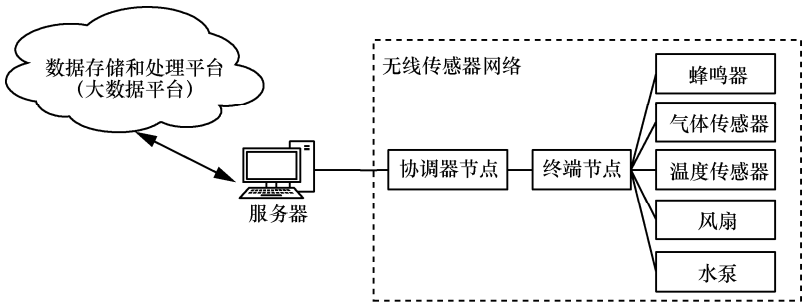

图 4　基于物联网的空气质量监测框架

在这里的分类过程同样是分析空气质量历史数据，根据历史数据中的空气质量状况和空气中各种成分含量之间的关系进行分类。但与空气质量预测中分类不同的是，此处只有 2 个类别：空气质量良好和空气污染严重。

分类过程这里就不再详述了。当确定了训练样本及其分类之后，就可以使用贝叶斯方法得到对未来数据进行预测

$$P(x\mid D)=\int_{\Theta}P(x,\Theta\mid D)=P(x\mid\Theta,D)P(\Theta\mid D) \quad (9)$$

其中，D 是训练样本，Θ 是概率模型参数，x 是待预测数据。当模型确定后，数据来自于独立同分布的抽样，所以 $P(x\mid\Theta,D)$ 可以简化为 $P(x\mid\Theta)$。

这里有一个问题：朴素贝叶斯方法设定样本的各属性是相互独立的，但决定空气质量的参数并不存在这样的特点。虽然本文考察的历史数据中每一个样本的空气指数，但空气中的各种成分 PM10、PM2.5、SO_2、CO_2 等共同确定着空气质量指数，而且各种空气成分也相互影响。为了解决这个问题，体现各属性之间的关系，引入了加权的统计量方法，假设经过加权后的样本属性是独立的。

假设空气中各成分含量的值记为 A_i、最终的空气指数记为 A_q，χ^2 加权的原理即为利用式(10)计算每个属性 A_i 与空气指数 A_q 之间的相互关系。

$$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e} \quad (10)$$

也就是 A_i 对 A_q 的权重

$$w_i=\mid\chi^2(A_i,A_q)\mid=\left|\sum\frac{(f(A_i,A_q)-\frac{f(A_i,A_q)}{N})^2}{\frac{f(A_i,A_q)}{N}}\right| \quad (11)$$

其中，$f(A_i,A_q)$ 代表 A_i 和 A_q 的关系强度。这就是基于 χ^2 的属性加权。

5.3 空气质量监测过程

在空气质量监测系统中，首先预先在某些区域设置空气成分采集节点（即空气质量数据获取终端），这些终端可以实时采集该区域的空气中各种成分的含量。系统数据采集及处理的过程如下。

1) 用户查询某区域的空气质量状况，这时需通过空气质量预报平台发起请求。

2) 用户请求通过网络被发送到指定区域的空气质量监测节点。

3) 指定的监测节点收到请求之后，立即采集空气各成分的信息。

4) 空气质量成分信息通过远程服务器被上传到 HDFS 系统。

5) Hadoop 数据处理平台调用朴素贝叶斯算法分析空气质量成分，并形成空气质量良好或空气污染严重的结论。

6) 分析结果被传送给用户。

该处理过程可以如图 5 所示。

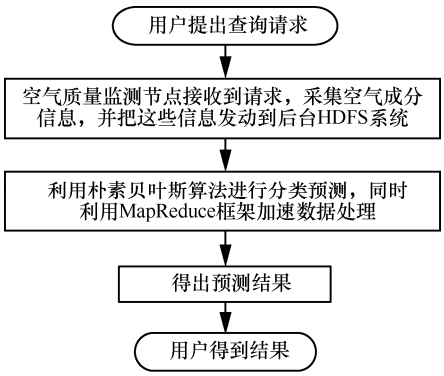

图 5 用户查询的处理过程

在空气质量监测系统中，处理用户主动查询之外，系统本身也会每隔一定时间就对所有的空气质量监测节点进行一次采集，并把生成的分析结果保存在 HDFS 系统或远程服务器，已备用户查询时快速返回结果。

6 海量空气指数数据的快速处理及预报平台

6.1 基于 Hadoop 的海量数据处理

智能空气质量指数预测研究利用 Hadoop 加速基于贝叶斯分类的海量数据分析，其基本框架如图 6 所示。

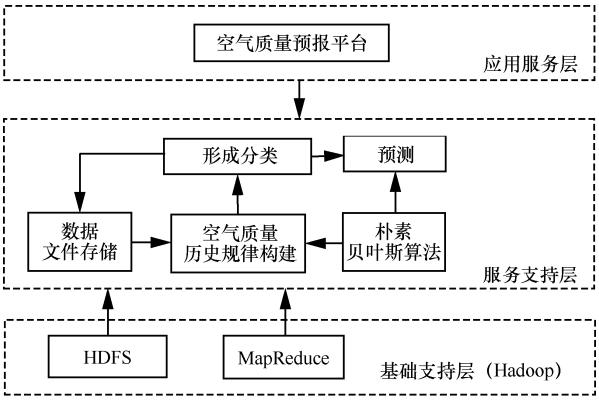

图 6 基于 Hadoop 的空气质量预测框架

6.1.1 数据的分布式存储

首先把获取的空气质量历史数据通过导入程序导入到 HDFS 内，此为预测分析的初始数据，基于 HDFS 的空气质量预测数据存储架构如图 7 所示。

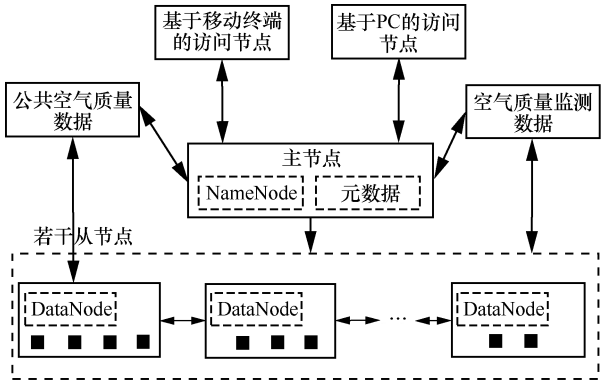

图 7　基于 HDFS 的数据存储架构

存储架构中主要有主节点、数据节点、数据块、客户访问节点、公共空气质量数据获取节点和空气质量检测节点等。HDFS 的工作过程如下。

1) 当有用户（通过访问节点）提出查询请求时，请求信息（包含查询时间和地点）被通过主（master）节点发送给各个从（slave）进行处理；主节点收集处理结果并返回给用户。为了加快处理速度，在实际使用中可以对空气质量历史数据进行预处理，实现分类处理，把分类信息保存在 HDFS 系统中，这样可以提高用户的请求处理速度和结果返回速度。

2) 若有公共空气质量数据或空气质量监测数据需要存储，相关节点会向主节点提出存储请求，主节点选择从节点并把存储位置返回给请求存储的节点，然后请求存储节点就可以直接把数据发送给分配的从节点进行存储。这其实是 HDFS 保存数据的标准流程。

6.1.2 数据的分析和处理

根据数据在 HDFS 系统的存储情况，利用 MapReduce 加速贝叶斯分类的决策过程，对未来的天气进行预测。基于空气质量历史数据的分类事实上有 2 个操作，一是通过对空气质量历史数据的分析得到各个空气质量变化规律；一是对空气质量变化规律应用朴素贝叶斯分类方法进行分类。这 2 个操作都可以使用 MapReduce 框架并行加速处理。

MapReduce 的计算分为 Map 和 Reduce 2 个过程。Map 过程把对全部数据的庞大计算任务拆分成多个任务碎片，并根据每个任务碎片所需要分析的数据的存储位置把计算需求分配给集群中的一个计算资源，每个计算资源处理一部分任务碎片并得到初步的处理结果，这些结果就是 Reduce 过程的输入。这个过程是对历史数据进行判断，得出空气质量情况的过程。Reduce 过程收集各个任务碎片的处理结果，统计并计算概率, 形成最终的处理结果，这些信息进入 MySQL 数据库中保存起来，这个过程也可以由多个计算资源进行合并完成。Map 和 Reduce 2 个过程都实现了数据处理的并行化，提高了数据的处理速度。基于 MapReduce 的分类过程如图 8 所示。

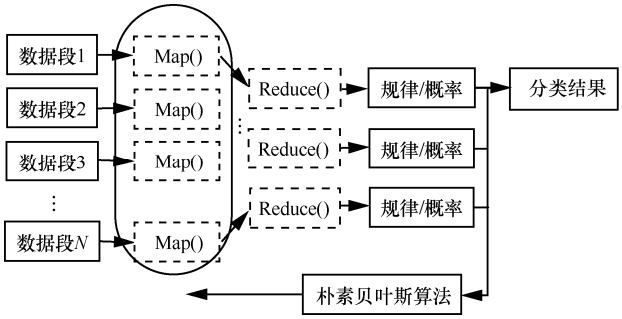

图 8　基于 MapReduce 的分类过程

空气质量的预测就是针对待考察的样本利用贝叶斯方法进行分类、划入某个类别的过程，计算该样本的后验概率 $P(\omega_1 \mid x)$ 和 $P(\omega_2 \mid x)$，并比较 2 个值的大小确定预测结果，这个计算过程在第 3 节已经介绍。

6.2 空气指数预报平台

智能空气质量指数预测预报研究的预报平台被设计为 3 部分：PC 平台、Android 平台和 iOS 平台。PC 平台的预报通过浏览器访问，iOS 平台的预报仍在开发中。这里重点介绍一下基于 Android 平台的空气质量预报。

基于 Android 平台的空气质量预报程序一打开就会查看用户偏好设置中用户是否启用了自动更新空气质量信息的服务，如果已经启用，程序会自动开启 TimeService 这个后台服务，根据用户预先的设置进行空气质量信息更新。它主要包含显示界面和简单数据存储 2 个模块。

6.2.1 界面设计

根据功能需求，初步的 Android 平台空气质量预报客户端包含 3 个主要界面。

1) 当地城市空气指数显示界面可以分为上、中、下 3 个部分，最上面的部分用来显示用户选定

的当前城市的城市名、当日空气质量情况、AQI 指数等，如图 9 所示。

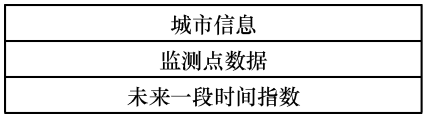

图 9 当地城市空气指数显示界面

2) 各地城市空气指数详情界面大体可分为上下两部分。上部分显示的是全国各城市空气质量较差的排名情况，也就是污染最为严重的城市排名。下部分显示的则是全国各城市空气质量较好的排名情况，反映出空气质量相对优良的城市分布状况。

3) 根据以往空气指数数据的分析结果展开预测，展示接下来一段时间内，空气质量可能发生的变化情况，给即将出行的人们提供了一定的参考作用。

6.2.2 简单数据存储

为了提高反应速度，基于 Android 的空气质量预报客户端会把一部分常用数据存储用到本地。这些数据的存储有 2 种方式：SharedPreferences 和 Files。

1) SharedPreferences 是 Android 平台上一个轻量级的存储类，用来保存应用的一些常用配置。空气质量预报客户端用 SharedPreferences 保存用户的偏好设置，如当前城市、重要城市、计时更新数据等。

2) 空气质量预报客户端用 Files 数据储存方式来存储空气指数信息。它通过调用互联网服务获取空气指数信息，该调用采用 Soap2 协议，通过标准 XML 文件流交互信息，得到的空气监测信息为一个文件输入流对象，保存之后是一个 XML 文件。XML 文件解析处理后保存到数据库中，这些数据可能以后一次也用不上，它们就利用 Files 数据存储方式来存储。

7 实验及结果

本节将介绍智能空气质量指数预测预报研究的实验验证环境、数据获取、运行界面等内容。

7.1 数据集

在研究过程中，用于实验验证的原始空气质量数据主要来自两方面：本地（江苏省盐城市）监测点的采集数据和来自于北京市环境保护检测中心的发布数据。其中，北京市环境保护检测中心的发布数据的发布数据包含从 2014 年 1 月~2016 年 5 月的空气质量相关数据共收取了 23 万多条，数据格式如图 10 所示。数据主要存储与 MySQL，当进行数据分析时上传到 HDFS 中。

7.2 计算环境

Hadoop 环境搭建需要多台计算机组成集群才能搭建。根据实际条件，研究中选择在利用虚拟机构建 Hadoop 运行环境：基于 VMware Workstation 创建 3 台的虚拟机，并使之构成一个小型 Hadoop 集群，进行系统的开发调试，在这个分布式集群上开发的程序可以完美的移植到的实际的 Hadoop 集群中。

研究中主要使用了 Jfinal、Hadoop 等框架，利用 MapReduce 实现贝叶斯算法对空气质量数据进行分析和预测，通过 Java Server Pages（JSP）和 Android 对相关的数据进行展示。

7.3 运行结果及查询界面

本节给出了先验概率和后验概率的计算结果（如图 11 所示）、基于 PC 平台的空气质量查询界面（如图 12 所示）、基于 Android 的空气质量查询及推送界面（如图 13 和图 14 所示）等。

date	hour	type	东四	天坛	官园	万寿西宫	奥体中心	农展馆	万柳	北部新区
20150922	3	SO2	26	17	19	13	26	12	20	4
20150922	3	SO2_24h	25	24	19	19	21	23	14	9
20150922	3	NO2	96	85	64	64	88	95	52	71
20150922	3	NO2_24h	69	63	70	65	95	72	81	48
20150922	3	O3	21	29	59	50	72	14	50	2
20150922	3	O3_24h	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)	(Null)
20150922	3	CO	1.4	1.7	1.8	1.8	2.1	1.9	1.7	1.2
20150922	3	CO_24h	1.4	1.3	1.4	1.5	1.8	1.4	1.4	1.1
20150922	4	SO2	24	13	17	13	17	10	13	4
20150922	4	SO2_24h	25	25	20	19	22	23	14	9
20150922	4	NO2	79	119	62	61	123	73	89	72
20150922	4	NO2_24h	70	65	69	65	95	72	81	49

图 10 获取数据的存储格式

name	forecast	chinese_name	空气较好后验概率====0.333002851813216
万寿西宫	0.24084	AQI	重度污染后验概率====0.02170507955967
定陵	0.18477	AQI	空气较好后验概率====0.35740150514033
东四	0.25721	AQI	重度污染后验概率====0.02270447113424
天坛	0.23322	AQI	空气较好后验概率====0.35433998676576
农展馆	0.24928	AQI	重度污染后验概率====0.025271772191712
官园	0.23697	AQI	空气较好后验概率====0.337065030408288
海淀区万柳	0.23898	AQI	重度污染后验概率====0.023431810529736
顺义新城	0.23873	AQI	空气较好后验概率====0.348835779270264
怀柔镇	0.21386	AQI	重度污染后验概率====0.01863529626728
昌平镇	0.19018	AQI	空气较好后验概率====0.37558788343272
奥体中心	0.24423	AQI	重度污染后验概率====0.020341482139572
古城	0.24562	AQI	空气较好后验概率====0.363625818260428
万寿西宫	0.26164	PM10	重度污染后验概率====0.017609972207452
定陵	0.2275	PM10	空气较好后验概率====0.384378100192548
东四	0.26303	PM10	重度污染后验概率====0.025224397613982
天坛	0.24135	PM10	空气较好后验概率====0.340498058286018
农展馆	0.24073	PM10	重度污染后验概率====0.02736070272123
官园	0.27018	PM10	空气较好后验概率====0.32742690197877

图 11　先验概率和后验概率数据的保存和输出结果

空气监测

城市	地区	pm2.5	pm2.5_24h	首要污染物	污染级别	采集时间	未来预测
北京	万寿西宫	100	53	颗粒物(PM10)	轻度污染	2016-04-09T13:00:00Z	空气较好可以户外运动
北京	定陵	16	34	null	优	2016-04-09T13:00:00Z	空气较好可以户外运动
北京	东四	127	72	细颗粒物(PM2.5)	中度污染	2016-04-09T13:00:00Z	空气较好可以户外运动
北京	天坛	129	62	细颗粒物(PM2.5)	中度污染	2016-04-09T13:00:00Z	空气较好可以户外运动
北京	农展馆	106	61	颗粒物(PM10)	轻度污染	2016-04-09T13:00:00Z	空气较好可以户外运动
北京	官园	67	64	颗粒物(PM10)	轻度污染	2016-04-09T13:00:00Z	空气较好可以户外运动

图 12　基于 PC 平台的空气质量查询

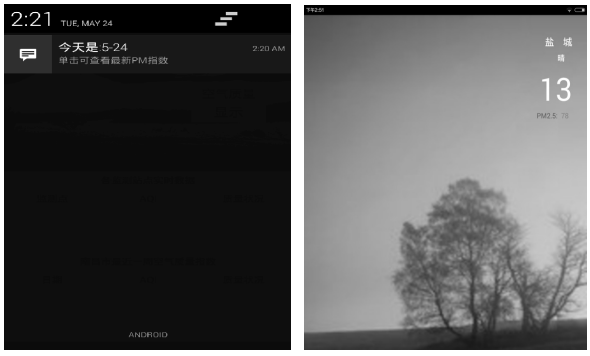

图 13　基于 Android 平台的空气质量简单推送

从计算结果和运行界面来看，智能空气质量指数预测预报研究可以给出高概率正确的预测结果，其预报系统基本可用，初步达到了研究目标。

8　结束语

本文研究了基于热门信息技术的智能空气质量指数预测预报，提出了一种基于大数据分析的空气质量预测系统，从实验验证接过来看，具有一定的可行性。接下来，将从以下几个方面改进该研究工作。

1) 目前仅实验了经典的贝叶斯分类方法，接下来将验证其他热门的分类方法，通过对比选出更适合对空气质量指数数据进行分析的智能决策方法，提高分类的可靠性。

2) 界面的设计比较粗糙，接下来将优化基于 PC 和 Android 的界面、开发基于 iOS 的查询界面，以提高系统的可用性。

3) 进一步优化数据的存储，提高访问数据的速度，减少计算过程中对数据的复制。

参考文献：

[1] 刘毅, 孙秀艳. 雾霾天为何增多[N]. 人民日报, 2011-12-19.

[2] 李龙飞. 世卫警示：雾霾危及健康[N]. 陕西日报, 2013-11-13.

[3] 韩雪萌. 雾霾：中国经济转型中的环境挑战[N]. 金融时报, 2014-02-22.

[4] 李亚红, 肖思思. 直面雾霾下的“肺腑之忧”[N]. 科技日报, 2015-12-22.

[5] 李小飞, 张明军, 王圣杰, 等. 中国空气污染指数变化特征及影响因素分析[J]. 环境科学, 2012, 33(6): 1936-1943.
LI X F, ZHANG M J, WANG S J, et al. Variation characteristics and influencing factors of air pollution index in China[J]. Environmental Science, 2012, 33(6): 1936-1943.

图 14　基于 Android 平台的 3 种空气质量查询界面

[6] 王艳平, 谢正苗. 城市空气环境质量变化趋势中长期预测研究[C]// 中国环境科学学会学术年会, 2009.

[7] 王晓彦, 陈佳, 朱莉莉, 等. 城市环境空气质量指数范围预报方法初探[J]. 中国环境监测, 2015(6).
WANG X Y,CHEN J, ZHU L L, et al. Discussion on the methods of urban ambient air quality index range forecasting[J]. Environmental Monitoring in China, 2015(6).

[8] XU Y Z, YANG W D, WANG J Z. Air quality early-warning system for cities in China[J]. Atmospheric Environment, 2016:148.

[9] MIRANDA A I, FERREIRA J, SILVEIRA C, et al. Teixeira A cost-efficiency and health benefit approach to improve urban air quality[J]. Science of The Total Environment, 2016, (569-570): 342-351.

[10] CHEN Y L, WANG L Z, LI F Y, et al. Air quality data clustering using EPLS method[J]. Information Fusion, 2016, (36): 225-232.

[11] 朱军, 胡文波. 贝叶斯机器学习前沿进展综述[J]. 计算机研究与发展, 2015, 52(1): 16-26.
ZHU J, HU W B. Recent advances in Bayesian machine learning[J]. Journal of Computer Research and Development, 2015, 52(1): 16-26.

[12] 李正杰. 基于 Hadoop 平台的数据挖掘分类算法分析与研究[D]. 南京: 南京邮电大学, 2016.
LI Z J. The analysis and research of data mining classification algorithm based on Hadoop platform[D].Nanjing: Nanjing University of Posts and Telecommunications,2016.

[13] DIENES Z. How Bayes factors change scientific practice[J]. Journal of Mathematical Psychology, 2016, (72): 78-89.

[14] TAHERI S, MAMMADOV M. Learning the naive Bayes classifier with optimization models[J]. International Journal of Applied Mathematics and Computer Science, 2013, 23(4):787-795

[15] 刘智慧, 张泉灵. 大数据技术研究综述[J]. 浙江大学学报(工学版), 2014, 48(6): 957-972.
LIU Z H, ZHANG Q L. Research overview of big data technology[J]. Journal of Zhejiang University(Engineering Science), 2014, 48(6): 957-972.

[16] IQBAL R, DOCTOR F, MORE B, et al. TEMPORARY REMOVAL: big data analytics: Computational intelligence techniques and application areas[J]. International Journal of Information Management, 2016, (9).

[17] CHANG V, RAMACHANDRAN M, WILLS G, et al. Editorial for FGCS special issue: big data in the cloud[J]. Future Generation Computer Systems, 2016, (65): 73-75.

[18] SRINIVASAN S. Cloud computing evolution[M]. New York: Springer, 2014.

[19] XUE C Y, LIU F, LI H H, et al. Research and design of performance monitoring tool for Hadoop clusters[M]. India: Springer, 2014.

作者简介：

刘燕（1986-），女，江苏盐城人，盐城工学院讲师，主要研究方向为数据处理、环境监测等。

张永平（1979-），男，河北邯郸人，博士，盐城工学院讲师，主要研究方向为大数据技术、压缩感知、物联网等。

朱成（1989-），男，江苏常州人，主要研究方向为物联网技术。

皋军（1971-），男，江苏盐城人，博士，盐城工学院教授，主要研究方向为人工智能、机器学习、模式识别等。

刘其明（1965-），男，江苏盐城人，硕士，盐城工学院副教授，主要研究方向为软件工程、算法流程与分析。