第 43 卷 第 6A 期
2016 年 6 月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

Spark 在人类基因领域的应用

丁东亮 吴东月 于福利
(天津理工大学自动化学院 天津 300384)

摘 要 人类基因组作为一种具有高价值的、弥足珍贵的大数据信息,亟待人们进行高效、准确的分析处理。由于传统 Hadoop 云框架的数据处理存在高延迟的致命缺点,Spark 云平台应运而生。基于 Spark 云平台的人体基因组数据系统将为疾病的早期发现或治疗以及降低婴儿的出生缺陷等做出巨大贡献。

关键词 人类基因,大数据,Spark
中图法分类号 TP301 **文献标识码** A

Application of Spark in Human Genome

DING Dong-liang WU Dong-yue YU Fu-li
(School of Electrical Engineering, Tianjin University of Technology, Tianjin 300384, China)

Abstract The human genome as a kind of the high-value and the precious big data is badly in need for efficient and accurate analysis. In view of the high-latency fatal weakness of the traditional cloud framework Hadoop, the cloud platform Spark emerges in response to the needs of times. The human genome data system based on Spark will make great contributions to the early detection or treatment of the disease and birth defects.

Keywords Human genome, Big data, Spark

1 引言

1.1 人类基因组也是一种大数据

一提到大数据的概念,人们首先想到的就是互联网上琳琅满目的数据,往往忽视了与人类关系最密切的基因组数据。1953 年,沃森和克里克提出了 DNA 双螺旋结构,之后人类便致力于挖掘 DNA 的全部信息,探求生命的全部秘密。1990 年,人类基因组计划正式通过,其目的在于确定人类 DNA 的总体结构,弄清其中各种基因的结构、功能、位置和相互关系,从整体上认识遗传信息的组成及其调控方式,促进生命科学和医学的发展[1]。

据统计,人类 23 对染色体上共含有约 30 亿对碱基对,完全测序的人类个体基因组数据量为 100~1000GB,即使进行压缩处理,数据量仍然约有 3GB,其中包含的数据信息量可以达到一个小图书馆程度的数据量。另外,考虑到从对某一致病基因的开始检查到最终确诊,期间需要进行一系列定性、定量和定位分析,而有的基因可能天天变化,有的则要每月检查,有的则按季度检查,甚至有的按年检查,这其中又包含了大量的信息。因此,人类基因组是一种含有海量信息的数据,这是大数据的第一个特点[2]。人类基因组计划涵盖了基因组学、蛋白质学、代谢组学等多学科的基因数据,其中大多的数据都是非结构化的。因此,人类基因组是一种类型和结构多样化的数据,这是大数据的第二个特点[2]。目前,人们已经明确认识到 3000 个左右的基因与特定的疾病有关,利用现代化设备进行的基因组测序工程每天都能产生百万兆的新数据,平均一个星期就能发现一个新的致病基因。因此人类基因组是一种快速产生并更新的数据,这是大数据的第三个特点[2]。

综上所述,人类基因组也是一种大数据,并且对人类而言是具有高价值的、弥足珍贵的大数据信息,因为一个人的健康情况全体现在这个大数据里面。

面对人类基因组如此庞大的、非结构化的、快速更新的数据信息量[3],在传统计算机上进行的数据挖掘技术显得束手无策,要在合理时间内挖掘出具有实际价值的基因信息,结合云计算平台的并行化数据处理方式将为该问题的解决提供一种有效的途径。

1.2 数据挖掘

数据挖掘是指在海量的、不完全的、有噪声的、模糊的、随机的数据中提取具有隐含关系的、人们感兴趣的、具有一定深入研究价值的信息的过程,是统计学、数据库技术和人工智能技术等多学科的综合[4]。数据挖据过程通常包括 3 个步骤,即数据准备、数据挖掘和结果的表达和解释,如图 1 所示。

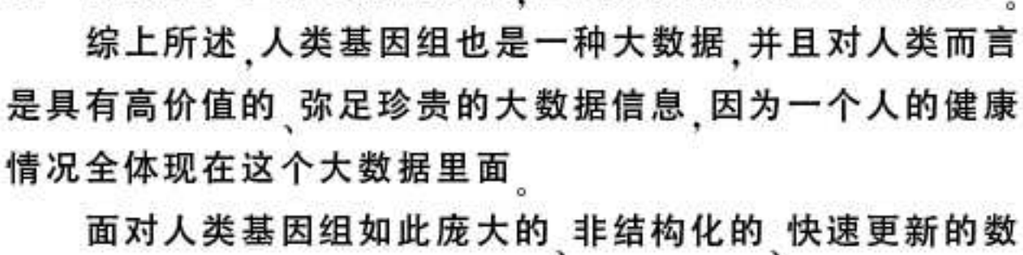


图 1 数据挖掘过程

数据挖掘的主要任务是进行数据关联分析、聚类分析、分类分析、异常分析和演变分析等,其最大价值在于利用提取出

丁东亮(1992—),男,硕士生,主要研究方向为生物大数据,E-mail:1310463647@qq.com;吴东月(1982—),男,博士生,讲师,主要研究方向为模式识别与智能系统;于福利(1976—),男,博士生,教授,主要研究方向为模式识别与智能系统。

来的信息改善甚至改变现有的预测模型[5]。数据挖掘技术具体到人类基因组方面的应用价值，就是要对特定基因（如癌症基因）的特性进行归纳，之后利用挖掘出来的基因特性信息对相应疾病进行及时的治疗或有效的预防，并预测致病或致残基因及特定人群患病的概率，从而在一定程度上降低出生缺陷，并据此研究新药物等。

2 云计算

云计算是一种“按需给予”的服务模式，利用虚拟化、一致性管理、负荷均衡、网络技术、分布式文件存储等技术将多台普通计算机组接成为一个拥有高计算力、高容错性、高存储能力、高拓展性、方便易用的运算平台[6]。其基本原理是通过网络技术将存储和计算处理程序分送到大量的分布式计算机中，并借助于相应的应用程序服务，允许用户将资源切换到需要的应用上，根据需求访问计算机和存储系统。要借助于云计算技术实现对海量数据的分析与处理，需要满足 3 个条件，即海量数据分析的存储和访问、海量数据分析的管理与组织以及海量数据分析的并行处理[7]。

为了进一步提高数据挖掘的准确性，加快数据处理的平均运行速度，并行化数据处理方式应运而生。并行化数据处理方式，即首先将海量的待处理数据集划分为无差别的 n 个数据子集，然后将 n 个数据子集发送给对应的多个子处理器，子处理器单独对数据进行相应处理，最后将每个子处理器的处理结果整合，得出整个数据处理的结果。与串行化处理方式相比，并行化处理将显著缩短数据的平均运行时间。

美国 DNANexus 公司借助于谷歌的云计算和大数据平台，建立了一个开放式的 DNA 数据库；IIlumina 公司推出了基因测序云计算平台 Base Space，该平台允许用户免费存储一定数量的测序数据。在国内，华大基因推出了第一个具有自主产权的云服务产品 Easy Genomics TM。该平台集基因组学领域内常用的和华大基因特有的数据及参数为一体，结合云存储和高性能计算技术，能够以更低的成本、更高的效率完成大量的基因数据处理及分析，但是其对具体使用的平台没有详细介绍。

2004 年，谷歌公司发布 MapReduce、GFS、BigTable 的云计算处理平台，为大数据框架技术的研究奠定了坚实的基础；之后，基于 MapReduce、GFS 的设计思想，提出了 Hadoop 开源计算平台。Hadoop 是一种对海量数据进行处理的并行计算模型，主要包括分布式计算和分布式存储两部分。分布式计算 Hadoop 依托 MapReduce，分布式存储部分依托 GFS。

MapReduce 的设计思想是将自动分割要执行的问题分为 Map（映射）和 Reduce（归化）的方式，在自动分割后通过 Map 程序将数据映射成不相关的区块并分配给大量计算机处理，从而达到分散运算的效果，再通过 Reduce 程序将结果汇整，输出开发者需要的结果[8]。MapReduce 在每次执行的时候都需要从磁盘中读取数据，数据处理完成后还要把结果数据存放到磁盘上，其执行过程如图 2 所示。MapReduce 的分布式计算思想，使得 Hadoop 能够轻松面对海量数据的批量处理问题，同时，Map 和 Reduce 算子实现简单，编程开发者仅需要解决数据逻辑的问题，而不必考虑分布式系统的具体实现机制。

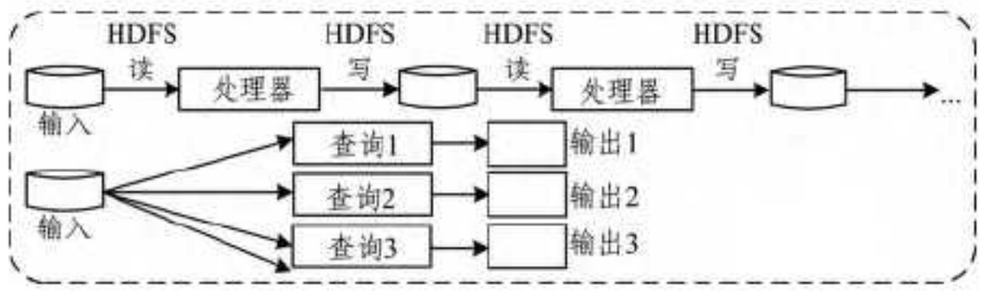


图 2 MapReduce 执行过程

GFS 的数据存储设计思想中，一个 GFS 集群是由一个主服务器和多个子服务器组成的，文件被划分成固定大小的块，存储在多个子服务器中，这样可以保证数据的高可用性。主服务器管理整个文件系统的所有元数据。客户端与主服务器的交互仅需要基本的操作元数据，其他数据的获取和存储则是直接与对应的子服务器通信完成。

Hadoop 最大的优点就是操作简便，另外 Hadoop 还提供了 Streaming 编程工具，使得用户直接使用 C++、python 等语言就可以实现 Mapper 和 Reducer。2008 年，Hadoop 成为 Apache 的顶级项目，之后便逐渐成为分布式计算和海量数据处理的主流平台[9]。2009 年，Schatz 等人首次在基因组数据中使用 Hadoop，此后，Hadoop 和 MapReduce 成为生物信息学中使用最广泛的并行数据处理框架。

但是 Hadoop 平台也存在着很多问题。由于 Hadoop 平台是基于 MapReduce 思想提出的，MapReduce 是一种批量数据处理的构架，因此在面对实时或流式的数据访问时，Hadoop 并不适合。其次在 MapReduce 思想中，将 Map 中间计算的结果存入到本地磁盘，再通过相应机制发送给 Reduce 进行数据的分析处理，这样的工作模式使得 Hadoop 适合于离线处理数据，而不适合需要大量网络通讯的任务，同时并不是所有的数据处理任务都可以抽象为 Map 和 Reduce 两个算子，并且因为其模型的计算较抽象，对于结构不一致的数据也不能很好地进行自动分割。另外，数据库高级索引系统的缺乏使得 Hadoop 对某些数据类型的分析效率较低，当系统中没有相应索引时，就要在整个数据集中寻找，极大地延长了数据处理的平均运行时间。因此，如何充分发挥 Hadoop 的海量数据处理能力，同时又有效避免 MapReduce 思想带来的高延迟的致命弱点，是开发云计算平台应该着重考虑的问题。

3 Spark 云平台

针对 Hadoop 云平台存在的诸多问题，人们进一步开发了 Spark。基于内存计算的分布式框架 Spark 是由加州伯克利 AMP 实验室研发的一款开源通用并行云计算平台，基于“One Stack to rule them all”的理念成功成为了一体化、多元化的大数据处理平台，其体系架构如图 3 所示[10]。Spark 起源于 Hadoop 的开源社区，计算过程包含了 MapReduce 的思想，但是其并不局限于 MapReduce 的两个阶段式范型。

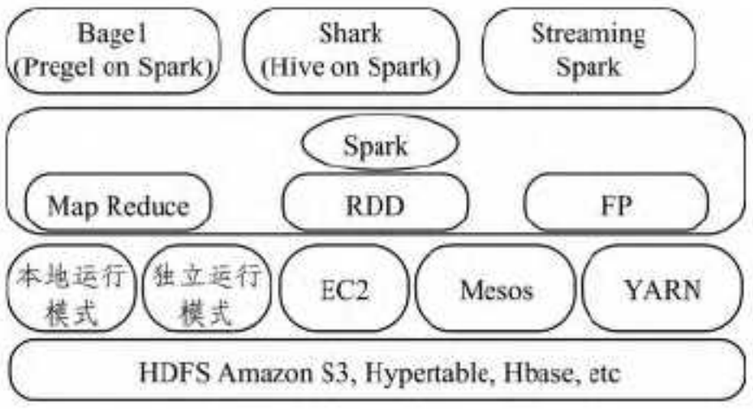


图 3 Spark 的体系架构

3.1 Spark 的特点

MapReduce、RDD、FP 是 Spark 的 3 个主要特点，核心是基于内存计算的抽象对象 RDD(Resilent Distributed Dataset，弹性分布式数据集)，它允许用户在执行多个查询时显式地将工作集缓存在内存中，后续的查询能够重用工作集，这极大地提升了查询速度。另外，RDD 增加了可供使用的数据集操作类型，如 map、filter、flatMap、groupByKey、reduceByKey、union、join 等 transformation(转换)，充分发挥多节点的计算性能，同时 RDD 还提供了 count、collect、reduce、lookup、save 等多种 action(动作)。Transformation 是将原来的 RDD 构建成新的 RDD；action 是通过 RDD 对数据进行处理分析，并将处理结果返回到驱动程序或者保存到外部存储系统。Spark 基于内存的计算特点，在某些应用上的实验性能要超过 MapReduce100 多倍，即使是基于磁盘性能的 Spark 也要比 Hadoop 快 10 倍以上[11]。

RDD 之间有两种依赖关系，即窄依赖和宽依赖。窄依赖是指一个父 RDD 最多可以被一个子 RDD 引用，宽依赖是指一个父 RDD 可以被多个子 RDD 引用。在容错机制方面，RDD 是通过对丢失数据的重建来实现的。另外还可以通过对转换操作的限制来降低容错开销。RDD 尽管并不是一个通用的内存共享抽象，具备了良好的描述能力、可伸缩性和可靠性，能够广泛适用于数据并行类应用。

3.2 流数据处理平台

流数据处理平台主要是针对需要接受大量的、不间断的数据(称为流式数据)，且可以迅速完成数据处理并响应的系统[12]。Spark 体系中采用了当前最为流行的流数据处理平台——Spark Streaming(流处理)。Spark Streaming 构建在 Spark 之上，是 Spark 的核心子框架之一，是一种准实时大规模流式数据处理框架，数据处理的延迟可以达到大约 1s，能够满足实时查询的要求，也能共享 Spark 平台的许多优势，其运行原理示意图如图 4 所示。

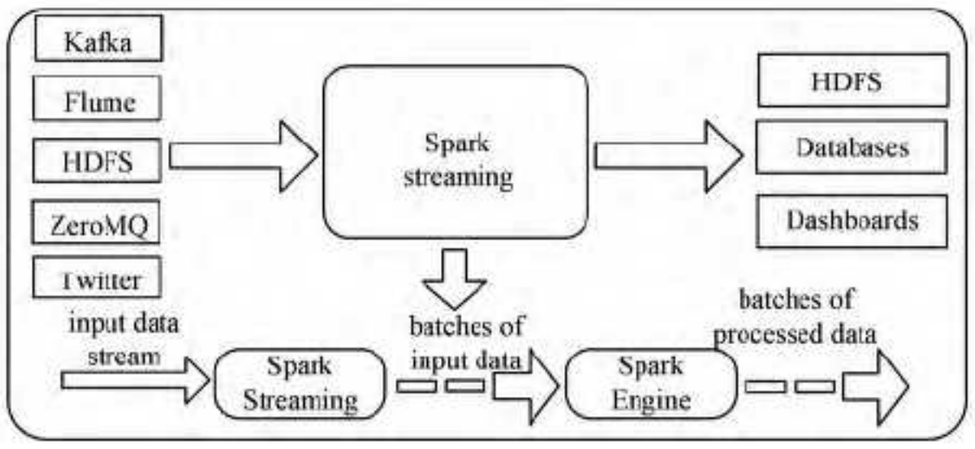


图 4 Spark Streaming 运行原理示意图

Spark Streaming 把 Kafka、HDFS、Socket 等系统作为流处理的数据来源，将输入的数据流用时间切片的方式分为一个个小的 Batch，然后将这些 Batch 交给 Spark 引擎去处理。

3.3 资源管理系统

考虑到 Spark 本身不具备对计算任务的管理能力，需要加入第三方资源管理平台来完成调度分配任务[13]。Spark 自带的 Standalone 模式能够满足绝大部分纯粹的 Spark 计算环境中对集群资源管理的需求。在 Spark 集群中运行多套计算框架时，尽管其自带的 Yarn 模式资源调度不能进行动态调整，但可以在 Hadoop2.6 下得到解决，并且能够很好地与 Spark 结合，为 Spark 提供任务调度的能力。其运行流程图如图 5 所示。

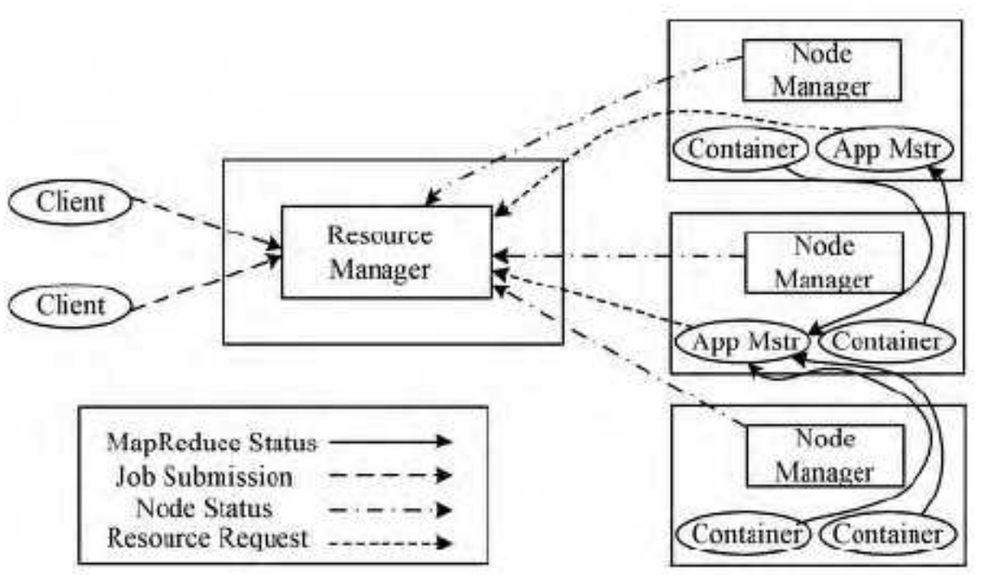


图 5 Yarn 运行流程图

ResourceManager 负责将集群的资源分配给相应的应用；NodeManager 是计算节点，负责启动应用所需要的 Container，并把资源的使用情况反馈给 ResourceManager；Container 是资源分配和调度的基本单元。Yarn 采用的是事件驱动并发模式，在该模式下将各种处理逻辑抽象成事件和调度器，将事件的处理过程用状态机表示。

综上所述，相比于 Hadoop，无论是性能还是扩展性，Spark 都更具优势，Spark 正在加速成为一体化、多元化的大数据处理中心的首选和唯一的计算平台。图 6 是一个基于 Spark 的人类基因组测序云平台架构。

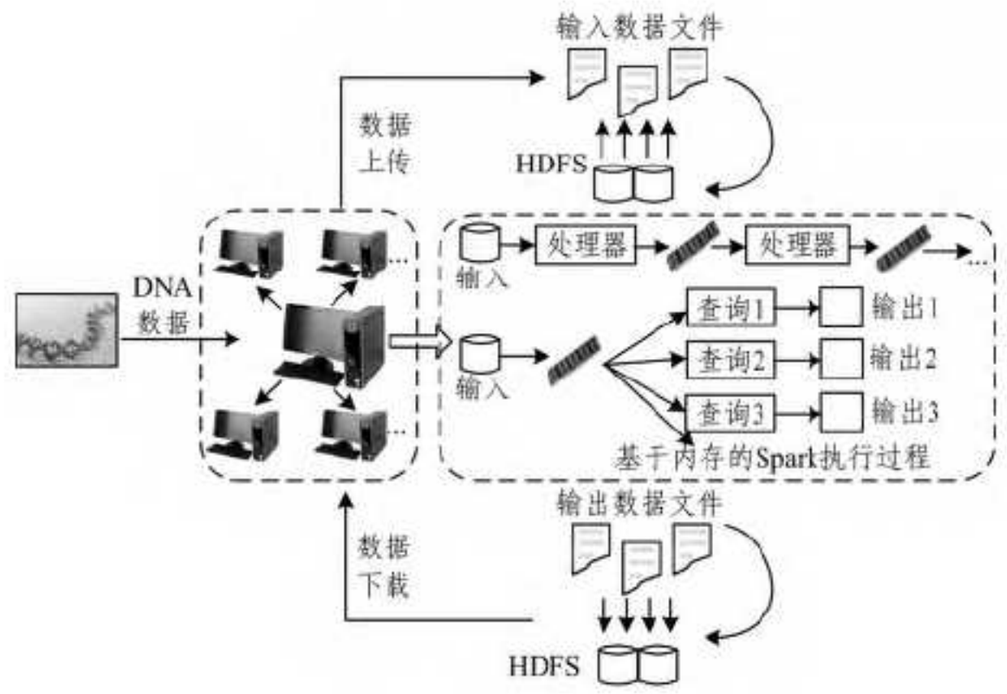


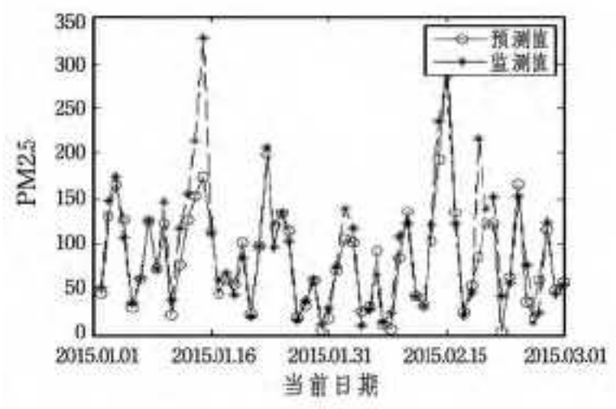
图 6 基于 Spark 的人类基因组测序云平台架构

4 结论

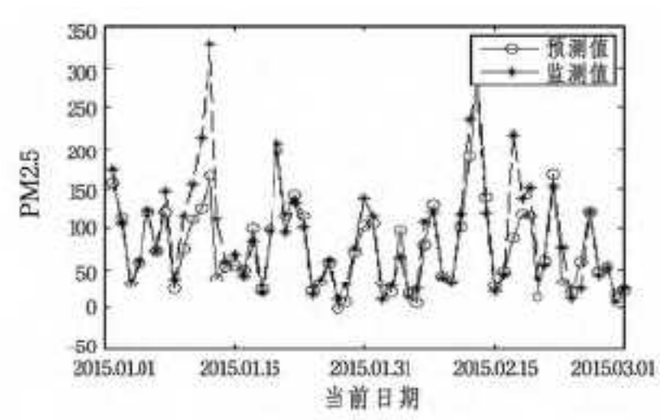
信息量巨大，更新频繁，数据类型和结构复杂，成为制约人类基因组工程发展的最大挑战。目前广泛使用的 Hadoop 云计算平台又存在高延迟的致命缺点，而且除了 Hadoop 存在的缺陷，云计算本身也存在一定的问题。首先，利用云计算来分析处理基因组数据，就必须先把数据放进去，即使网速很快，海量数据的上传也需要一定的时间。其次，云计算作为一种相对新颖的事物，生物医学研究者对其仍然持有怀疑的态度，在云环境下，一些敏感的数据和病人信息发生泄漏的可能性很大，因此云计算的可靠性、隐私安全等问题受到挑战。但尽管如此，云计算仍然是解决基因组问题的最有效方法。

利用 Spark 云平台对人体基因组数据进行有效的分析处理，一方面可以使人们发现常见致病基因的特征，如癌症、肿瘤疾病等，从而采取相应的预防或早期治疗措施，从临床角度来看，胃癌能否有效地治愈，与其发现的早晚有密切关系，早期治愈率极高[14]；另一方面可以有效降低婴儿的出生缺陷，

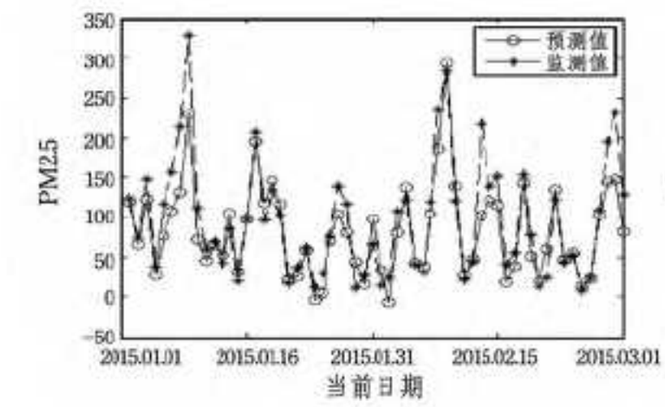
(下转第 528 页)



(a)第一天的 PM2.5 预测与监测结果



(b)第三天的 PM2.5 预测与监测结果



(c)第七天的 PM2.5 预测与监测结果

图 3　未来 7 天 PM2.5 预测结果

表 1 给出了其未来七天雾霾预报结果的统计值。从表 1 可以看出,在预测过程中,随着预报周期的变长,空报率和漏报率都有所增加,命中率有所降低,但是命中率都维持在 80%以上,预测精度较高,预报结果比较满意,从而验证了所采用预测模型及预测方法的有效性。

表 1　其未来 7 天雾霾预报结果的统计值

指标 \ 天数	第一天	第三天	第七天
空报率	0.1000	0.1000	0.1167
漏报率	0.0500	0.0500	0.0667
命中率	0.8500	0.8500	0.8167

结束语　所提模型通过多元线性回归,采用了在线更新的预测方式,在保证预测精度同时,无需大量预测数据。与传统的大数据离线预测模式不同,每天根据当天检测结果,不断更新模型,既保证了预测的精度,又无需大量的预测数据,并可以及时反映新情况的变化。根据本文的原理设计了雾霾预测小助手之类的软件,为后续手机 APP 设计提供了参考。

参 考 文 献

[1] 杨准. 雾霾现象成因初步探讨[J]. 科技创新导报,2014(34):21

[2] 王继志,杨元琴,周春红,等. 雾霾低能见度天气分析与预测方法研究[C]//2007 年中国气象学会年会论文集. 2007:145-149

[3] 张琳,胡雪红. 德州市雾霾客观预报方法[J]. 农技服务,2010,27(11):1493-1493

[4] 艾洪福,石莹,等. 基于 BP 人工神经网络的雾霾天气预测研究[J]. 计算机仿真,2015,32(1):402-405

[5] 侯琼煌,杨航. 基于三次指数平滑模型的雾霾天气分析与预测[J]. 环境保护科学,2014(6):73-77

[6] 王勇,黄国兴,彭道刚. 带反馈的多元线性回归法在电力负荷预测中的应用[J]. 计算机应用与软件,2008,25(1):82-84

[7] 李军成,陈国华,石小芳. 基于灰色多元线性回归的粮食产量预测[J]. 安徽农业科学,2010,38(16):8281-8282

[8] 周晨,冯宇东,肖匡心,等. 基于多元线性回归模型的东北地区需水量分析[J]. 数学的实践与认识,2014,44(1):118-123

[9] 周永生,肖玉欢,黄润生. 基于多元线性回归的广西粮食产量预测[J]. 南方农业学报,2011,42(9):1165-1167

[10] 李莉,孙永霞. 基于均值化主成分分析的雾霾环境分析与研究[J]. 计算机应用研究,2015(5):1373-1375

(上接第 504 页)

例如先天性盲人疾病就是一种基因病,可以通过基因分析对其进行有效筛选。在未来,基因测序服务能够为个性化健康管理提供科学依据。

结束语　针对人类基因这样的大数据分析处理问题,文中介绍了一种新的云计算平台 Spark。Spark 较传统的 Hadoop 平台有了很大程度的改进,极大地缩短了数据处理的运行时间,同时集成了一些资源管理平台,使得对数据的分析处理更加方便、简单,基于 Spark 的人类基因系统的研究具有广阔的应用前景。

参 考 文 献

[1] 赵广荣,杨冬,白姝,等. 现代生命科学与生物技术[M]. 天津:天津大学出版社,2008

[2] 严霄凤,张德馨. 大数据研究[J]. 计算机技术与发展,2013,23(4):168-172

[3] 孙磊,胡学龙,张晓斌,等. 生物医学大数据处理的云计算解决方案[J]. 电子测量与仪器学报,2014,28(11):1190-1197

[4] 胡秀. 基于 Web 的数据挖掘技术的研究[J]. 软件导刊,2015,14(1):149-150

[5] 米允龙,米春桥,刘文奇. 海量数据挖掘过程相关技术研究进展[J]. 计算机科学与探索,2015,9(6):641-659

[6] Zhang Shu-fen, Yan Hong-can, Chen Xue-bin. Research on Key Technologies of Cloud Computing[C]//2012 International Conference on Medical Physics and Biomedical Engineering. 2012:1791-1797

[7] 卢小宾,王涛. Google 三大云计算技术对海量数据分析流程的技术改进优化研究[J]. 图书情报工作,2015,59(3):6-11

[8] 崔杰,李陶深,兰红星. 基于 Hadoop 的海量数据存储平台设计与开发[J]. 计算机研究与发展,2012,49(21):12-18

[9] 林清滢. 基于 Hadoop 的云计算模型[J]. 现代计算机,2010(7):114-116

[10] 王家林. 大数据 Spark 企业级实战[M]. 北京:电子工业出版社,2015

[11] Wu Leng-dong, Yuan Li-yan, You Jia-huai. Survey of Large-Scale Data Management Systems for Big Data Applications[J]. Journal of Computer Science and Technology, 2015, 30(1):163-183

[12] 陈虹君. 基于 Hadoop 平台的 Spark 框架研究[J]. 电脑知识与技术,2014,10(35):8407-8408

[13] 胡俊,胡贤德,程家兴. 基于 Spark 的大数据混合计算模型[J]. 计算机系统应用,2015,24(4):214-218

[14] 所剑,王大广,刘泽锋. 早期胃癌诊断和治疗[J]. 中国实用外科杂志,2011,31(8):717-719