

硕 士 研 究 生 读 书 报 告

徽标

描述已自动生成

题目 大数据相关文献阅读

作者姓名 陈文儒

作者学号 21951428

指导教师 张志猛

学科专业 软件工程

所在学院 软件学院

提交日期 2020年10月1日

Big data related literature reading

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Zhimeng Zhang

By

Wenru Chen

Zhejiang University, P.R. China

2020

摘要

本文探讨了近年来与大数据相关的技术。在复杂的大数据环境下，传统的数据处理方式和数据处理平台效率和效果不佳，大数据相关技术就显得尤为重要。文章将就大数据相关的技术，从全局到局部进行比较，对技术的特点优势缺陷进行讨论。

**关键词**：大数据

**Abstract**

This paper discusses the technologies related to big data in recent years. In the complex big data environment, the efficiency and effect of traditional data processing methods and data processing platform are not good, so big data related technology is particularly important. This paper will compare the technologies related to big data from the global to the local, and discuss the characteristics, advantages and disadvantages of the technology.

**Keywords：big data**

1引言

我们每天生活都被大数据所包围，移动手机每天的上网数据、线下超市每日的交易记录、每日医保的消费记录……随着计算机计算机网络、云计算、移动互联网等的发展，每天数据的产生量都以空前的速度增加着。传统的数据存储和处理方式，对于如此庞大的数据，有效率低下，成本偏高，可拓展低等缺点。大数据中，往往蕴含着巨大的价值，因此针对大数据的存储、处理方式就显得尤为重要。

各行各业，各国各地，大数据的形式、项目等不同，使得产生了许许多多的模型、框架，人们为了这些数据处理地更准确、可靠而产生了大数据应用。然而，处理大数据并不是一件容易的事，就算是有现成的技术，选择何种技术，如何对数据进行操作也是有些困难的；并且，在做大数据时应考虑很多因素：兼容性、部署复杂性、成本、效率、性能、可靠性、支持和安全风险。

在接下来的内容中，将介绍近年来大数据的相关技术，以及作者相关的切身体会。第二章节将介绍大数据和大数据相关应用及其难点，第三章将介绍Hadoop相关内容，第四章将介绍亲身处理大数据的经历，第五章为小结。

**2 大数据相关**

在当今的商业环境中，数据管理可能是决定成功与否的主要因素，大多数企业已经开始意识到数据战略的重要性。大数据不仅仅是单一的一种技术或方法，相反，大数据跨越了商业和技术的众多领域的范畴。大数据包含很广，结构化、非结构化、半结构化的数据都可以组成大数据的一部分。大数据不仅仅包含数据，它还包括着高级算法、先进技术。

**2.1 大数据的定义**

大多数数据学家用一下三种主要特征定义大数据：

1. Volume：大量的移动设备、和应用产生的庞大的数字数据。
2. Velocity：数据的产生和处理十分迅速，并且能从中提取出十分有用的信息和相关内容。
3. Variety：大数据是从分布式的多样来源产生的，具有多样性。

**2.2 大数据的应用**

以下是几个大数据的应用：

智能电网：智能电网至关重要的一点是通过复杂的智能仪表、传感器、控制中心等来管理国家电力能源消耗，以及监管智能电网的操作。通过对大数据的分析有助于识别有风险的变压器和检测所连接设备的异常行为。因此网格计算是最佳的选择。它对生成的大数据实时分析，并对场景建模，从而能更好的管理优化资源。

互联网医疗：从医院获取的远程数据、个人上传的数据，这使医疗服务能够个性化。医生能够在线监测患者症状，以便调整处方；根据人群症状调整公共卫生计划，疾病进化和其他参数。它也有助于优化医院业务降低医疗成本支出。

智能电网：在德国，为了鼓励利用太阳能，会在家庭安装太阳能，除了卖电给你，当你的太阳能有多余电的时候还可以买回来。通过电网收集每隔五分钟或十分钟收集一次数据，收集来的这些数据可以用来预测客户的用电习惯等，从而推断出在未来2~3个月时间里，整个电网大概需要多少电。有了这个预测后，就可以向发电或者供电企业购买一定数量的电。因为电有点像期货一样，如果提前买就会比较便宜，买现货就比较贵。通过这个预测后，可以降低采购成本。

推荐系统： Gracenote公司拥一种技术，它能在行车时了解真实喜好听的歌曲。它采用智能手机和平板电脑内置的麦克风识别用户电视或音响中播放的歌曲，并可检测掌声或嘘声等反应，甚至还能检测用户是否调高了音量。这样，Gracenote可以研究用户真正喜欢的歌曲，听歌的时间和地点。Gracenote拥有数百万首歌曲的音频和元数据，因而可以快速识别歌曲信息，并按音乐风格、歌手、地理位置等分类。

**2.3 大数据面临的难点**

大数据挖掘带来了许多机遇，然而，人民在进行数据挖掘的过程中，面临着许多问题。例如：数据的获取、数据的容量问题、大数据搜索问题、数据分析、数据管理、数据可视化等等。此外，尤其在分布式数据驱动应用上的数据安全和隐私问题。下面是几个现在大数据中不可忽视的问题：

1. 大数据管理

数据学家面临着一项挑战：如何以尽可能少的计算机软件和硬件资源收集，一体化和存储数据。而另一项挑战则是数据管理，数据管理是大数据分析处理的基础。大数据管理影响着数据清理、数据分析等后续步骤。大数据管理是为了确保数据可靠，从而使得数据便于获取、方便管理、准确存储和受到保护。

1. 大数据清理

传统数据管理有五个步骤：清理、整合、编码、存储、处理。原始数据可能存在噪声、错误、不完整数据。一个可靠的数据分析结果，往往需要，准确的数据集作为支持。那么在大数据的清理数据中，难点就在于，如何在如此庞大的一个数据集中找出问题数据，如何分辨出可靠数据、有用数据。

1. 大数据聚合

如何将外部数据源和分布式大数据平台与组织的内部基础设施同步是个难点。外部数据可以包括第三方来源、市场波动信息、天气情况和交通状况、来自社交网络的数据、客户评论和公民反馈。例如，做慢病预测时，将天气指数与肺病的内部数据对照，可以加大模型的可靠性。

1. 数据集不平衡

对不平衡数据集进行分类，这一问题在过去几年里引起了广泛的关注。在实际应用中可能会有不同分布的类别，把实例数少到几乎可以忽略的类称为少数类或正类，而把有大量实例的类称为多数类或负类。这个问题在医学诊断、软件缺陷检测、财务、药物发现或生物信息学等各个领域，十分凸显。经典的学习方法不适用于不平衡数据集，这是因为模型的构建是基于全局搜索的，而不考虑实例的数量。实际上，一个模型的体现的规则通常是大多共性的，而不是特定的，因此在模型构建过程中忽略了少数群体。拿自身经历来说，有一个需求：神经网络输出维度为4000多个，但正类仅10来个，余下全是负类，那么这在训练时就遇到了问题，网络模型往往会向负类发展。而解决这一难题的一个方法是使用FocalLoss增加正类的权重。

1. 大数据与机器学习

机器学习的目的时从数据中学习到有用的东西，并作出较为准确的决定，通常分为有监督学习、无监督学习以及强化学习。大数据与机器学习与的应用十分广泛，例如推荐系统、人脸识别、道路检测等等。

数据流学习：在某些应用中，如微博、股票、智能电表等，会产生的大量数据集，而数据挖掘能在这些数据流中提取出有用的模型以及巨大的价值。然而，传统的数据挖掘技术如关联挖掘、聚类和分类等在动态环境下应用于此类大数据集时缺乏效率、可扩展性和准确性。由于数据流的大小、变化速度和它的可变性，不可能将它们永久地存储起来然后进行分析。因此，研究人员需要找到新的方法来优化分析技术，在有限的时间内用有限的资源（如内存）处理数据实例，并产生实时准确的结果。此外，数据流的可变性会给传入的数据流带来不可预测的变化。

深度学习：目前，深度学习已经成为机器学习和模式识别领域中一个非常活跃的研究领域。它在计算机视觉、语音识别和自然语言处理等预测分析应用中发挥着重要作用。深层学习是基于层次学习和不同层次的复杂数据抽象的特征提取，因此它适合于简化对大数据量的分析、语义索引、数据标注、信息检索和区分性任务（如分类和预测），将原始数据（如图像的像素值）转换成适当的内部表示或特征向量的特征提取程序，学习子系统（通常是分类器）可以从中检测或分类输入中的模式。然而，尽管有这些优势，大数据仍然给深度学习带来重大挑战：海量大数据、异质性、数据高速产生、噪声标签和非平均分布。

此外还有增量和集成学习、粒度计算等，面临大数据表现出不同的困难。

**3 大数据与Hadoop**

Apache Hadoop是一种著名的大数据技术，它的目的是避免传统技术处理和分析大数据时所遇到的低性能和复杂性。Hadoop的一个主要优点是它能够快速处理大型数据集，这得益于它的并行集群和分布式文件系统。事实上，与传统技术不同的是，Hadoop并不是在内存中复制整个远程数据来执行计算，而是在存储数据的地方执行任务。因此，Hadoop减轻了网络和服务器的通信负载。Hadoop的另一个优点是它能够在保证容错性的同时运行程序，为了保证这一点，它通过在服务器上复制数据来防止数据丢失。Hadoop平台的强大功能基于两个主要的子组件：Hadoop分布式文件系统（HDFS）和Map-Reduce框架。此外，用户还可以根据自己的目标和应用需求（如容量、性能、可靠性、可扩展性、安全性）在Hadoop之上根据需要添加模块。

下面将介绍数据存储、数据处理和数据查询；此外还有数据获取，数据分析，和数据管理。

**3.1 数据存储**

1. Hadoop分布式文件存储系统 (HDFS)

HDFS是一种数据存储系统，它可以在一个集群中支持多达数百个节点，并提供经济高效和可靠的存储能力。它可以处理结构化和非结构化数据，并保存体积庞大的数据（存储的文件可能大于1 TB）。HDFS的主要优点是它可以跨异构硬件和软件平台进行移植。此外，HDFS将数据运算转移到存储数据的设备或者相邻设备，有助于减少网络拥塞并提高系统性能。

1. Hbase

HBase是一个分布式的、面向列的非关系型开源数据库，HBase中的所有数据文件都存储在Hadoop HDFS文件系统上。HBASE的目标是存储并处理大型的数据，更具体来说是仅需使用普通的硬件配置，就能够处理由成千上万的行和列所组成的大型数据。HBASE是Google Bigtable的开源实现。HBASE的优势有： 1）线性扩展，随着数据量增多可以通过节点扩展进行支撑；2）数据存储在HDFS上，备份机制健全；3）通过zookeeper协调查找数据，访问速度快。

**3.2 数据处理**

1、Map-Reduce

Map-Reduce通过高效、经济的机制简化了海量数据的处理。它能够编写支持并行处理的程序。它将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数：Map和Reduce。它采用“分而治之”策略，一个存储在分布式文件系统中的大规模数据集，会被切分成许多独立的分片（split），这些分片可以被多个Map任务并行处理。

**3.3数据查询**

1、Hive

Hive是一个数据库系统，旨在简化Hadoop的使用，hive允许在用户更熟悉的结构化数据库中表示数据。hive的数据模型主要基于表，表示HDFS目录，并被划分为多个分区。hive还提供了一种类似sql的语言hiveql，允许用户访问和操作存储在HDFS或HBase中的基于Hadoop的数据。因此，蜂巢是适用于许多商业应用。

**4 大数据与机器学习在医疗领域**

如今的生活，几乎处处收到了大数据和机器学习的影响。人们在日常的网络生活中无意之间透露出来的数据被加以利用。很多商户可以知道用户这段时间喜欢买什么东西；使用搜索引擎时，能把人民心中所想的大概率放到前面；逛电影网站的智能推荐也精准无比。

甚至大数据和机器学习在医疗领域的应用也取得惊人成果。一个病理医生需要大量的时间和专业的训练，才能对病理图像进行准确的判断，并且其一天的完成量也是有限；而用大数据和机器学习的方法，只要一段时间的训练，就能以较高的准确度完成训练，并且速度比人工方法快很多。在医保领域，异常的医保诊断，不是老道的医生难以辨别，而用大数据和机器学习的方法，也能够准确的找出问题数据。大数据和机器学习的方法，在医疗诊断、预测、异常检查等等方面，均有不错的表现。

下面是两个亲身经历的项目：

**4.1 癌症病理图诊断**

对于病理图像诊断，人工阅片耗时耗力，所以用大数据和机器学习结合的方法辅助诊断是十分有必要的。针对宫颈癌病理切片图像进行研究，采用基于卷积神经的方法，对病理图像进行分类诊断。

这里介绍的，是对宫颈细胞切片数字图像进行检查，判断出病理特征。

1. 目标检测识别细胞核

数据集是临床采集的512\*512大小的宫颈细胞病理切片图像，其中对应的标注信息为手工标注的XML文件。

总共标注的图像数量为503，标注的细胞核个数为19693。

如图4.1，左为原图，右为标注图像展示（图像按比例缩小）。

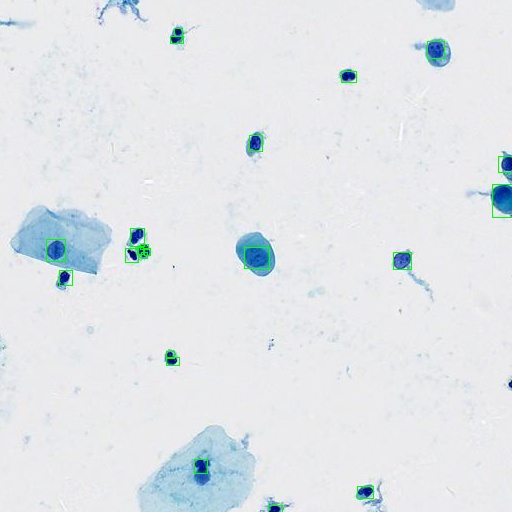
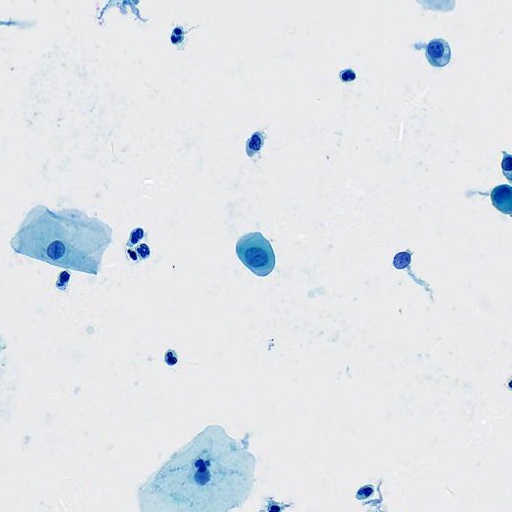
 

图4.1 数据展示

对于这样的目标，采用了Shot MultiBox Detector（SSD）目标检测算法。SSD基于一个前向传播的卷积神经网络，它会产生一批Bouding Boxes(即物体框)，这些Boxes都是固定的而不是通过训练得到，而后将得出每个Box类别和相应的Score（即可能性）。最后，再进行非极大值抑制，选择最优的结果，即为得到的最终预测结果。

最终结果如下：

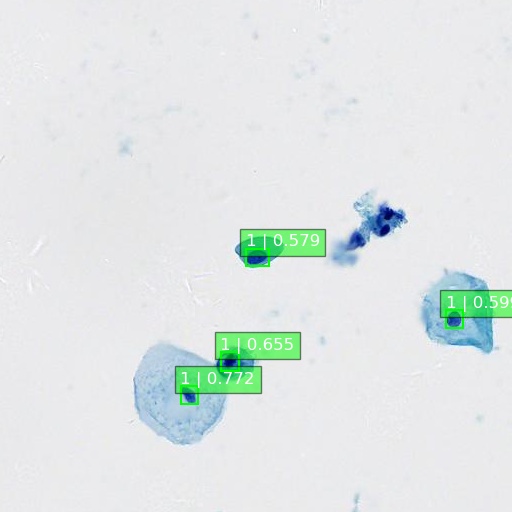


图4.2 结果图像展示

1. 分类算法对病理图像分类

使用的数据集是临床采集的32\*32、64\*64、128\*128、256\*256四个尺寸大小的宫颈细胞病理切片图像，每个尺寸下有八类细胞，分别如表3.1所示：

表3.1 细胞类别表

|  |  |
| --- | --- |
| 类别编号 | 类别 |
| 101 | 正常表层细胞 |
| 102 | 正常中底层细胞 |
| 103 | 粒细胞 |
| 104 | 腺细胞（子宫颈管细胞） |
| 105 | 非典型鳞状细 |
| 106 | 挖空细胞 |
| 107 | 高核浆比细胞 |
| 120 | 垃圾 |

每个尺寸下有11772张图像，总计有47088个图像。

不同尺寸图像如图3.13。

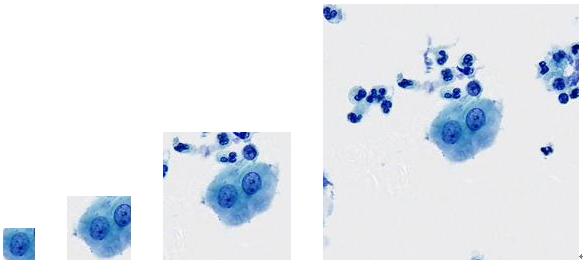


图4.3 不同尺寸

不同类别图像如图3.14。

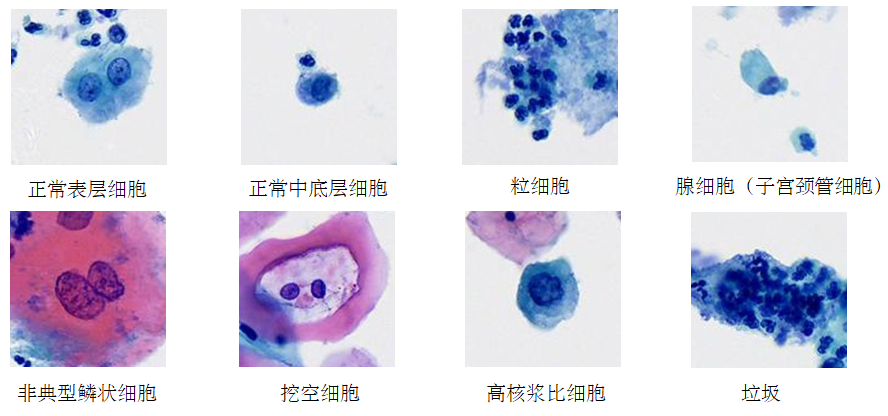


图4. 4 不同类别

这里采用了ResNeXt作为分类算法的模型。ResNeXt可以说是脱胎自ResNet，但却优于ResNet。它的目的是保持ResNet的高可移植性优点，同时继续提高精度。传统提高网络性能的方法不是增加网络宽度就是增加网络深度，而ResNeXt可以做到相同网络复杂度的情况下提高精度。ResNeXt与相似复杂的ResNet相比，计算量仅仅为后者一半。

最终结果如下：

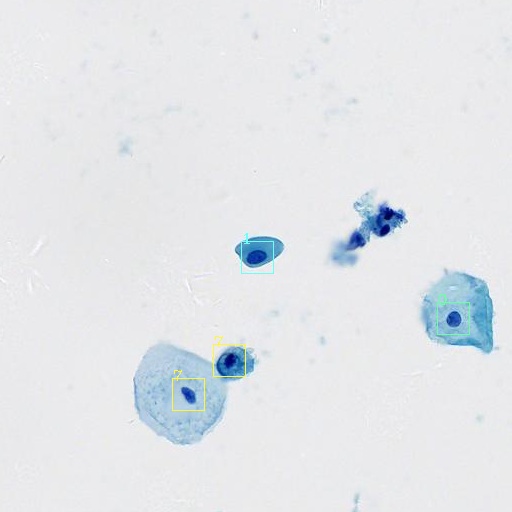


图4.5 结果图像展示

最终，当两个模型训练完毕以后，使用预测功能时，数十张图片，数百个细胞的批量识别，仅仅用了几秒的时间。

**4.2 医保用药异常检测**

随着社会发展，社会福利也随之提升，近年来，医保的普及度和抵扣力度大大提升。随之而来一个问题：医保的基金是否用到了正确的地方？可能，在某些地方存在着，恶意刷医保、故意开贵药的情况。这时候，就需要一个方法来检测这些异常行为。

由于每日，不同的医院产生的医保数据可能非常复杂并且数据巨大，传统人工检测，或传统算法难以胜任如此庞大的数据。所以针对医保用药的异常检测，可以使用机器学习的方法进行。

这里采用的数据集是浙江省近三年的医保数据，原始数据总共有7416570条记录。每条记录形式为疾病组合及其对应的处方药方。如图4.6：



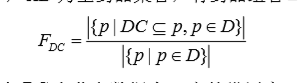
图4.6 数据集展示

据不完全统计（数据来源方提供），本数据集中全部的疾病约1-2W种类别，药品约2-4W种类别。药品组合的不同药方（已去重）有140W+种。原始数据，由于是来自浙江省内不同医院，所以差别大， 且没有进行过清理。这无论对传统算法，或者机器学习都是一个挑战。

首先，对于原始数据，进行了数据清理。因为对于原始数据，可能存在异常数据、不规范数据、不可用数据等等。使用一个疾病编码——疾病名称对应表，与数据进行对照，剔除第一部分异常数据。而后，使用药品编码——药品名称对照表，与数据进行对照，剔除第二部分异常数据。然后，对于不规范的数据，再对格式进行调整。最后，将药品进行压缩，把相同药，不同厂家的药品合并为一种。最终得到可用数据4824118条，疾病8811种，药品4049种。

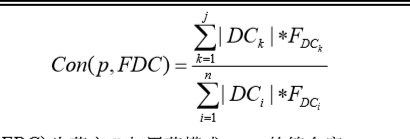
1. 尝试聚类算法

根据数据特点，常用聚类算法难以胜任，所以这里采用了基于常用度（FDC）算法。公式4-1为常用度公式。

 公式4-1

DC为某药品组合（药方），D是所有药品组合，的意思是包含DC的药方数占总药方数的比例。常用药方：FDC，大于一定值的药方。

用药符合度公式5-2：

 公式4-2

药方 p 与用药模式FDC 的符合度是指用药模式FDC 中包含在药方 p 中药品组合的加权和与总体药品组合的加权和之比。这里所取权重是药品组合中元素的数目和常用度的乘积。

异常药方：

设 FDC为用药模式，p 为一个药方，若 Con（ p ，FDC） γ < （其中 γ 为一个工作人员根据实际需要规定的阈值），则药方 p 为相对于用药模式 FDC 的异常处方。

但该方法出现了个问题：1、指标的设置，对不同病种的影响不同；例如一个只有10条数据的疾病诊断，常用度为0.1，可能全部为正常数据。而一条10W条的数据，同样的常用度为0.1， 可能有60%的异常数据。

1. 使用深度学习

由于数据分布不均匀，聚类算法难以实现，因此使用神经网络进行尝试。由于药方组合有140W，所以，用药品作为输出，网络的输入维度8000\*1，输出维度4000\*1。

这里采用了改版的wavenet，具体的网络结构不再赘述。通过wavenet对概率的一种把控，将造出疾病和药品之间的联系，最后一层改为输出sigmoid，根据没中过药品对应的概率来判断异常情况。

**5 小结**

目前大数据处理有多种平台与技术的支持，是否正确的选择处理大数据的方式策略，将意味着项目的成功与否。本文，对大数据进行了一些基本介绍，讲了大数据在生活中的应用，大数据面临着哪些挑战和难题；介绍了大数据的一个平台——Hadoop的相关内容；还用实践经历体会了大数据的种种难点。

大数据相对于传统数据，其实本质上是相同的，不过大数据相对来说更多了数据的数量、数据的多样性，数据的流动性变快等等；这些复杂的变化使的大数据处理起来不是那么容易。针对不同的大数据，采取的方法也不同，采取合适数据处理方式的是至关重要的。

文末的实践经历所体会的难点1、数据多，2、数据分布不均匀。对模型训练可能几天时间，但是数据清洗、对数据处理的方法的尝试却占去了大部分时间。大数据具有大前景，但大数据面临难题也十分众多，也因此大数据是一项大挑战。如何正确选择处理方法，需要不断尝试。

参考文献

[1] Oussous A, Benjelloun F Z, Lahcen A A, et al. Big Data technologies: A survey[J]. Journal of King Saud University-Computer and Information Sciences, 2018, 30(4): 431-448.．

[2] Beam A L, Kohane I S. Big data and machine learning in health care[J]. Jama, 2018, 319(13): 1317-1318.．