C. S. R. Prabhu

雾计算，深度学习和大数据分析-研究方向

Springer

雾计算，深度学习和大数据分析-研究方向

C. S. R. Prabhu

雾计算、深度学习和大数据分析-研究方向

Springer

C. S. R. Prabhu
印度政府电子和信息技术部国家信息中
心（NIC）（退休）印度新德里

ISBN 978-981-13-3208-1　　ISBN 978-981-13-3209-8（电子书）
https://doi.org/10.1007/978-981-13-3209-8

国会图书馆控制号：2018961192

© Springer Nature Singapore Pte Ltd. 2019
本作品受版权保护。 出版商保留所有权利，包括整体或部分
材料的权利，特别是翻译、重印、插图的再利用、
朗读、广播、微缩胶片复制或以任何其他实体方式复制，以及传输
或信息存储和检索、电子适应、计算机软件，或通过类似或不同的
已知或今后开发的方法。
在本出版物中使用一般描述性名称、注册名称、商标、服务标志等，并不意味着即使在没有具体声明的情况下，这些名称也不受相关保护法律和法规的限制，因此可以自由使用。

出版商、作者和编辑可以安全地假设本书中的建议和信息在出版日期时被认为是真实和准确的。 出版商、作者或编辑对本文所含材料不提供任何明示或暗示的保证，也不对可能存在的任何错误或遗漏承担责任。 出版商在已发表的地图和机构 affiliations 中保持中立。

这个 Springer 品牌由注册公司 Springer Nature Singapore Pte Ltd. 出版。
注册公司地址为：152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
Singapore

前言

今天的互联网正与大量的物联网设备或传感器连接在一起预计到2020年将有500亿台设备连接到互联网。 物联网驱动的全球经济将面临许多技术挑战。

由于物联网设备的计算和存储能力非常有限，需要从云端获取增强资源的支持。 这导致网络过载，带宽有限。 云服务的性能取决于位于云数据中心的服务器，这些数据中心位于离物联网设备所在的边缘或现场位置较远的核心网络附近。 这导致端到端延迟比交付云服务所需的值高出两个数量级。 这种延迟问题进一步加剧了物联网设备实时生成的大量数据，如果实时分析，可能具有很大的价值。 然而，如果将数据发送到云端，数据量过大会超过存储系统和分析应用程序的处理能力。 将数据传输到云端并获取分析结果所造成的延迟使得满足实际应急响应和健康管理等实际情况的需求变得不可行，因此需要一种新的范式，即雾计算。本书提供了对雾计算技术的全面概述。在多个定义之后，介绍了各种雾计算架构，以满足在健康、交通、智慧城市、智慧村庄、智能家居、增强现实、设备对设备连接、交通管理等各种应用领域的需求。对雾中应用管理中的延迟感知应用管理和资源协调问题进行了调查。 详细调查了雾分析，介绍了即将推出的模型和产品，如智能数据和雾引擎。 处理分布式雾生态系统中出现的安全和隐私问题，并介绍了一些现有的方法，如身份验证和加密。 挖掘和提取海量输入数据中的有意义模式，用于决策、预测和其他推理，是大数据分析的核心。除了分析海量数据外，大数据分析还面临其他独特的挑战

机器学习和数据分析，包括原始数据的格式变化，快速流动的流式数据，数据分析的可信度，高度分布式输入源，嘈杂和质量差的数据，高维度，算法的可扩展性，不平衡的输入数据，无监督和未分类的数据，有限的监督/标记数据等。在大数据分析中，还存在着充足的数据存储，数据索引/标记和快速信息检索等关键问题。因此，当处理大数据时，需要创新的数据分析和数据管理解决方案。在大数据分析的背景下，从深度学习算法中学到的知识尚未得到充分利用。某些大数据领域，如计算机视觉[242]和语音识别[241]，已经广泛应用深度学习来改善分类建模结果。 深度学习从大量数据中提取高级、复杂的抽象和数据表示的能力，特别是无监督数据，使其成为大数据分析的有价值工具。具体而言，深度学习可以更好地解决语义索引、数据标记、快速信息检索和判别建模等大数据问题。传统的机器学习和特征工程算法无法高效地提取大数据中通常观察到的复杂和非线性模式。通过提取这些特征，深度学习使得在处理大数据规模时可以使用相对简单的线性模型进行大数据分析任务，如分类和预测。最后，介绍了雾计算这一新兴领域中多样化的方向和广阔的研究机会。 还提供了在大数据分析中部署深度学习技术所涉及问题的研究展望。

关键词 雾计算 · 雾分析 · 雾安全 · 移动雾 · 延迟

印度新德里 C. S. R. Prabhu

目录

"关于作者

"**C. S. R. Prabhu**博士在印度政府和各个机构担任过重要职位。"他曾任印度国家信息中心（NI**C**）的总干事，印度电子和信息技术部，新德里，并在塔塔咨询服务（T**CS**），**C**M**C**，TE**S**和TEL**C**O（现塔塔汽车）的各个职位上工作。他还是亚洲生产力组织（A**P**O）项目的国际资源教员，并代表印度参加了A**P**O在日本大阪举办的Vent**ura** 2004国际论坛。他曾在佛罗里达中央大学任教和研究，并在NA**S**A卡纳维拉尔航天局担任顾问的短暂时间。

他被一致选举为印度计算机学会（CSI）海得拉巴分会主席。他目前在安德拉邦农业大学（KL University）担任顾问，并在海得拉巴的凯沙夫纪念技术学院（KMIT）担任研究与创新主任。

他在1978年获得印度理工学院孟买分校电气工程专业的硕士学位（M.Tech），专业为计算机科学。此前，他在1976年获得贾瓦哈拉尔尼赫鲁技术大学的电子与通信工程学士学位。他指导过博士研究生和硕士学位学生，并发表了多篇论文。

摘要

这本书《雾计算、深度学习与大数据分析-研究方向》全面介绍了物联网（IoT）、物联网应用、雾计算的定义、架构、实时或延迟敏感应用管理、雾分析、雾安全和隐私问题。 此外，它还探讨了在雾计算背景下与大数据分析相关的深度学习技术。最后，它指出了雾计算、深度学习和大数据分析融合的许多研究方向。

第1章
引言

摘要 本章介绍了雾计算。它提出了物联网、智能城市和其他物联网应用所面临的技术挑战，以及大数据和物联网之间的协同作用。它介绍了雾计算范式作为解决方案的出现方式、雾计算的特点、雾架构、IoX、云端、ParStream、Prismatic Vortex等产品。它深入探讨了雾平台的组成部分，以及智能家居、智能村庄、健康监测和支持、农业、智能车辆和增强现实应用等雾应用。

1.1 基于物联网的新经济从2015年开始出现

1.1.1 物联网的出现

如今，越来越多的物联网设备和传感器正在连接到互联网上。这些数字以前是无法预料的。 物联网或称为物联网一切，预计将连接越来越多的消费电子设备、家用电器、医疗设备、摄像头以及各种温度、压力或湿度传感器等，除了手机和工业物联网设备。 数字令人震惊：根据爱立信的预测，预计将有多达500亿台设备连接。这将在不到10年的时间内带来20万亿的市场机会。据估计，到2025年，物联网将对经济产生影响，每年产生的收入和运营节省高达11万亿美元，占世界经济的11%，普通公众用户将在2025年之前部署多达1万亿台设备，即物联网驱动经济的技术变革。 在互联网上的大数据演进阶段，我们有物联网或称为物联网一切[175]或一切互联网（IOE）作为最近和最新的趋势和发展的重要推动力。 物联网可以被描述为电信行业、软件行业和硬件行业（包括设备制造行业）等不同活动领域的交互和互操作领域，为工业和商业各个领域带来巨大的机遇。

© Springer Nature Singapore Pte Ltd. 2019

C. S. R. Prabhu,雾计算、深度学习与大数据分析-研究方向, https://doi.org/10.1007/978-981-13-3209-8_1

为了提供一个正式的定义，'物联网是一个无缝连接的嵌入式对象/设备网络，具有标识符，其中使用标准和互操作的通信协议（手机、平板电脑和个人电脑也包括在物联网的范围内）可以进行机器对机器（M2M）通信，而无需任何人为干预。'物联网（IoT）[176]是由数万亿个传感器设备的输入驱动的，这些设备与智能子系统接口，并与数百万个信息系统相连。 物联网（IoT）将推动业务和客户利益的前所未有的新视野，这将需要越来越多的智能系统，这些系统可以自动驱动IT/IoT行业和供应链公司中越来越多的商机。

连接到互联网的设备数量已经超过了70亿的人类数量，并且预计到2020年，全球连接到互联网的设备数量将达到30-50亿 [178]。

1.1.2 智能城市和物联网

为了为城市和农村地区的居民提供更好的生活质量，物联网设备如传感器正在智能城市和智能村庄中得到部署。全球范围内，智能城市已经被提出、设计并正在大量实施。

物联网在智能城市实施中的应用包括以下主题：

1. 智能停车
2. 智能城市照明
3. 智能交通系统 [7]
4. 智能垃圾管理
5. 远程护理
6. 女性安全
7. 智能电网
8. 智能城市维护
9. 数字标牌
10. 水管理

智能家居 [179] 物联网服务有助于通过更轻松、更便捷地监控和操作家电（如空调和冰箱）来提升个人生活方式。

在彼此交互时，物联网设备产生大量的数据。 处理这些数据并整合物联网设备，可以建立一个系统。 智能城市系统和智能村庄系统的发展基于物联网和大数据。这样的系统开发和实施包括数据生成和收集、聚合、过滤、分类、预处理、计算，并在决策时完成。

行业应用

除了智能城市，智能和远程控制设备可以帮助解决农业、健康、能源、安全和灾害管理等不同行业面临的各种问题。

这是电信行业一方面和系统集成商另一方面的机会，通过实施物联网应用来增加收入，这种技术可以提供各种服务。 此外，IT行业可以提供直接服务，也可以提供与物联网无关的分析相关服务或其他服务。

1.1.3 物联网的阶段和利益相关者

1.1.3.1 物联网的阶段

可以确定物联网实施的四个阶段：

阶段1：确定适用于应用程序的传感器。
阶段2：开发应用程序。
阶段3：服务器应接收应用程序的传感器数据。
阶段4：使用数据分析软件进行决策过程。

所有主要国家都在实施物联网应用方面采取了行动。

1.1.3.2 利益相关者

物联网实施的关键利益相关者包括以下几个：

1. 公民；
2. 政府；和
3. 行业。

每个利益相关者都必须展示合作产生结果的承诺。
利益相关者在每个阶段或步骤的参与是必不可少的。制定促销政策对于回答以下问题至关重要：“哪些数据将为公民提供服务？”物联网应明确制定战略，以实现“价值提升”和“成本降低”的目标模式。

1.1.3.3 实际下采样

从实际情况来看，将物联网从数十亿的愿景下采样到现实中的数千个物联网设备，它们松散地相互连接

是要实现的。 此外，许多这样的设备与移动设备连接或嵌入其中，具有P2P、2G/3G/4G/5G和Wi-Fi连接功能。

提供云中心的连接，用于数据收集和适当的大数据分析，可能还可以进行个性化。 开放的生态系统，没有供应商锁定是必不可少的。

1.1.4 分析

分析可以针对来自传感器设备的‘小’数据或来自云端的‘大’数据进行。 根据需求，两者需要结合起来。

1.1.5 从边缘到云端的分析*[179]*

我们不能将所有数据都推送到云端进行分析。 在智能手机和潜在的间歇通信的背景下，我们需要决定何时/何时将‘大’数据的子集推送到手机，何时/何时将‘小’数据的子集推送到云端，以及如何自动进行协作决策。 通信和计算能力、数据隐私和可用性以及最终使用的应用程序决定了这些选择。

*1.1.6*物联网（*IoT*）中的安全和隐私问题和挑战

安全和隐私问题在物联网及其应用中带来挑战。

在基于物联网的数据安全中，存在着三个众所周知的关键问题：（a）数据机密性，（b）数据隐私和（c）信任。 除了众所周知的与用户相关的安全问题，如身份验证、完整性和访问控制，还存在以下问题： 此外，身份保护和隐私对用户来说也很重要。

根据一些研究方法，上述安全和隐私问题可以在物联网生态系统的多个层面上进行处理，包括：（i）传感器或硬件层面，（ii）数据通信层面，（iii）上下文注释层面，（iv）上下文发现层面，（v）上下文建模层面和（vi）上下文分发层面。

为了全面了解，我们必须按以下方式检查安全问题：

1. 不同类型的物联网威胁
2. 协议
3. 隐私

4. 身份管理
5. 信任和治理
6. 容错性
7. 动态信任
8. 安全和隐私管理
9. 网络安全
10. 治理
11. 信任

隐私：我们试图通过防止个人、财务或技术细节的披露或公开暴露来提供隐私。 这可以通过实施数据控制技术来实现，例如- 加密- 匿名化- 聚合- 同步- 集成。

身份识别/访问控制：这包括控制人员/物体对受限区域的合法入侵。 这可能涉及车辆的识别和定位、湿度和温度的测量、产品的跟踪、敏感区域的监视参数管理等。信任：信任可以建立，并包括以下内容：- 人与智能物体之间的相互信任；- 提供安全保证；- 提供透明度。

如果实施上述原则，将在全球范围内形成一个生态系统，使需要的人能够及时获得信息。

在敌对环境中，智能物体具有保护自身的能力。

可以通过询问节点来检查节点的可信度。可靠性：在一个过程（例如制造过程）中，需要确定信息的可靠性并编制结果。 这需要适当的感知和测量手段、度量标准、计量校准、处理诊断、异常检测和定期维护。 此外，为了实现更高的整体可靠性，需要开发和建立自动和灵活的自适应控制系统。

确保物联网系统隐私的机制：在普遍存在的物联网应用系统中，敏感数据可能以分布式方式存储。 因此，有必要建立一个适当的控制机制，根据各自的敏感级别控制和管理数据的披露给第三方。

数据隐私需要在数据收集、数据传输和数据存储的各个阶段考虑。

以下技术可以提供可能的解决方案：

(1) 匿名化
(2) 块密码
(3) 流密码
(4) 哈希函数
(5) 随机数生成函数和
(6) 轻量级公钥基元

访问隐私涉及到可能的访问信息的方式，这可能是个人的。需要部署具有不同约束条件的高效策略和机制来管理各种类型的数据。

1.1.7 访问

我们需要确保将网络靠近传感器。 成千上万个传感器的广泛普及将导致电力和通信带宽的限制。 与依赖可能大规模部署定制传感器网络相比，借助现有标准可能更有价值。 点对点数据可以用于最后一英里连接的目的。 我们还可以集成高功能的网关和云。 非对称架构可能更可取。 技术的普及在不同国家甚至同一国家的不同地区并不均衡，反映了基础设施、接入、成本、电信网络、服务、政策等方面的不平等和差异。因此，需要相应地提供价格合理的传感器设备、网络成本和分析解决方案。

1.1.8 成本降低

在物联网的成功实施中，重复使用已部署的基础设施将是不可避免的要求。在发展中国家或贫困国家，部署可重复使用的设备和传感器以供新颖的使用方式更为可取，因为采用先进基础设施的先进国家的模式将无法为发展中国家或贫困国家提供相关或具有成本效益的解决方案。

1.1.9 机会和商业模式

物联网设备和网络中流动的数据将为分析师提供无尽而广阔的机会。 传感器技术的有效相互作用

设备、电信基础设施和通过它们流动的数据导致了新的商业模式的出现。

技术模型的成功或失败取决于产生收入的商业模式。 对于谷歌在互联网上，收入的繁荣来自使用谷歌用户的个人数据进行广告，以换取免费的信息搜索和检索服务提供给最终用户。 随着物联网（IoT）的出现，捕获的信息将变得庞大、微观和快速。 对于所有这些活动的使用、重复使用、共享和付款都很重要。 可以提供数据经纪服务，可以利用开放数据与企业共享。 这可能会产生诸如更大的共同利益之类的回报，除了鼓励用户与企业共享开放数据政策的明确回报，无论是货币、同行认可还是更大的共同利益。

1.1.10 内容和语义

以人为中心，内容和语义构成了决策的核心数据。 基于内容的挖掘和分析对内容变得更有意义和相关。 因此，必须有一种机制来捕捉能够描述系统行为和社会行为的语义。

有时，智能代理或软件代理可能会取代人类用户，因此可能会了解个人信息。 例如，我们有Siri和Cortana这样的软件代理。它们将与服务提供商、公用事业公司和供应商公司的其他代理进行交互。

1.1.11 基于数据的物联网商业模式

语义内容可以扩展M2M数据交换中的数据内容。 基于数据的商业模式将不断发展。物联网（IoT）带来的新技术和应用领域以低功耗、更便宜的设备为特点，基于强大连接的更多计算。 这项技术提供了一个窗口，可以以微观的细节来观察我们的生活和环境，几乎可以监测到一切。

物联网业务模型可以是水平的或垂直的，可以与供应链集成，对最终用户来说可以成为一个价值主张。 这样的价值主张是多样化且数量众多的，有许多用户需求或愿望需要满足。

第一步是开发成本效益高的传感器和执行器，进行低成本和微观级别的观察。 下一步是构建一个业务模型，可以进行数据的收集、存储、经纪和持续时间。第三个业务模型将针对数据的分析组合。 端到端的系统集成是第四个业务模型。

总结起来，基于物联网的业务模型的发展将需要集中的努力来支持架构的模块化，通过基于服务的模型提供能力，以解决真实的客户问题。

1.1.12 物联网的未来

1.1.12.1 技术驱动因素

技术驱动因素的未来包括低成本和准确的传感器和执行器开发，以及它们的网络技术以及计算和内存能力。 因此，我们可以说，理解物联网技术是通过理解设备、它们的网络和它们的计算配置来实现的。 设备和传感器不断被微型化，它们的成本也在下降。90年代国防部推动的“智能尘埃”技术基于使用MEMS和电子纺织品、织物或可穿戴计算机的设备，这些都是物联网设备和传感器的驱动因素。

智能织物及其在体育或医学中的商业应用已成为会议议题。

1.1.12.2 未来可能性

预计到2020年，大部分迄今为止的手动流程将被自动化流程和产品所取代，供应链将嵌入传感器和主动设备。 集成设备的可穿戴设备或服装可以在体育和艺术（舞蹈）训练中发挥作用，也可以用于医疗环境中的患者监测。

1.1.12.3 挑战和关注点

物联网传感器网络中的隐私问题如何解决？ 虽然让太阳镜识别房间里的一个人可能是个好主意，但通过基于云的监控应用程序来监控自己是否正确？ 许多新的倡议和应用想法带来了商业机会，但伦理、安全和隐私问题也浮出水面。 物联网有无数无尽的工业和商业应用：运输和卡车车队跟踪、快递和邮件跟踪、环境感知和监测，这些都涉及到集成在智能手机中的特殊设备和传感器。在体育中，球和球员可以附着传感器以获得更好的进球准确性。 在机场、仓库和运输过程中追踪和定位货物和包裹。 带有传感器的漫游机器人可以减少仓库中的库存量。 监测数据中心的温度，监测昆虫，

蜜蜂、候鸟、手势识别、视频游戏、虚拟现实（VR）等都是可能的应用。 城市和农村消费者应用，如智能家居、智能农场和智能牧场，可以在线监测温度、湿度等参数，以提高基于反馈的高效用电或用水。 在欧洲、韩国和美国，智能电网、智能城市和智能村庄等应用都使用物联网设备和传感器进行监测和基于反馈的行动。 石油和天然气行业早已部署了这些设备很长时间了。

不受隐私问题影响的应用可以是：交通监测、桥梁负载、智能飞机机翼根据气流和温度调整等。

与手动城市导航中心相比，物联网可以提高市民和游客的生活质量，指示最佳路线，提供导游，实现汽车共享自动化，实现自动驾驶汽车，还可以应用于野生动物监测、鱼类监测、鸟类监测等。

物联网在各种不同环境中的全球范围提供了机会，可以促进我们的日常生活，节约环境，更好地监控和实施法律和秩序（警察应用）。 列表是无穷无尽的。

关键因素是能够将传感器和设备嵌入到重要和相关的位置，以监测各种现象并将其连接到网络以监测收集的数据。 是什么决定了对物联网的投资？ 为公民、游客、运动员、工业企业、交通系统、警察和安全需求等提供“微服务”都可以带来收入。 在实施和推出之前，需要解决法律和政府障碍（如果有的话）。

1.1.13 大数据分析和物联网

物联网中连接设备的显著和实质性增加将导致企业预计需要管理、分析和采取行动的数据规模呈指数级增长。 因此，物联网成为大数据的自然合作伙伴，因为它提供了大数据分析所需的数据量。

正如霍华德·鲍德温所说：“我们将从各个方向获得数据，包括家电、机械设备、火车轨道、船运集装箱、发电站等等。”所有这些实时数据都需要进行处理和分析，以感知可行动的信号。

物联网仍处于初级阶段。 很快，数据将开始从传感器和设备中流动。 可行动的洞察可以是识别客户的购买习惯或机器性能的效率。 例如，LexisNexis拥有一个开源的HPCC大数据平台，通过集成机器学习和商业智能，实现数据集成、处理和结果交付的解决方案。

1.1.13.1 大数据与物联网集成的基础设施

将大数据与物联网集成取决于基础设施环境。
这包括存储和云基础设施。 许多组织试图转向平台即服务（PaaS）云来托管和分析庞大的物联网数据，因为自行维护此类存储将非常昂贵。 PaaS云应提供可扩展性、灵活性、合规性和有效的复杂架构，以存储来自物联网设备的云数据。 如果数据是敏感的，可以部署私有云架构。 否则，可以部署公共云服务，如AWS（亚马逊）或Azure（微软）。 许多云服务提供和提供自己的物联网平台，用于实施物联网应用程序。

1.2 物联网驱动经济的技术挑战

首先，物联网设备的特点是存储和计算能力有限。 这导致了大量的云服务出现，为这些智能设备提供必要的支持。

这导致现有网络过载，带宽不足。
网络功能虚拟化（NFV）和软件定义网络（SDN）架构的联合采用被认为是一个有希望的解决方案。
基础设施即服务（IaaS）的云范式也能够通过互联网为远程用户提供虚拟计算资源。然而，这种云服务的性能受到承载这些云服务的数据中心
中各个服务器的性能负载的限制。 为云提供服务的数据中心通常位于大型网络的核心基础设施附近。 但物联网设备和传感器明确地位于农场、道路、工厂和住宅等现场级别。

因此，我们面临着非常大的延迟问题，无法以满意的QOS或QOE水平实时提供服务。这种情况变得更加复杂，因为物联网设备生成的数据实时
体积前所未有，但需要实时分析才能获得其真正的价值。

然而，生成的数据量超过了存储系统和分析应用的承载能力。 云服务（如IaaS）可以通过提供按需和可扩展的存储和处理服务来帮助满足物联网的需求。

然而，如上所述，在实际生活中的健康监测、紧急响应和管理以及其他延迟敏感的应用中，将这些大量数据传输到云端并将结果返回给应用程序所造成的延迟是不现实或令人满意的，因此是不可接受的。

将大量数据发送到云端也是不可取的，因为它会饱和和过载整个网络带宽，影响整体性能，从而对所有其他物联网应用产生不良影响。

在当前趋势如实时医疗基于物联网应用、智慧城市、智慧村庄等实际情况中，已经注意到整体网络流量会受到不利影响。

1.3 雾计算范式作为解决方案

为了有效应对上述挑战，最初提出了边缘计算范式，利用其计算资源进行本地存储、初步数据处理，以减少网络负载或拥塞，并实现本地化数据分析，从而实现快速数据驱动的决策过程。 然而，不幸的是，边缘计算设备的资源非常有限，这将导致资源争用并增加处理延迟。

因此，出现了一种称为雾计算的新范式，它将边缘设备与云资源无缝集成在一起，以克服边缘计算的所有限制。

因此，雾计算通过利用云资源并协调地理分布的边缘设备，避免了边缘资源争用。

1.4 雾计算的定义

雾计算已经以多种方式被定义：

1. 雾计算是云平台从核心到边缘的扩展
   网络Bonomi [12]，Dastjerdi [13]。

   这个定义过于简单，没有注意到雾计算的普遍性以及其增强的网络能力，例如提供托管环境和改进设备之间的交互支持。

2. Rodero-Merino [9]提出了一个不同的自主定义：雾计算是一个场景，其中大量异构（无线和自主）的普适和分散设备相互通信和合作，并与网络一起执行存储和处理任务，而不需要第三方交互来支持基本网络功能或支持在沙盒环境中运行的新服务和应用，用户租用部分设备来托管这些服务，并获得激励。

上述定义未能解决雾与云的关键联系。

然后，我们有了第三个定义，由Yi [15]提出。

3. 雾计算是一种地理分布式计算架构，其资源池由一个或多个边缘设备（包括边缘设备）组成，不仅仅是无缝支持云服务，而是在网络边缘协同提供弹性计算、存储和其他服务，无论是在远程位置还是在大量附近的客户端。

我们可以预期未来会有更多新的定义。

1.5 雾计算的特点

虽然计算、存储和网络的弹性资源对于云和雾都是常见的，但我们可以确定雾计算范式的独特特点，使其成为云的非平凡扩展，如下所示：

1. 位于边缘：作为靠近边缘的位置，雾具有支持实时处理的延迟敏感应用的能力。
2. 位置感知：与广泛分布的云服务不同，雾计算所提供的服务需要具备推导其位置并跟踪最终用户设备位置的能力。

3. 实时交互和服务交付：与基于云的批量处理不同，雾应用程序确保实时服务交付，在云服务中由于延迟和网络过载而无法实现。

4. 边缘分析：与集中式分析相对，雾计算可以支持在本地分析敏感数据（而不是将数据全部发送到云端进行分析）。

5. 可扩展性：云可能无法处理所有实时数据。. 雾计算解决了云端过载无法实现实时分析的问题。 并非所有分析都必须在云端进行（这要求所有边缘的数据都要到达云端）。 在边缘本身进行预处理并仅发送最重要和相关的信息到云端通常更合适和足够。

总结一下，雾计算是一种分布式计算范式，为网络边缘提供类似云服务的功能。

雾计算利用云和边缘资源以及自身资源。

本质上，雾计算技术通过利用靠近最终用户的客户端或边缘设备在本地处理物联网数据，执行大量的存储、通信、中央、配置和管理。 雾的中介方法从其与传感器等边缘设备的紧密接触中获益，同时利用云服务（作为IaaS）提供的按需可扩展性。

雾计算还涉及在分布式云和边缘设备上运行的数据处理或分析应用程序的组件。 因此，雾计算促进了数据中心与终端或边缘设备之间的计算、网络和存储服务的管理和编程。 此外，雾计算本身的特性支持用户的移动性、资源和接口的异构性，以及分布式数据分析，以满足广泛分布的边缘应用的实时响应和极低延迟的要求和期望。

1.6 雾计算的架构

Cloudlet、IoX和Paradrop是早期的雾计算架构（尽管雾计算当时还没有完全出现）。

*1.6.1 Cloudlet*架构*[11]*

如图1.1所示，Cloudlet是一个资源丰富的雾节点的实现示例。

如上所示，最底层是操作系统和从云端下载的数据。 下一个更高的层是虚拟化层。 最高层是虚拟机，可以托管多个应用程序。

*1.6.2 IoX*架构

IoX是基于思科路由器的架构。 应用程序托管在操作系统和虚拟机监视器上。 该平台可以启用脚本和代码开发。

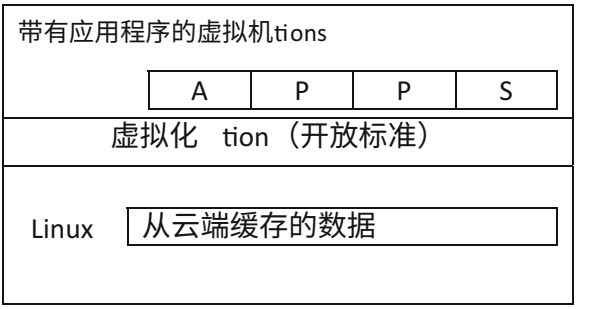


图1.1 Cloudlet

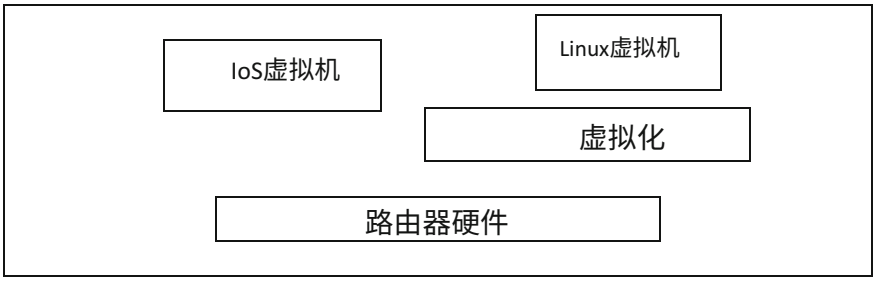


图**1.2** IoX架构

新的（自己的）操作系统也可以安装在昂贵的系统上，关闭对公众的访问，工作在（图1.2）上。

1.6.3 本地网格的雾计算平台

它是嵌入式软件或安装在网络设备和传感器上的。 它标准化并保护了所有供应商的所有设备之间的通信，从而最大程度地减少了定制和服务成本。 本地网格平台位于边缘和云之间的设备上，并提供可靠的设备间通信，无需经过云端。 这使得应用程序能够在边缘进行实时决策，而无需处理与云端通信的高延迟。 此外，所有本地网格设备都可以通过开放的通信标准与云端通信，实现了雾的概念作为云的扩展。

在本地网格平台上运行的应用程序可以利用雾和云之间的相互作用来解决更复杂的问题。 本地网格的雾计算平台与本地网格的虚拟远程终端单元（vRTU）一起打包并提供，vRTU将边缘设备之间的通信转换为兼容的开放标准。 vRTU可以安装在现成的产品上，也可以安装在OEM的定制解决方案上，赋予设备RTU功能，并提供一个管理所有边缘设备的单一点，从而降低定制和维护成本。

1.6.4 ParStream

ParStream是一个实时物联网分析平台。 思科与ParStream合作，在雾计算上构建了一个快速、可靠和高度可扩展的基础设施，用于分析。ParStream提供的物联网大数据分析平台取决于其拥有的专利技术

数据库技术，ParStream DB，是基于列的内存数据库，具有高度并行和容错架构，使用了专利的索引和压缩算法。 作为内存数据库，它非常适合具有有限磁盘空间的雾设备。 任何查询都可以在边缘执行，并以分布式方式进行分析。 它可以部署在支持雾计算的设备上，如思科IoX。

1.6.5 ParaDrop

这适用于Wi-Fi接入点或机顶盒，靠近边缘的雾节点。 这
产品适用于家庭使用，作为网关服务的入口点。

1.6.6 棱镜漩涡

这是一个为物联网而设计的普适数据共享平台。它提供可扩展的端到端、无缝、高效、安全和及时的数据共享访问物联网支持的边缘、网关和云。

Vortex利用DDS 2.0标准进行互操作数据共享，并扩展其以支持互联网规模、移动性和Web 2.0应用。 Vortex与常见的物联网消息传递协议（如MQTT和COAP）无缝交互。Vortex还提供细粒度访问控制支持和对称和非对
称身份验证。 每个物联网设备都连接到执行所有Vortex软件模块的Vortex边缘设备，其中每个平台执行实现全球共享DDS所必需的功能。

一个具有连接到其的物联网设备的Vortex边缘设备在此上下文中形成一个域（DDS实体），称为雾域。
配备这些设备，Vortex支持多种部署模型：

1. 雾+云：雾域内的物联网设备以点对点的方式相互通信。. 在这样的两个或多个雾域之间，它们必须通过云进行通信。

2. 雾+云-链路+云：与前一模型类似，同一雾域内的设备进行点对点通信，而不在同一雾域中的设备通过云使用处理相关安全问题的云链路交换数据，并控制所公开的信息。

3. 联合雾：每个雾域都有一个在V**orte**x设备上运行的V**orte**x云链接。. 联合雾是由云链接实例联合的一组雾域，它们控制着雾域之间的信息交换。

1.7 设计一个强大的雾计算平台

强大的雾计算平台的设计目标如下：

1. 延迟：所有雾应用都应具有低延迟。. 它们应确保任务的运行时间非常短，卸载时间低，并帮助进行搜索和快速决策的时间。

2. 效率：在利用资源和能源方面的效率非常重要，比云场景更为必要，因为（a）一些雾节点可能资源有限，计算和存储能力受限，或者（b）雾节点是手持设备（仅由电池操作），如移动设备、可穿戴设备和传感器。

3. 泛化：由于雾节点和客户端可能非常异构，在雾客户端层面上具有类似的顶层抽象应用。. API应该是通用的，以应对现有的API协议（M2M协议、智能家电API和智能车辆API）。

1.8 设计雾计算平台的挑战

在设计雾节点时可以确定以下挑战：

（a）虚拟化技术的选择-这将决定雾节点的效率、速度和灵活性。

（b）延迟-需要将延迟最小化，因为雾应用程序必须实时响应。

以下考虑因素对于降低延迟很重要：

（I）数据聚合：通过使用各种技术来减少数据量是必不可少的。

（II）调度和配置：如果资源受限的雾节点没有及时进行资源调度和配置，将导致延迟。这需要使用优先级和移动模型进行正确的调度。 这需要使用优先级和移动模型进行正确的调度。

(III) 翻滚或故障：在节点移动、翻滚或节点故障的情况下，雾计算将受到影响。 可以部署诸如块指向、重新调度和复制等缓解策略。

(c) 网络管理：对于雾功能，网络管理至关重要。将软件定义网络（SDN）或网络功能虚拟化（NFV）集成到雾计算中将是具有挑战性的。

(d) 安全和隐私：需要在平台的每个层面上部署访问控制和入侵检测系统，并得到适当的支持。

1.9 平台和应用

1.9.1 雾计算平台的组成部分

一般雾计算平台的组成部分可以按照以下方式确定：

(1) **身份验证和授权**
每个尝试进入雾计算环境的用户都需要通过适当的新颖身份验证方案进行身份验证和授权。
这将有助于识别用户访问模式、移动模式和可信的安全设备。

(2) 通过卸载进行管理：有时卸载可以解决许多问题。
(3) 跟踪：为了基于位置提供服务，我们需要跟踪（**a**）邻居，（b）移动用户详细信息和（**c**）物理位置。
(4) 监控：在雾计算平台和云基础设施中，系统监视器成为提供关键参数（如负载、使用情况、功率和其他参数）以支持决策和成本计算的关键组成部分。
(5) 资源管理：分布式资源管理包括与资源发现和分配相关的任务。

(6) 虚拟机调度：由于系统使用情况、工作负载统计数据以及位置和移动详细信息的融合输入，需要设计新的虚拟机调度策略以找到最优解。

1.9.2 应用和案例研究

见图1.3。

1.9.2.1 健康数据管理和医疗保健

雾计算通过存储在本地雾节点（如智能手机或智能车辆）中，帮助患者保护自己的医疗数据的隐私。Cao [16]开发了一种快速的雾计算辅助分布式分析系统，用于监测中风患者的跌倒情况。

基于雾计算的实时跌倒检测系统将跌倒检测任务分配给边缘设备和云端进行处理。 部署了一整套跌倒检测算法，包括加速度测量和时间序列分析方法，以及用于促进跌倒检测过程的滤波技术。

在真实世界的数据条件下，该系统实现了高灵敏度和高特异性，同时具有高效的响应时间和能源消耗。

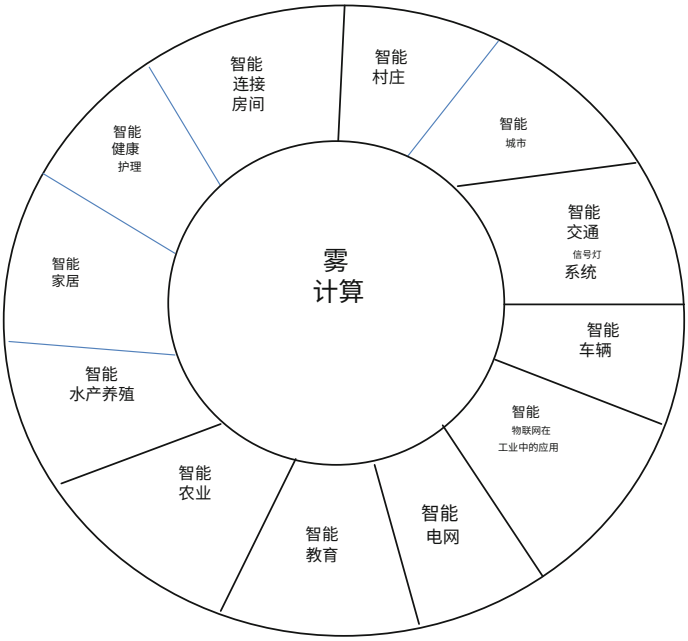


图**1.3** 应用

Stanchev等人提出了一个三层架构的智能医疗保健，包括一个角色模型，分层云架构和一个雾计算层，为医疗保健和老年护理应用提供高效的架构。雾层通过提供低延迟、移动支持、位置感知和安全性来改善架构。 使用业务流程建模符号(BPMN)和面向软件的架构(SOA)，在雾环境中将服务与智能设备进行了标识。 以智能传感器为基础的医疗基础设施被展示为一个使用案例。

1.9.2.2 智能村庄医疗保健

在一个村庄中，只设立了一个初级卫生中心(PHC)，由国家政府运营，有医生和支持人员、诊断中心等，提供基本的诊断和临床设施。 在智能村庄中，通过使用智能健康设备，如智能心电图(ECG)，增强了初级卫生中心的设施，其中一个物联网启用的心电图与本地雾服务器连接，上传当天的心电图数据(最多50个)。

随后，雾服务器将每周的数据上传到云端。 任何时候，心电图上传的数据都可以被心脏病专家访问和检查，如果需要的话。

如果当地医生觉得有必要，可以转诊给心脏病专家进行干预。 这导致了遍布全球的专科和超专科医学专家进行远程医疗干预。

1.9.2.3 智能家居

智能家居是一个可以部署许多物联网功能的家庭。 在智能家居中，可以安装和操作大量物联网设备。 如果部署了多个异构设备，它们的接口和集成将成为一个问题。 雾计算可以通过将所有设备合并到一个平台中来解决这个问题。

让我们考虑以下智能家居安全应用：智能电子锁、视频录制、通过光传感器、占用传感器和运动传感器等传感器进行监控。

除非所有这些传感器都由同一供应商制造，否则集成它们的功能将变得非常困难。 因此，雾计算介入其中。

1.9.2.4 智能车辆和车载雾计算

车载雾计算部署智能车辆，并具有许多可能的应用，例如

(a) 城市交通中的智能交通监控和智能交通信号灯调度，以减轻拥堵和预防交通警察的智能交通管理的一部分。

(b) 通过监测停车场的占用情况并向汽车驾驶员提供指导信息，进行停车设施管理，以及提供停车位的可用性。

(c) 基于基础设施的车载雾计算，例如罗等人的Tube，一种具有路边雾节点的自主车载内容分发。

具有连接车辆的自主车载云[19]可以形成具有云支持的雾节点。

1.9.2.5 增强现实应用

增强现实应用必然需要非常低的延迟，否则用户体验将受到干扰和损害。 因此，雾计算对于增强现实应用具有巨大的潜力。

Zoo等人基于雾计算和链接数据构建了一款增强型脑-计算机交互游戏。当一个人玩游戏时，从大脑中捕获脑电图信号（使用他/她的脑电图传感器进行分析），将其分类为不同的类别，从而帮助识别玩家的脑状态。

脑状态信号的处理和分类是一项需要非常高的计算负载的重要信号处理任务，需要实时进行，这是一项非常困难的任务。

通过同时使用雾服务器和云服务器的组合，可以在雾服务器上进行连续实时的脑状态分类，同时根据传感器收集的数据在云服务器上定期调整分类模型。

Ha等人提出了基于雾谷歌眼镜设备的可穿戴认知架构系统，以帮助具有心智敏锐度的人们。 部署的认知设备资源有限，因此，这个应用程序的计算密集型工作负载需要被卸载到外部服务器上。 这种卸载的响应应该是实时的，否则用户体验将是失败的。

将计算密集型任务卸载到云端会导致长延迟，因此附近的设备被利用起来。这些设备可以与云端通信，用于处理容忍延迟的任务，如错误报告和日志记录。 这种情况是一个典型的雾计算场景，即在边缘进行延迟关键分析，在雾中进行延迟容忍计算，从而将雾视为云的延伸。

第二章
雾应用管理

摘要通常，物联网和雾应用由物联网设备组成，这些设备执行常见活动，如从物联网设备接收数据，预处理接收到的数据，使用驻留在雾服务器上的雾分析工具对接收到的数据进行本地分析，并处理感兴趣的事件以实时响应物联网应用需求。 在本章中，将调查各种应用管理方法。 介绍了基于延迟感知的应用管理、分布式应用部署、分布式数据流方法和资源协调方法。

2.1 引言

通常，物联网和雾计算应用包括物联网设备，这些设备执行一些常见的活动，例如从物联网设备接收数据，对接收到的数据进行预处理，使用雾分析工具（驻留在雾服务器上）对接收到的数据进行本地分析，处理感兴趣的事件以实时响应物联网应用需求[1]。

2.2 应用管理方法

与云相比，雾节点的资源非常有限，并且它们是异构和分布式的。

因此，在雾环境中开发和部署大规模应用程序，需要将应用程序建模为一组轻量级、相互依赖的应用模块[2]。

每个应用模块基本上包含执行物联网应用的典型功能所需的指令；接收数据，处理接收到的数据（来自传感器或物联网设备），并进行分析和/或实时响应到相关应用程序。

© Springer Nature Singapore Pte Ltd. 2019

C. S. R. Prabhu，雾计算、深度学习与大数据分析-研究方向，h*ttps://doi.org/*10.1007/978-981-13-3209-8_2

因此，任何应用模块都包含执行上述步骤或组件所需的指令，每个指令生成相应的特定输出。 然后，根据数据依赖性，输出被发送到另一个模块作为输入。 为了执行，每个模块都需要一定数量的资源，如CPU、内存、带宽等。因此，应用模块以及它们分配的资源构成了不同应用程序的数据处理元素。 为了分布式开发大规模应用程序，上述应用程序的分解将是高效和有效的。

为了减少集中式云端的开销，Vaquero和Rodero-Morino [17]讨论了分布式应用程序开发策略，Hong等人 [18]提出了模型编程平台，以及关于应用程序执行过程中雾节点的协调 [3]。

2.3 性能

在雾生态系统中，应用程序中与性能相关的问题是什么？ 延迟相关的问题，如节点之间的通信、应用程序服务交付截止日期、不同应用程序的数据接收频率，都是问题。 这些参数影响雾应用的服务质量（QoS）、体验质量（QoE）、资源利用率和能源消耗。

2.4 延迟感知的应用程序管理

在雾计算出现之前，已经讨论了云或移动云应用程序中针对延迟敏感应用的各种高效应用程序管理策略[4, 5, 6和7]。

在雾计算出现之后，也正在发展考虑截止日期和各种应用程序的输入接收频率的延迟感知应用模块管理策略，以实现应用程序的截止日期驱动的QOS提供和资源优化，以维持雾中的能源优化[8]。

2.5 雾中的分布式应用程序开发

存在不同的雾应用管理方法，如下所示：

1. ‘Droplets’-将应用程序的最小部分根据需求分布在雾节点上。

当数十亿个地理分布式物联网设备连接到云并开始产生大量的服务请求时，雾计算范式成为云的必要补充。 在雾场景中，当计算元素分布且异构时，[9]提出了一种策略，将应用程序部署在雾节点上，形成‘Droplets’，这是应用程序的最小部分，可以在资源受限的雾节点上分布。

2. 移动雾中的动态节点发现[10]：这种方法基于雾节点的分层定位和分布式应用程序的运行时可扩展性。 移动雾基本上提供了一个API，用于启动动态节点发现过程，将感知数据发送到垂直和水平放置的节点。

2.6 分布式数据流方法

当应用程序分布式，并且执行过程中的应用逻辑可以表示为有向图时，可以根据情况从时间到时间分配雾资源。

2.6.1 延迟感知的雾应用管理

在云环境中，提出了一种延迟感知的迭代算法[4]，用于云应用的共部署。该算法周期性地重复如下步骤：首先临时部署计算实例中的应用程序。 然后，根据应用程序的延迟敏感性，迭代地确定适当的实例。

这个过程会周期性地重复，因为应用程序用户的数量和访问频率可能会随时间的推移而变化。 此外，[5]提出了一种针对移动云计算的延迟感知应用部署。 提出了一种异构资源共享架构。 一个协调器管理所有传入的请求和资源，以优化各种不同应用程序的服务器延迟。

2.7 资源协调方法

资源发现和资源调度通过协调器以最优方式进行协调。存在各种方法和算法来实现这种协调。 在复杂事件处理（CEP）中，基于基于计划的运算符迁移的算法[6]被准备出来，以实现高效的应用迁移。 首先，创建一个时间图模型来识别可能的迁移目标，并选择一个目标实例

基于源的最短路径，然后运行协调算法，在所选实例中容纳迁移操作。

为了实现对通信延迟敏感的虚拟化数据中心中的应用迁移，迁移管理器在迁移时识别[7]虚拟化实例之间的通信模式。 此外，迁移管理器负责根据它们当前的流量对虚拟化实例进行排序，选择一个合适的实例作为迁移目标，并检查其可用性。

提出了一种调度器求解器[10]，用于基于差分频率多切片的虚拟化实例中应用程序的调度。 与其他调度器不同，调度器求解器将CPU切片分成许多微小切片，然后根据微小切片的频率进行调度。 从而增加了应用程序的CPU访问概率。

以上所有方法都是基于云计算的方法，尽管它们的概念和逻辑可以用于雾计算应用程序的调度和管理方法。

第三章
雾分析

摘要雾服务器上的分析能够快速处理传入数据，以实时响应生成数据的物联网设备。 本章探讨了将传统浅层机器学习和深度学习技术应用于大数据分析时涉及的研究问题。 它调查了全球在将传统的无监督和半监督算法以及关联规则挖掘算法扩展到大数据场景方面的研究进展。 此外，它还讨论了将深度学习应用于大数据分析的计算机视觉、语音处理和文本处理等领域的应用，例如语义索引和数据标记。

3.1 引言

虽然物联网对大数据做出了重大贡献，但雾计算架构对物联网至关重要。如果大数据技术完全依赖于云计算，那么很明显，由于大量数据（来自物联网传感器和设备）和网络的高延迟以及云计算所需的互联网带宽，从边缘到云端传输庞大数据并从分析中得出决策再返回边缘将是操作上不可行的。

3.2 雾计算

雾计算是一种新的技术范式，旨在减少实际上传到云端的数据的复杂性、规模和大小。 对传感器和物联网设备输出的原始数据进行预处理是必不可少的，这是一种有效的方式来减少云端大数据的负载。

位于边缘设备附近的雾服务器提供了预处理甚至完成本地分析的可能性，以便为实时本地边缘需求做出快速决策。 只有聚合或摘要数据，大小较小，需要发送到云端。 这将带来从中获益的好处

© Springer Nature Singapore Pte Ltd. 2019

C. S. R. Prabhu，雾计算、深度学习与大数据分析-研究方向，h*ttps://doi.org/*10.1007/978-981-13-3209-8_3

雾计算包括本地快速处理、用于地理可减免和延迟敏感应用的存储，大大减少了网络和互联网上的通信开销，从而大大减少了需要发送到云端的数据的体积和速度。

增强现实、互动游戏和事件监控等应用需要数据流处理，这是与传统大数据应用生态系统中已有的数据库相对应的一种处理场景。

3.3 流数据处理

传感器和物联网设备产生的流数据具有大量连续数据、快速变化和需要快速实时响应的特点，由于其庞大的体积，无法存储数据流。 对整个要存储的数据进行分析可能是不可行的。 流数据的示例包括传感器数据、物联网设备数据、RFID数据、安全监控、电信通话记录、网页日志、网页点击、信用卡交易流程、网络监控、流量数据、股票交易数据等。 数据流的特点是瞬态性，具有连续查询、顺序访问、有限主存、内存数据库管理系统和每秒多GB到达率的实时响应要求。

3.4 流数据分析、大数据分析和雾计算

流数据挖掘和分析以及实时数据挖掘为雾系统的分布式流数据挖掘提供了理论基础，例如特征提取和雾系统的分类。 最近的开源软件产品，如'Tensor flow'，显著促进了在雾服务器甚至移动边缘设备（'移动Tensor flow'）中实现先进的数据挖掘和机器学习算法，如深度神经网络。 尽管有这样的实现，但仍存在未解决的挑战-如何在不影响性能的情况下在多个雾服务器和边缘设备之间进行负载平衡？ 此外，我们还有众所周知的流处理引擎，如Apache Storm和Spark Streaming，可以在雾服务器上执行。因此，我们可以说我们不需要为雾分析开发新工具。 即使如此，仍有一些未解决的问题需要解决：雾流媒体的API。虽然我们有Apache Hadoop生态系统组件，如Apache Mahout用于机器学习和Spark GraphX用于图处理，并且它们可以用于雾流应用程序，但某些特定应用程序的API缺失-微分方程求解器和用于基于雾的实时网络控制应用的控制误差估计器。

3.4.1 机器学习用于大数据、流数据和雾计算生态系统

机器学习技术在涉及流数据、物联网、雾计算和其他大数据场景的大量实际应用场景中非常有用。 例如，随着时间的推移，与健康相关的数据或生物信息学数据不断产生和积累，导致了非常庞大的大数据场景。

应用程序，如3D成像、基因组学、生物识别和其他物联网设备和传感器读数，都推动了数据呈指数增长，成为大数据。进一步的实时数据应用，如快速检测感染、引导正确和适当的治疗（而不是通用药物），都成为可能。 在新生儿护理中，实时数据流正在部署以跟踪感染的重大威胁-这是明显的物联网和雾计算应用。 如果对来自不同专业的大量数据进行分析，医疗保健行业可以实现革命性变革。

为了将传统或增强的机器学习技术或方法应用于分析流数据、物联网数据和大数据，以下是理想的特点：

(1) 可扩展性：部署的机器学习算法应能够处理大量数据，且空间复杂度和存储开销有限。

(2) 高速鲁棒性：机器学习算法应能够实时处理输入数据流，无论数据的数量、速度还是密度，都不会导致性能下降。

(3) 增量性：传统的机器学习方法无法处理随时间动态增长的数据。 对于大数据的目的，机器学习算法应能够成功处理数据随时间不一致的到达，而不会产生额外的成本，并且不会降低质量。

(4) 分布式数据：机器学习方法应能够处理分布式处理，每个节点上的部分数据以及部分处理后，将所有部分结果合并为一个结果。数据可以在大数据场景中分布在全球各地。

特征选择：在流数据或物联网、雾计算和大数据中，维度或特征数量太大。

特征选择的目标是识别最重要的特征并丢弃冗余特征。 通过仅选择最关键和关键的特征，预测分析算法的效率将大大提高。 因此，大维度的影响将显著减少，从而大大提高所涉及的机器学习算法的性能效率，从而加快学习过程，从而提高模型的可解释性。 在传统的机器学习算法中，特征选择是由人类执行的。 特征工程，手动任务是学习过程中非常重要的一部分梯度直方图（HOG）[182]和尺度不变特征

SIFT（尺度不变特征变换）[183]是计算机视觉领域中非常著名的特征工程方法。
如果能够部署自动化工具进行特征工程，将会极大地减少人力工作。

为了从超维大数据情境的数据集中识别出最重要的特征，我们需要采用特征选择和降维技术。
所选特征可以用于在短时间内对大量数据进行即时决策。 对于特征选择来说，大数据的大量、高速和真实性给性能、可靠性、鲁棒性、普适性、非线性、成本和实施复杂性带来了挑战。

以生物信息学为例，我们可以看到‘特征向量’在蛋白质序列分析中是一个关键指标（用于识别具有特定特征的蛋白质序列），具有非常高的维度。 它具有非常多的特征，因此导致了高度复杂性和降低的预测准确性- Bhagyamathi发表了一种新的特征选择方法[42]，而Barbu [43]则通过降维过程准备了一种退火技术。

增量学习方法考虑了随时间逐渐从数据样本中选择的特征子集。 Zeng [44] 提出了一种增量特征选择方法。

3.4.1.1 监督学习

通过使用带有标签的训练示例，在监督学习中对算法进行训练。 基于从可用训练实例中获得的知识，算法预测测试实例的类标签。 监督学习模型可以是回归的连续模型或分类模型。

然而，所有这些都是在大数据场景出现之前的事情。 然而，在大数据分析中，我们需要先进的监督学习方法，如多超平面模型机器(M)分类模型[49]、分而治之的支持向量机[50]和神经网络分类器。 在所有这些方法中，支持向量机被认为是最流行和非常高效的。

为大数据分析引入了修改后的SVM技术。Nei等人[51]提出了一种称为新原始SVM的修改后的SVM，用于大数据分类。
还开发了基于特定应用的技术，如下所示。

3.4.1.2 分布式决策树

梯度提升决策树(GBDT)[56]是为了以分布式方式并行化归纳过程而开发的。将GBDT转换为MapReduce模型很容易，在这种方法中，基于MapReduce的GBDT被用于水平数据分区，因为由于其高通信和I/O开销，HDFS不适用于该算法。

Calaway等人[58]准备了一种可靠的高速决策树(用于大数据)称为rXDTree。 它计算直方图以创建经验分布函数。 决策树以广度优先的方式构建。 这只能在并行计算(多核)环境中执行。 Hall等人[59]提出了一种修改后的决策树算法，它从一组并行构建的决策树中生成规则，并具有可处理的训练集。

但在这种情况下，训练数据集将比通常的要大得多，变得非常复杂，因此上述算法将非常有用。 只有当智能代理提供有关数据有用区域的提示时，才能使用该算法进行部署。 已知标签可用于训练决策树以对未知数据进行分类。

3.4.1.3 大数据的聚类方法

传统的聚类方法不能同时解决之前提到的所有大数据场景中出现的问题。

并行聚类方法正在出现，并似乎是处理大量数据的解决方案。

增量聚类技术可以处理高速大数据场景。
多视角聚类方法被开发用于处理具有多样性的大数据。

DBSCAN、DENCLUE、CLARA、CLARANS和CURE等方法适用于大数据场景。
K-mode和K-prototype方法[90]分别用于大规模分类和混合数据。 Ordonez提出的一种变体[91]可以减少内存需求。
同样，Bradley等人[92]提出了一个框架，通过迭代地从大型数据集中进行采样，在每次迭代中改进模型，最终产生聚类。

在Wave Cluster [93]中，空间域数据被转换为频域数据。

Li等人[94]提出了一种并行处理（SIMD）系统中的分区和链接分层聚类方法。
Zhao等人[95]使用MapReduce架构提出了一种并行k-means算法。 通过找到分布在多个系统上的聚类并合并结果，PDBSCAN [96]提供了一种用于聚类的分布式算法。
P-cluster [97]通过对对象进行分区来实现误差最小化。
PBIRCH [98]是BIRCH的并行版本，适用于共享无关架构，其中传入数据持续分布在多个处理器之间。
Chakraborthy [99]提出了增量k-means算法，用于计算聚类的新中心。

Widyantoro [100]提出了增量分层聚类算法。
IGDCA [101]提出了基于密度的聚类算法。
在无监督多视图学习的场景中，Kailing [102]提出了一种以多视图为范围的聚类方法。

Zeng等人 [103]提出了在不同特征空间中进行分离聚类的方法。

Chaudhuri等人 [104]和Kumar等人 [105]提出了一种多视图聚类方法，将多视图投影到较低维空间。

3.4.1.4 大数据场景下的分布式并行关联规则挖掘技术

顺序关联挖掘技术无法满足维度、大小的可扩展性，也无法讨论数据的地理分布。

因此，在接下来的几年中，开发了几种分布式、并行、高性能的关联规则挖掘方法，如下所示：

Count Distribution [131]中的并行化是针对关联规则挖掘中的传统顺序apriori算法。 该算法通过仅在处理器之间交换计数来最小化通信。

然而，该算法存在一个限制，即在每个处理器上复制整个哈希树，因此不能充分利用总内存。

PDM [132]基于DHP [123]。 FDM [133]是基于计数分布[131]的快速分布式挖掘。 FDM的并行版本称为快速并行挖掘FPM [134]。

然而，需要注意的是，即使将串行方法转换为并行方法，它们会获得速度，但仍保留其固有的缺陷。 为了实现更快的执行，也已经部署了MapReduce框架[135]。

3.4.1.5 动态关联挖掘

在所有关联规则挖掘技术中，固有的假设是正在处理的数据集是静态的。 事实上，这是不正确的。 所有事务性数据库都是动态的。 由于事务引起的数据变化可能使先前进行的关联规则挖掘的结论无效。

快速更新（FUP）[136]在定期更新的数据集中计算大项集。

Borders [137] 算法基于 [138] 中引入的消费者边界集。

在关联规则挖掘中，当前算法无法处理大量的数据。 基因组中一组基因的复杂动态行为以及它对其他基因的表达的影响由GRN表示。已经尝试了多种方法来推断基于稳态时间序列数据的GRN。 然而，它们都无法处理动态时间序列数据 [112]。 迫切需要一种可扩展的GRN重建方法，可以用来推断可靠的GRN。通过比较正常和疾病网络，可以确定潜在的药物靶点 [110]。

3.4.2 深度学习技术

深度学习技术旨在对数据中的高级抽象进行建模。 它们可以使用有监督和/或无监督学习算法来实现这一目的，以学习多个层次的抽象。 这些技术 [246] 是指一类机器学习算法，其在分层架构中进行多个阶段的非线性信息处理，用于模式分类和特征学习，自动地（而不是在传统机器学习中手动进行）。 深度学习还与表示学习相关，其中从较低级别的特征或概念定义了一个层次结构的高级特征。 Deng [246] 提供了一个分类法来创建深度架构，将其分为三类：（1）生成模型，（2）判别模型和（3）混合模型。

根据邓的调查报告，深度学习是指一类具有多个阶段的非线性信息处理的机器学习算法，在层次结构中用于模式分类和特征学习。 在现有文献中，深度学习还与表示学习相关，其中从低层次的特征或概念定义了一个层次结构的高层次特征。 然后，调查报告通过将现有的深度架构和算法分为三类：生成模型、判别模型和混合模型，提供了一个分类法。 生成模型通过观察变量和隐藏变量的高阶相关性特性来描述观察数据和相关类别的联合概率分布。 判别模型通过描述在观察数据条件下类别的后验分布来区分模式。 混合模型是判别模型在很大程度上通过更好的优化或/和正则化来辅助生成模型从数据中学习参数的模型。 深度学习也可以被理解为对浅层架构的扩展，解决了一些受限问题，如广义线性模型、支持向量机、多层感知器、最大熵模型、条件随机场、高斯混合模型和隐马尔可夫模型。 对于深层架构所解决的典型问题，标准统计方法的浅层架构往往会产生棘手的算法和无用的特征学习和类推结果。

邓的调查报告指出，大多数常用的深度学习模型，如自编码器、深度置信网络、卷积神经网络，都是生成模型，通过在无标签训练数据上进行无监督预训练来提取输入特征中的结构和规律，然后通过顶层执行判别任务。 此外，这些生成模型通过引入一堆受限玻尔兹曼机（RBM）来避免全局优化的困难，采用一种贪婪的逐层学习算法在数据上优化参数。 玻尔兹曼机（BM）被定义为一个由对称连接的类似神经元的单元组成的网络，它对是否打开或关闭做出随机决策。RBM是BM的一种特殊情况，它有一层可见单元和另一层隐藏单元，没有可见-可见或隐藏-隐藏的连接。

同一层内的连接。 此外，预训练步骤有助于减轻许多浅层架构训练数百万参数时观察到的过拟合问题。 因此，深度学习模型对于端到端学习复杂系统、嵌入领域知识和解释不确定性非常有用。

深度学习真的安全吗？ 即使是浅层机器学习的例子真的安全吗？

Goodfellow和Shlanes [247]指出，深度学习模型容易受到对抗性示例的攻击。 对抗性示例是通过对交叉验证数据应用小的有意的最坏情况扰动来生成的。 因此，扰动的输入会以高置信度产生错误的输出。 确定此类脆弱性的主要原因是深度学习模型在高维搜索空间中的线性特性。 模型的脆弱性可以从线性变为非线性。

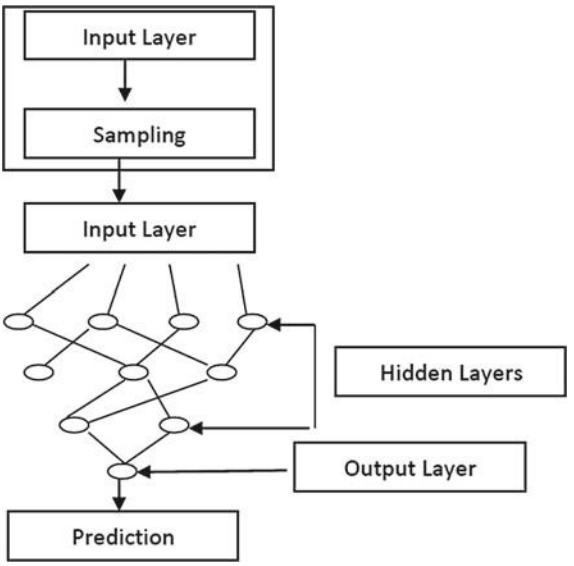
大数据中的大量数据对深度学习构成了严峻挑战。大数据包括大量的输入、大量的输出以及非常高维的特征或属性。 使用单个处理器和其存储器训练深度学习算法是不可行的。 可以部署分布式并行系统框架。 可以部署CPU和GPU集群以提高训练速度，并且算法可以以非常高的准确性进行。 需要部署数据并行和模型并行策略。 另一个挑战是大数据环境中与不同来源相关的数据不完整性和噪声标签。 大多数数据没有标签。

当大量的无监督数据可用于深度学习算法时，可以实现学习，从而以贪婪的方式逐层生成数据表示，直到最高层[187, 188]。 当非线性变换函数作为特征提取器堆叠起来时，通常在深度学习中发生的情况，所得到的结果更好，具有改进的分类建模[189]。 不仅如此，通过使用生成概率模型[190]，我们可以获得更好质量的生成样本。 因此获得的数据表示可以是不变的[191]。 如今的计算机视觉[188,189,197]速度要求部署深度学习算法以获得更高的性能[192–196]。

深度学习模型可以扩展为可扩展的并行算法，以获得更卓越的性能[106]。 部署深度学习算法的主要目标是从输入数据[184, 204, 205]中提取全局抽象或高级抽象。 传统的机器学习算法（决策树、支持向量机等）没有这样的能力。 深度学习技术可以直接从输入的无监督数据中生成高级抽象，无需人为干预。 数据的表示可以是分布式的，可以使用深度学习技术识别出输入原始数据的许多良好和大量的抽象特征配置。 配置数量与提取的特征数量成指数比例。 除了已知的模式，还可以通过新学习的数据模式配置来学习新模式。

深度学习算法在层次化的抽象模式中操作，可以基于较低层次的抽象模式进行识别和连接。 更高层次的抽象有一个好处，就是它们对输入数据的局部变化保持不变。

图**3.1** 深度学习架构



当然，这种情况并不总是发生。 学习不变特征是模式识别中非常重要的目标。 例如，尽管面部朝向不同，学习面部的不变特征对于面部识别至关重要。 通过这种不变表示，可以分解数据中的变化因素。

深度学习架构是具有连续层的深层架构，每个层对输入应用非线性变换，并将输入表示为输出。

在图3.1中，输入数据的分区被处理以识别模式。 在多个层次中，使用中间层进行处理。

在多层次中，深度学习中的数据表示为高效处理大量数据（无法通过传统机器学习技术处理）提供了支持。 数据可以是异构的、非结构化的，并且来自多个来源。 正如已经提到的，深度学习由非线性转换功能层堆叠而成。如果数据经过更多层次的转换，那么构建的非线性转换就会更加复杂。 数据通过它们的转换来表示。 因此，我们可以说深度学习包括能够在具有多层表示的深度架构中学习数据表示的算法。 因此，最终构建的表示仅仅是输入数据的高度非线性函数。 非线性转换提取了输入数据中的潜在解释因素。 线性转换只会导致另一个线性转换。

非线性转换能够更好地学习数据中固有的复杂特征。 如果我们以人脸识别任务为例，深度学习算法将按阶段进行学习：首先在第一层中，只学习不同方向的边缘。 在第二层中，算法将比较这些边缘，以找出更多关于脸部的细节，如嘴唇、鼻子、眼睛、耳朵等。接下来，在第三层中

通过这一层，它可以尝试使用更多的颜色、阴影、形状等细节进行人脸识别。这些最终特征可以作为人脸识别的特征部署。

然而，在实际的深度学习过程中，并不一定总是按照上述预定的顺序进行。 深度学习算法在每一层执行非线性变换。 这些变换将数据中的变化因素分解开来。

如何将上述内容翻译为深度学习算法中的适当训练准则是一个研究问题[185]。

深度结构最初由Hinton [187]在2006年提出，采用逐层贪婪学习的方式进行无监督数据输入。 当输入的感知数据作为第一层的学习数据时，第一层的训练就开始了。 第一层进行学习，并为输入数据识别出一种表示。 这个识别出的表示被用作学习数据输入到下一层。 这个过程会多次重复，使用多个设计好的层数。 最终，在所有层之后，这个深度网络就被训练好了。 整个网络的输出可以用于诸如分类之类的分析任务。 如果任务是分类，那么在其上方会放置另一个监督层。最后，通过将监督数据输入到整个网络中，整个网络就会被训练好。自编码器和受限玻尔兹曼机（RBM）是一起使用的两个构建块，用于构建堆叠自编码器[188, 206]。深度置信网络（DBN）[187]是通过堆叠自编码器和RBM构建的。

自动编码器也被称为自动关联器[207]，包括三层；输入层、隐藏层和输出层。 它们在隐藏层中学习输入的一些表示，以便通过使用隐藏层对输入数据进行重构以产生输出表示（以这种方式可以基于中间表示在输出层中重构输入）。因此，输出本身是输入，并进行传输。通过学习，自动编码器可以通过使用已知的技术（如随机梯度下降）来最小化重构的错误，这与神经网络节点、多层感知器中执行的相同过程非常相似。虽然自动编码器对于线性变换很容易，但也可以执行非线性变换。而具有线性变换的自动编码器只能产生维度降低，具有非线性变换的自动编码器可以从数据中提取有意义的表示。Bengio将它们称为正则化自动编码器[185]。从无监督输入数据中学习的另一种单层学习单元类型是并行限制玻尔兹曼机（RBM）。RBM用作构建深度置信网络的单元输入。RBM包含一个可见层和一个隐藏层。限制是连接仅在不同层之间，并且在同一层的单元之间不允许任何交互。玻尔兹曼机使用称为收缩发散算法[209]的算法工作。一般来说，由于非局部最优解的存在，以及无法处理大量无监督未标记的数据，几个网络表现不佳。

因此，深度置信网络（DBN）[107]被开发出来，以便它可以处理有标签和无标签的数据进行学习。 无监督的预训练需要学习数据分布，而监督的微调用于局部最优搜索。 对于异构的多个输入，如音频和视频数据，Nigam[108]提出了深度学习，它在学习多个抽象和捕捉多个抽象之间的相关性方面非常有效。

深度学习的应用包括：（a）模式识别，（b）计算机视觉，（c）自然语言处理和（d）语音识别。
对于物联网和雾计算，流数据以高速连续生成。 因此，需要实时处理。 增量数据应该由深度学习算法处理，以用于流数据处理。

可以考虑使用强化新网络（RNN）和传统新网络（CNN）进行部署。

3.4.3 深度学习与大数据

在大数据场景中，可以使用深度学习算法处理大规模的无监督或无标签数据，以找到有意义和显著的模式[184, 185, 205]。
通过这种方式从无监督/无标签数据中学习后，可以使用监督/有标签的数据点来训练传统的判别式机器学习模型，这些数据点可能数量较少。 与传统的浅层学习架构相比，深度学习算法被广泛认可具有更好的性能能力，可以发现输入中的全局（非局部）关系或模式[184]。

应用深度学习技术还有其他好处，例如：（1）利用从相对复杂和抽象的数据表示中获得的知识，之后可以使用相对简单的模型；（2）对于图像、语音和文本等异构数据类型，深度学习技术被认为表现更好，并且能够实现自动化特征工程。

通过深度学习技术获得的抽象表示的高层次，可以从相同的数据中获得关系和语义知识。

上述三个特征在大数据分析中非常有用。因此，深度学习是大数据分析的一种非常有效的工具。语义索引和数据标记是其中的例子。 在信息检
索场景[201]中，语义索引在社交网络、计算机视觉、语音、国防和安全系统、欺诈检测、网络流量监控等大数据场景中变得非常重要和有益。

在所有这些系统中，需要对涉及的大量数据进行语义索引，以便信息检索变得简单和快速。
深度学习技术可以用于索引原始数据的目的。
这些技术可以部署，以生成高层次的抽象数据表示。 关于数据的这种表示（或本体等价物）

可以有效地用于索引。 这些表示可能会揭示数据中的实体和关系。 这将实现更高效的数据存储组织，因为具有相似表示的数据项可以被紧密地存储在一起，从而更快、更容易地进行定位和检索。然而，深度学习算法找到的抽象表示应该有意义地显示数据中存在的关联（包括关系和语义）。

如果搜索或查询基于多种自然语言的内容，多语言文本处理会面临挑战。 通过使用标准中间形式（对象知识模型）[245] 来表示知识，可以实现跨语言搜索，而不依赖于任何自然语言，通过使用相应的形态分析进行转换。 中间形式的这一层将位于多语言文本的上一层。 在层次结构的下一级中，可以生成一个知识图，作为中间形式（OKM）中表示的知识框架的等效形式。 在第三个也是最后一个层次中，可以处理生成的知识图。 处理知识图将揭示从多种语言来源中提取的（跨语言）知识的内部结构。 处理给定文本输入中的实体、活动等，可以帮助各种应用，如摘要、不同内容的比较，以及其他文本分析。

完整文本文档的向量表示对应于从中提取的表示，可以极大地方便搜索和检索信息。 换句话说，从中提取的复杂数据表示确实包含有关原始数据的关系和语义性质的信息；它们可以有效地用于语义索引。 如果一个向量可以表示给定的文本片段，那么不同文本之间的比较就变得容易了。 通过比较表示它们的向量，可以找出文本片段之间的相似性。 显然，这比仅仅比较正在处理的文本中的原始数据要好得多。

文档的表示压缩了文档内容或主题的基本独特特征。 词频是几个文档分类和检索系统的关键参数。TFIDF [212] 和 BMZJ [213] 就是这样的系统。 在所有这些系统中，单词被视为维度，独立出现的单词高度相关。

通过应用深度学习技术，对这种高维文本数据（其中所有单词都是维度）进行降维是可能的，从而从中重复出现的较少关键词中识别出文档的语义特征。 通过应用Hinton [214]的生成式深度学习技术，可以学习到文档的二进制代码等效（向量表示）。 对于具有最低层的单词计数的高维数据文档，可以在最高层学习到文档的二进制代码。

一个128位的代码用于表示单个文档。 通过比较它们等效的二进制代码之间的汉明距离，可以评估表示几个这样的文档的代码的语义相似性。语义上相似的

文档在汉明空间中更接近。文档的二进制代码可用于检索文档。 如果提供了一个输入文档，它将被转换为其二进制代码等效，并且可以计算、选择和检索其汉明距离接近的所有具有二进制代码的文档。

二进制代码所需的存储空间很小。 可以使用快速位计数等技术计算二进制汉明距离。

甚至可以生成更短的二进制代码。 通过强制在学习层次结构的最高层中使用少量变量来实现这一点。 这样的较短二进制代码可以直接用作内存地址，以存储在内存中，并且可以创建一个由相似文档的二进制代码组成的'汉明球'。 因此，通过将所有相似文档的短码或长码存储在给定地址附近，创建了一个内存地址的汉明球。

每个文档都以一段内存中的单词来描述，以这样的方式，使得该内存地址单词周围的小汉明球包含所有其他语义相似的文档。 这种技术被称为语义哈希[215]。 即使文档集大小很大，也可以通过采用这种语义哈希方法来独立于文档数量进行高效的信息检索。

这些技术非常有吸引力，因为所有相似的文档都可以一次性检索，因为所有相似的文档将通过非常少的位与查询文档的等效二进制代码不同，即汉明距离很小。 在大数据存储和搜索中，使用局部敏感哈希（LSH）。 但是，语义哈希将比局部敏感哈希更高效和更快。为了获得更好的性能，我们可以使用半监督数据对深度学习模型进行训练。 有一项使用有监督和无监督数据训练模型的研究可供参考[216]。 通过使用标记的有监督数据进行训练，模型将具有先前的知识，从而在学习无监督数据时表现更好。

已经证明[216]，与传统的浅层机器学习算法相比，深度学习算法在学习紧凑表示方面表现更好。
紧凑的表示更好，因为它们需要更少的空间并且表现更好。

谷歌的‘Word2vec’产品能够生成与输入文档等效的向量。 ‘Word2vec’可以处理大规模文本语料库，并生成单词向量作为输出。 在训练过程中，将文本数据作为输入，Word2vec从输入中构建词汇表。 学习单词的向量表示。

然后，可以将单词向量文件用作自然语言处理（NLP）和机器翻译中的特征。 已经开发并提出了适用于大型或非常大型数据集（高达16亿个单词）的单词向量学习技术[217]，其词汇表包含数百万个单词。 神经网络被训练用于学习单词的分布式表示。 为了在大型数据集上训练神经网络，这些模型已经在分布式框架‘Dist Belief’[218]上实现。 从这样大规模的文本中学到的单词向量还可以找到微观级别的语义关系。 因此，使用‘Dist Belief’[219]也可以实现机器翻译。

深度学习可以用于学习文本中词出现的复杂非线性表示，从而实现对输入文本的高级语义特征的学习，这是线性模型无法实现的任务。 这样的活动需要大量的输入数据。 从这样的输入中产生一个带有标签的输出数据是一项困难的任务。 但是，如果使用深度学习，可以通过使用有限的监督标记数据进行训练来处理大量的未标记输入数据进行学习。 上述提取的数据表示对于有效的文档检索和搜索非常有用。

这些技术可以扩展到非文本大数据场景，如图像处理、计算机视觉和速度处理，用于对这些异构数据类型进行语义索引。 微软、谷歌等主要供应商已经发布了基于深度学习的各种大数据场景服务。 微软研究音视频索引系统（MAVIS）提供基于深度学习的音频、视频文件的快速搜索。

互联网上的数字图像集合、GPS、医学图像、军事数据集合、社交媒体等都非常庞大，以至于难以搜索图像。
因此，谷歌推出了谷歌图像搜索服务，可以根据图像文件名和文档内容搜索互联网，但与内容本身无关[221, 222]。 为了克服这些限制，提供更好的搜索机制，需要对图像进行标记，并从图像中提取语义信息。深度学习技术为从这些图像中推导出高级抽象提供了机会，以根据其内容语义来识别每个图像。 这些抽象可以用于图像标记和注释。 卷积神经网络（CNN）和强化神经网络（RNN）可用于标记、识别和分类各种图像：这种标记和注释对于根据查询进行高效图像检索将非常有用。

Hinton [197]在Image Net计算机视觉竞赛中展示了CNNs，其表现优于现有的图像目标识别技术。 此外，Dean等人[218]在Image Net上展示了进一步的成功。受限玻尔兹曼机（RBM）[187]、自编码器[206]和稀疏编码[221]是从Image Net类型的信息输入中提取特征的其他成功方法。

然而，它们只能检测到基本的低级特征，如边缘和斑点。 Google和斯坦福大学构建了一个非常大的深度神经网络，可以学习非常高级的特征。使用9层自编码器对从互联网随机下载的1000万张图像（200 ×200）进行了训练。
该模型有10亿个连接，并且训练持续了3天。
为此，使用了1000台机器和16000个核心的集群。 它能够作为人脸检测器、猫检测器和人体检测器进行功能。 它能够从Image Net数据集中识别22000个对象类别。 尽管上述实验是由Google在原始的无监督、无标签的数据输入上进行的，

在正常的计算机视觉实践中，只有[225]可以通过使用标记的图像来学习。然而，标记的图像数据是稀缺和罕见的。

Socher等人的进一步研究[226]部署了RNNs，用于预测图像的树状结构。这是第一个在分割和注释（复杂图像）方面取得良好结果的深度学习方法。DNNs能够比场景分类中现有的其他方法更好地预测层次结构、场景图像的树状结构。Kumar等人的研究[227]部署了RNNs，创建了一个有意义的搜索空间，也可以是基于设计的搜索。Le等人的研究[228]展示了深度学习技术如何用于视频数据标记和动作场景识别（使用独立变量分析），并取得了更好的性能结果。 这种方法还表明，如果可以从视频数据中提取特征，同样的方法也可以直接用于其他领域。

正如上面所看到的，深度学习技术在从大量输入的无监督原始数据中学习高级复杂抽象方面非常有用，反过来，可以在大数据场景中应用计算可行和线性模型进行分析，从而获得好处。

大数据面临的主要挑战可以归纳为以下几点：

（一）增量学习。

目前的一个挑战是处理快速移动的增量输入数据流。
如何将深度学习应用于这种需求？ 需要深度学习算法能够从连续的数据流输入中进行学习。

Zhou等人[229]采用去噪自编码器[230]进行增量特征学习。 去噪自编码器是常规自编码器的一种变体，能够从受损的输入中提取特征，使提取的特征具有鲁棒性且不受噪声影响，可用于分类目的。

这是通过部署一个用于去噪的隐藏层来实现的。

Calandra等人[231]展示了自适应深度置信网络来学习在线、非平稳流数据。

（ii）高维数据

现有的深度学习技术在处理高维数据（如图像）时计算效率不高。 这是因为深层神经网络中涉及到的学习过程较慢。

Chen等人[232]采用了边缘化堆叠去噪自编码器（mSDAs）来提高高维数据的性能，而不是常规的堆叠去噪自编码器（SDAs）。

在这种方法中，噪声在SDA的训练过程中被边缘化，因此不需要使用随机梯度下降进行学习。

卷积神经网络（CNNs）也可以扩展到高维数据。 在ImageNet图像数据上，CNNs产生了最先进的结果。 （在256 $\times$256的图像上）[197, 206]。 CNN的架构不要求隐藏层中的神经元与前一层的所有节点连接，而只需与处于相同空间区域的节点连接。

随着数据在网络中向更高层移动，图像数据的分辨率也会降低。 这两个因素共同促成了CNN的高性能。上述方法还不足够。

我们需要新的方法来提高深度学习技术处理高维数据的性能。

（三）大规模模型

如何将上述章节中描述的深度学习的成功扩展到计算和海量数据集的大规模模型中?

实证结果表明，具有大量参数的大规模模型[223-235]具有很高的有效性，能够提取非常复杂的特征和表示[218, 236]。

Dean等人[218]提出了使用数十万个CPU核心训练具有数十亿个参数的深度学习神经网络的可能性。
我们上面提到了DistBelief，它可以使用数万个CPU核心来训练神经网络，以实现大规模模型的良好性能。 所采用的通信基础设施利用了模型并行性的两种模式（在机器内部的多线程和机器之间的消息传递）。

值得注意的是，DistBelief使用的非常昂贵的计算资源一般情况下对其他用户来说是不可能的。

Coates等人[236]部署了一个更便宜的GPU服务器集群，还使用了高速通信网络的商用高性能计算技术来协调分布式计算。 这个系统能够在短短几天内训练10亿参数的网络，并且能够使用16台机器处理110亿参数。 因此，这个系统对于希望探索大规模系统的每个人来说都是可负担的。

另外，在分布式雾节点的雾生态系统中，每个节点接收大量的物联网数据，部署深度学习技术会出现一个有趣的场景。

我们期待进一步的创新，以找到更好的大规模数据解决方案。

3.5 雾分析的不同方法

在接下来的章节中，我们将介绍雾分析的不同方法的调查。

A. ‘**智能数据** ’

‘智能数据’是由物联网设备和传感器生成的封装结构化数据的集合，包括一组元数据和一个附带的虚拟机（VM）-用于雾分析的提议。

B. 雾引擎

雾引擎是一个端到端解决方案，提供本地数据分析和通信能力，可以相互通信，也可以与云端通信。雾引擎被称为可定制、敏捷和异构的平台，集成到物联网设备中。 它使得数据可以在云端和网络边缘连接的分布式物联网设备网格中进行处理。 一个雾引擎可以与附近的其他雾引擎合作，从而在云端下方创建一个点对点网络。它提供了一个将数据卸载和与云端交互的功能。 流数据在雾引擎中进行本地分析，而多个雾引擎的数据被收集并传输到云端进行离线全局数据分析，根据需要。 可以确定多种雾引擎部署场景，取决于多个接收器、多个或单个分析器、多个或单个发射器。 雾引擎部署可以部分承担网络骨干和实用程序侧的数据分析负担，并减少对云端的依赖。当计算在本地完成时，只有经过雾引擎清理和分析的一小部分数据被传输到云端，从而大大减少了通过网络传输的数据量，从而大大减少了由于延迟引起的网络拥塞和延迟。

C. 其他产品

雾分析中的其他产品包括微软Azure Stack和Intlock的Cardio logAnalytics，它提供本地数据分析。Oracle提供基于需求的本地Oracle基础设施即服务（IaaS），使客户能够在自己的数据中心部署基于Oracle的系统。 IBM的本地数字分析是其数字分析加速器解决方案的核心网络分析软件组件。

D. ParStream

Cisco的ParStream能够在数据加载时立即进行实时数据分析。 ParStream具有许多吸引人的特点，包括：高度可扩展的分布式混合数据库架构，能够分析数十亿条记录；具有专利的索引和压缩功能，能够实时处理数据并最大程度地减少性能降低；使用标准的CPU和GPU执行查询；与R语言和其他机器学习引擎集成，支持高级分析；使用时间序列分析来分析具有大量历史数据的流数据；使用警报和操作来监视数据流，创建用户友好的程序生成警报，发送通知和执行操作；通过应用统计函数和分析模型从大量数据中推导模型和假设，使用先进的分析技术。

3.6 比较

以上所有产品都有各自的优点和缺点。 尽管它们都提供本地数据分析服务，但它们在提供基于雾概念的整体方法方面存在不足，雾概念是边缘和云之间的中间层。

3.7 边缘分析的云解决方案

亚马逊等云服务提供商（CSPs）提供的解决方案可用-亚马逊的AWA IOT通过HTTP、Web套接字、MQTT实现数据收集，并与云中的设备网关集成。亚马逊Quick Sight可用于机器学习目的。

微软提供使用HTTP、AMQP、MQTT和自定义协议进行数据收集的Azure IoT Hub；提供REST API集成；提供流分析和机器学习，使用Azure IoT网关（本地网关）。

IBM提供使用HTTP和MQTT进行数据收集的IBM Watson IOT，与REST和实时API集成。数据分析通过IBM的Bluemix数据分析平台提供。

谷歌提供了谷歌物联网，仅使用HTTP进行数据收集，与REST API和RPC集成：分析通过云数据流、大查询数据实验室和数据处理使用通用网关（本地）到云。阿里巴巴提供了阿里云物联网，使用HTTP，与REST API

集成，使用自己的分析解决方案，最大计算，并使用云网关到云。在本节中，介绍了雾分析的方法、技术和产品的调查。

第四章
雾安全与隐私

摘要雾中的安全和隐私问题与云中的问题有很大不同。 在雾服务器中，数据比物联网设备本身更好地受到保护。
本章深入探讨了雾安全和隐私的三个层面的挑战：（a）物联网设备级别的身份验证，（b）物联网设备与雾服务器之间的数据加密，以及（c）中间人攻击。 此外，还介绍了可能的故障恢复和备份机制。

4.1 引言

物联网设备形成了一个雾。因此，雾中的安全和隐私问题与云的问题有很大的不同。如果数据存储在雾服务器中，比起存储在物联网设备中，数据会得到更好的保护。 雾服务的提供者也会根据雾的实施方式而有所不同：互联网服务提供商（ISP）或无线运营商控制网关或基站，他们可以利用现有的基础设施建立雾。 希望将云扩展到边缘的云服务提供商也可以建立雾基础设施。 本地私有云所有者可以通过租用闲置资源将他们的云转换为雾。因此，雾服务提供商需要建立一个信任模型。物联网设备从现场接收数据并将数据发送到雾服务器进行处理。 在这个过程中，只有在有处理请求或仅用于存储时，雾服务器才会接收数据。 只有这些活动才会被视为雾环境的一部分。

但是雾节点之间的交互只会在它们需要管理网络本身或网络中的资源时发生。 因此，为了保护网络中的所有通信，(a) IoT设备与其雾服务器之间的通信需要得到保护，同时(b) 雾节点或服务器之间的通信也需要得到保护。

保护此类通信面临以下挑战：

(A) 对称密钥加密技术无法部署，因为IoT设备可能不知道雾网络的存在。

© Springer Nature Singapore Pte Ltd. 2019

C. S. R. Prabhu，雾计算、深度学习与大数据分析-研究方向，https://doi.org/10.1007/978-981-13-3209-8_4

(B) 在雾中部署非对称密钥加密也面临着独特的挑战，这对于IoT/雾环境来说是不可行的：在这种环境中维护公钥基础设施(PKI)几乎是不可行的。 此外，还需要解决其他挑战，例如在IoT设备通信的环境中，需要尽量减少消息开销。

就涉及到多个雾节点或服务器之间的互联而言，需要使用适当的加密技术来确保安全，因为涉及的多跳路径可能不可信赖。

4.2 认证

在雾节点的各个层次中，认证是必不可少的。如前所述，公钥基础设施（PKI）可能不可行。 近场通信硬币（NFC）可以在雾计算中有效地简化认证过程。 基于生物特征的认证（例如印度的Aadhaar卡）也可以在雾计算生态系统中有效地使用。 在网关的各个级别或设备本身的级别上，认证变得重要。 每个设备，如智能电网中的仪表或任何雾环境中的iPad，都应具有基于生物特征的认证或其他认证方式，以防止滥用、操纵或冒充。 例如，智能仪表可以加密数据并发送到雾设备，如家庭区域网络（HAN），在那里数据可以解密，结果可以聚合，然后将其传递给云端。 由于环境限制，无法部署或扩展其他认证机制，如PKI。作为替代，可以考虑使用3GPP、GBA、OMA、M2M协议。 建议部署GBA、OMA、M2M协议。

4.3 隐私问题

需要分析与设备相关的隐私问题，例如使用了哪个设备、何时使用、用于什么目的等等。 加密可以用来提供加密的结果，这些结果无法被各个设备解密。

4.4 用户行为建模

在雾计算中，可以通过用户行为建模来防止内部数据盗窃。通过将当前用户行为与标准用户行为进行比较，可以发现任何异常行为。

(C) 例如，当一个发生故障或已被入侵的物联网设备在雾服务器上不断请求处理和存储时，其他合法的物联网设备将被阻塞。 此时可以发动拒绝服务（DOS）攻击。

如果此攻击同时从多个节点发起，攻击的强度将非常高且倍增。

或者，可以欺骗多个设备发送虚假的处理和存储请求。 存在许多防御策略，但由于雾网络生态系统的完全开放性，它们可能不适用于雾计算。 规模是另一个问题，因为物联网设备的数量可能非常庞大，跨越许多雾节点的网络，而这些设备可能没有通过所有雾节点的身份验证。 相反，身份验证可以基于或依赖于第三方认证机构对设备进行认证。 但是，这种第三方认证只能确保验证请求是否由一个合法的节点发起。 即使如此，被入侵的节点也无法被检测到，因为它可能是一个合法的、经过第三方认证的设备。 地址欺骗很容易，因为这些地址实际上是无限的。

4.5 内部人员数据盗窃

在云系统中，以前出现了许多内部人员盗窃案例。 最终用户将不得不无条件地信任云服务提供商。 如果这种信任被误放或被内部人员盗窃破坏，最终用户将无能为力。

在云计算中，通过使用用户行为分析来防止恶意攻击。 在雾计算中，设计问题是：在哪个位置放置诱饵，以及如何设计诱饵信息以便按需提供并减少被窃取的数据量？

侧信道攻击：攻击者可以通过将恶意虚拟机靠近目标云服务器并利用侧信道攻击来试图破坏云。

4.6 中间人攻击

雾计算是脆弱的，中间人攻击是其脆弱性的一个例子。
在这种类型的攻击中，攻击者将自己置于通信网络中的两个方之间。 在这种情况下，作为雾设备的网关可能被入侵并替换为提供虚假或恶意访问路径的网关，这些路径提供欺骗性的SSID作为公共、合法的SSID。 攻击者可以随时将自己置于通信路径中，拦截和修改通过该路径传递的消息。 因此，攻击者可以控制网关，从而私人通信将被控制。

被劫持。 在雾计算中，中间人攻击可能非常隐蔽。 保护雾设备免受中间人攻击非常困难。用户可能能够从攻击中获取信息，但必须在阅读之前解密该信息。

多租户

在雾生态系统中的多用户（多租户）环境中会出现许多复杂问题。 身份管理、监控、性能、可扩展性、安全等问题。

基于管理员和租户的分析，应在雾计算环境中实施基于终端用户角色或身份的多因素认证机制。 安全可靠的网络平台可以为拓扑结构的微调、带宽分配和流量管理策略提供可编程环境。

4.7 失败恢复和备份机制

与任何备份和恢复系统一样，在雾环境中我们也需要提供可靠的数据备份和恢复机制。 现场数据应定期复制和镜像到离线存储。

根据具体应用需求，雾平台将具有高数据吞吐频率和相对较小的数据存储需求。
为了备份和恢复过程的有效性，我们需要专注于开发数据选择、映射、测试和可访问性角色的策略和策略。

在灾难情况下，雾系统中的数据可能会完全丢失。
因此，需要实施一种有效的备份和恢复策略计划，不能失败。 备份之前应该进行复制。 这将减少备份和恢复过程中的成本和资源消耗。

实践中还有许多其他可用于确保一致性、协调性和性能的替代机制，包括：

(a) 光纤通道，
(b) 高密度分布和耙技术（HSDRT），
(c) 奇偶云服务技术（PCS），
(d) 基于分类法的高效路由（ERGOT），
(e) 冷热备份服务替换策略（CBSRS），
(f) 共享备份路由器资源 (SBRR)。

在移动和无线雾生态系统的情况下，需要基于移动和现场安排的后端恢复系统。 或者，应提供基于大带宽的离线备份和恢复机制。

第5章
研究方向

摘要在本章中，针对物联网的雾计算领域，确定了大约25个研究问题，并提出了明确的研究方向。利用物联网数据的时间维度，为物联网数据添加语义，物联网数据的语义Web，互操作性和标准，功耗管理，数据管理问题，数据溯源问题，数据治理，应用资源管理和迁移问题，雾计算的编程语言，雾生态系统和移动雾生态系统的模拟。 此外，本章讨论了将深度学习技术应用于大数据分析中的开放研究方向，例如降维、改进数据抽象的公式以及可能的解决方案方向。

以下是物联网和雾计算领域的新兴研究方向。

5.1 利用物联网数据的时间维度用于客户关系管理（CRM）

物联网数据的时间视角可以用来帮助解决连续的选择，通过提供令人惊叹的客户体验，提高服务质量和体验质量。 例如，一个供应商协会可以利用每天购物场所占用信息的可用客户信息，向他们的客户提供激励措施或主动管理他们的库存，每天或季节性地。 这种模式变成了一种主动的模式，以影响关系来获得洞察力，发现新的关系，并很好地处理与客户的现有或旧有关系（CRM）。

机器学习和深度学习中的回归技术等分析解决方案可以利用物联网数据的时间维度。

© Springer Nature Singapore Pte Ltd. 2019

C. S. R. Prabhu，雾计算、深度学习与大数据分析-研究方向，https://doi.org/10.1007/978-981-13-3209-8_5

5.2 为物联网数据添加语义

通过为其特定情况和含义添加元数据，可以增强信息的价值。 这些元数据对于帮助客户在边缘设备或雾计算层级上处理和使用异构物联网数据非常重要。

元数据的词汇将被部署以创建本体论。 本体论将有助于将来自不同空间的物联网信息进行组合或合并。
需要努力为连接和共享不同应用和组织领域的信息创建特定的本体论。 本体论可以基于RDF、RDMS和OWL或OKM来集成URI。

5.3 迈向物联网的语义Web

一旦本体论在所有不同的应用和领域中得到明确定义，就可以将物联网设备和传感器产生的多媒体数据集成到语义Web中。

5.4 多样性、互操作性和物联网标准化

物联网世界观具有异构的传统、标准、阶段和倡议，如CoAP、MQTT、XMPP、STOMP、HTTP和AMQP。它们之间的整合是必不可少的。供应商锁定以及像IBM Watson、Microsoft Azure、GE Predix、Cisco Jasper和PTC Thing Worx（工业物联网）这样的平台。开源物联网解决方案活动，如'things board'。 Io'、'Kaa'、Device Hive在这些方向上是不错的尝试。

5.5 物联网中的数据管理问题

新机制，如数据湖，已经出现来处理从物联网中产生的大量信息。信息湖存储有组织的非结构化信息，没有预先设想这些信息项可能如何被后续使用。

在数据湖中，数据质量、元数据和可靠性等问题变得至关重要。

5.6 数据溯源

信息溯源与信息的真实性和可靠性有关，也与确定每个步骤中的所有者/所有者和数据修改者的可追溯性有关。 鉴于大数据提供了深入的洞察和分析，可能导致某种形式的自主激活，实际上，我们必须确保用于做出这些重要决策的数据项来自真实可靠的来源。 物联网中的信息溯源问题需要解决，除了信息管理问题。

5.7 数据治理和监管

物联网数据最初来自各种来源，如私人住宅、道路、农场、工厂。 它需要受到管理和监督。 信息管理和控制其利用需要有效执行。 物联网设备的所有者应该被赋予权利，以指示他们对其各自物联网设备产生的数据的偏好、意图和指令。 未来的研究应该提出可能的安排和结构，以识别物联网数据的所有者、物联网数据的消费者以及在这两者之间起作用的其他参与者的利益和关切。

5.8 上下文感知的资源和服务供应

基于位置、时间、应用程序的敏感性、社交网络交互上下文、设备电池寿命、网络流量等上下文进行服务供应。

5.9 可持续和可靠的雾计算

雾计算中的可持续架构面临许多问题，如服务器质量保证（QoS）、用户体验质量（QoE）、服务可重用性、能源高效的资源管理等。就可靠性而言，它取决于雾节点的一致性、高性能服务的可用性、安全的交互和容错性。

在这个方向上的研究在实现雾计算的可持续性和更高性能水平方面具有巨大潜力。

5.10 雾节点之间的互操作性

雾节点既可以作为具有通信功能的网络节点，也可以作为计算节点，具体取决于它们所扮演的角色的上下文。 它们还需要具备互操作性。 根据需求进行节点的自组织互操作性是必不可少的。到目前为止，还没有出现真正互操作的雾节点架构。

5.11 应用程序的分布式处理

雾中的节点可以分布并且可能具有异构资源。 虽然在文献中已经提出了几种用于分布式应用程序开发和部署的编程平台，但是延迟为基础的应用程序管理优化、数据流管理、QoS和QoE的保证、面向边缘的亲和性等问题仍然是开放问题。

5.12 雾内的功耗管理

随着活动的雾节点数量的增加（基于需求），雾网络内的功耗管理变得关键。

到目前为止，只有云数据中心的节能技术被调查过。 在雾网络中优化功耗是一个开放的问题。 为了达到这个目的，可能需要将任务从一个节点迁移到另一个节点。 需要研究在雾网络中进行功耗管理的优化策略。

5.13 雾中的多租户支持

在为多个租户分配雾资源方面存在着开放性的研究问题。

5.14 雾的编程语言和标准

从根本上讲，雾被提议将云服务扩展到边缘。
由于雾与云有很大的不同，因此雾的编程语言和标准将与云的不同。 识别雾环境中的新编程语言和标准非常重要。

5.15 雾中的模拟

‘Fog Sim’已经开发出来（在墨尔本大学）用于模拟雾生态系统。 然而，雾中模拟的设计和实现是一个未来的可能性。

5.16 移动雾：研究机会

1. 主动与被动服务迁移

当物联网节点从其关联的雾服务物理上移动时，需要继续保留雾计算的好处。 为此，雾服务必须与移动设备或物联网节点保持拓扑接近。 这意味着移动节点的水平或垂直切换（或从一个站点迁移到另一个站点）是能够引起拓扑距离显著变化的实际事件。 因此，迁移决策（何时何地）应该根据用户的随机移动性主动或被动地进行，考虑到这一事件作为基石。 在这种情况下，需要开展研究以开发相应的算法。

2. 利用上下文信息进行定向服务迁移

哪些参数决定迁移？ 它们是上下文相关的，因情况而异，并且还取决于实际目标，如QoS、带宽等。 考虑上下文作为一个参数将是有益的。 （例如，在医院场景中，医院的雾节点可以将患者在医院区域内的位置考虑为符合条件的目标节点之一）。

3. 为实现移动漫游而进行的雾联盟

当移动用户从一个位置移动到另一个位置时，原始的雾节点不存在，雾联盟方法可能能够提供更好的帮助。 属于联盟域的雾节点将提供更好的性能选择。
联盟是一个技巧，可以为用户提供一个与之前节点所属的同一联盟域更接近的雾节点，但是在研究中尚未解决形成和管理雾联盟所涉及的挑战。 其中一些挑战包括(i)管理雾节点之间的SLA，(ii)分布式或点对点雾联盟的架构，以及(iii)使用哪些技术。

4. 虚拟化和迁移技术

迁移的性能将取决于仔细选择虚拟化和迁移技术。 例如，为了扩展适合此目的的节点集合，需要找到一种适用于尽可能多物理节点的虚拟化技术来托管雾服务。

同样重要的是，需要确定一种迁移技术，以能够最小化迁移的总持续时间和随后的停机时间，并且最小化带宽消耗。 同时，同样重要的是，需要确定一种迁移技术，以能够最小化迁移的总持续时间和随后的停机时间，并且最小化带宽消耗。
所有这些问题都存在研究范围。

5. **与移动网络集成，迈向5G**

移动边缘计算（MEC）是一个标准的雾计算系统，依赖于欧洲电信标准协会（ETSI）。 它被认为是物联网服务和即将到来的5G网络的关键推动因素。 交互研究方向是作为ETSI、MEC系统的一部分准备移动支持解决方案，从而实现移动性和会话管理接口。 例如，长期演进（LTE）核心网络的移动性管理实体（MME）组件提供一些有用的功能，如（1）保留用户位置信息，（2）在初始注册过程中选择适当的网关，（3）管理LTE和2G/3G网络之间的切换，（4）漫游管理。所有这些都需要通过适当的研究来进行优化策略的整合。 以上所有内容都需要通过适当的研究来进行优化策略的整合。

5.17 部署与雾节点集成的深度学习用于雾分析

如果每个雾节点都要处理和分析来自其自身物联网设备的流数据，那么通过将深度学习与每个雾节点集成，雾分析可以更好地进行。 需要研究通过定义和部署新的架构将深度学习与雾节点有效地集成，用于基于雾的深度学习技术。

5.18 深度学习与大数据分析的研究方向

与传统的机器学习和特征工程算法相比，深度学习在解决海量输入数据中的数据分析和学习问题方面具有潜在优势。

更具体地说，它有助于从大量无监督数据中自动提取复杂的数据表示。 这使得它成为大数据分析的有价值工具，该工具涉及从通常是无监督和未分类的非常庞大的原始数据集合中进行数据分析。 深度学习中的分层学习和提取不同层次的复杂数据抽象为大数据分析任务提供了一定程度的简化，特别是用于分析

海量数据、语义索引、数据标记、信息检索和分类和预测等判别性任务。

在讨论文献中的关键作品并提供我们对这些特定主题的见解的背景下，我们可以关注与深度学习和大数据相关的两个重要领域：（1）将深度学习算法和架构应用于大数据分析，以及（2）大数据分析的某些特征和问题如何对适应深度学习算法提出独特挑战。

可以确定以下问题：

(i) 一个重要问题是在使用深度学习算法分析数据时是否利用整个可用的大数据输入语料库。 一般的重点是应用深度学习算法来训练基于部分可用输入语料库的高级数据表示模式，然后利用学习到的模式和剩余的输入语料库来提取数据的抽象和表示。 在这个问题的背景下，一个要探索的问题是通常需要多少输入数据来训练深度学习算法的有用（好的）数据表示，这些表示可以在特定的大数据应用领域中进行泛化。

(ii) 领域适应：在进一步探索上述问题时，我们回忆起大数据分析的多样性特征，它关注的是大数据中输入数据类型和领域的变化。在这里，通过考虑输入数据源（用于训练表示）和目标数据源（用于推广表示）之间的转变，问题变成了大数据分析中深度学习的领域适应问题。领域适应是深度学习研究的一个重要焦点[237, 238]（其中训练数据的分布（用于学习表示）与测试数据的分布（用于部署学习表示）不同）。

Glorot等人[237]证明了深度学习能够以分层学习的方式发现中间数据表示，并且这些表示对不同领域是有意义的，并且可以共享。

在他们的工作中，首先使用堆叠去噪自编码器从不同源域获得的无标签数据学习特征和模式。 随后，支持向量机（SVM）算法利用学习到的特征和模式应用于给定源域的标记数据，从而得到一个优于其他方法的线性分类模型。 这项领域适应研究成功应用于一个包含22个源域的大型工业级数据集。 然而，需要注意的是，他们的研究并没有明确编码数据在源域和目标域之间的分布转变。Chopra等人[238]提出了一个基于神经网络的领域适应深度学习模型，该模型力求通过考虑训练数据和测试数据之间的分布转变的信息来学习无监督数据的有用（用于预测目的）表示。 重点是分层学习

在训练和测试领域之间的插值路径上有多个中间表示。 在目标识别的背景下，他们的研究证明了对其他方法的改进。 上述两项研究提出了一个问题，即如何增加深度学习数据表示和模式的泛化能力，指出在大数据分析中泛化学习模式的能力是一个重要的要求，因为输入域和目标域之间经常存在分布偏移。

（三）语义索引和数据标记：另一个感兴趣的关键领域是探索提取的数据表示为大数据提供有用的语义含义所需的标准。早些时候，我们讨论了一些利用深度学习提取的数据表示进行语义索引的研究。 Bengio等人[185]提出了构成执行判别任务的良好数据表示的一些特征，并指出了关于在深度学习中学习良好数据表示的标准定义的开放问题。与更传统的学习算法相比，其中误分类错误通常被用作模型训练和学习模式的重要标准，为大数据训练深度学习算法定义相应的标准是不合适的，因为大多数大数据分析涉及从大量无监督数据中学习。 虽然在一些大数据领域中有监督数据的可用性可能有所帮助，但在大数据分析中，定义获得良好数据抽象和表示的标准的问题仍然很大程度上未被探索。此外，定义提取良好数据表示所需的标准引出了构成对语义索引和/或数据标记有效的良好数据表示的问题。 在一些大数据领域中，输入语料库包含标记和未标记数据的混合，例如网络安全[239]、欺诈检测[240]和计算机视觉[225]。 在这种情况下，深度学习算法可以采用半监督训练方法，以实现定义良好数据表示学习的目标。 例如，通过从未标记/无监督数据中学习表示和模式，可以利用可用的标记/有监督数据进一步调整和改进特定分析任务的学习表示和模式，包括语义索引或判别建模。 在数据挖掘中的半监督学习的变体中，主动学习方法也可以用于获得改进的数据表示，其中可以使用众包或人工专家的输入来为一些数据样本获取标签，然后可以用于更好地调整和改进学习的数据表示。

(iv) 由于数据来自各种格式和各种来源，需要进行数据集成。 Ngiam等人通过整合音频和视频数据开发了一种新的深度学习应用。 研究表明，(a) 通过未标记的数据学习单一模态表示，(b) 共享表示能够捕捉多个模态之间的相关性。

模态。 提出了一种多模型深度玻尔兹曼机(DBM)，它将实值密集图像数据和稀疏词频的测试数据等非常不同的数据模态融合在一起，以学习统一的表示。 在不同来源提供冲突信息的情况下，如何有效和高效地解决冲突并融合来自不同来源的数据是一个开放的研究问题。 是否可以有效处理扩展的模态，而不是现有深度学习算法的双模态？ 在异构数据中，深度学习特征融合的可能性是在哪个层次上实现的？

(v) 深度学习领域的低成熟度需要进一步的广泛研究。 特别是，我们需要更多的工作来适应与大数据相关的深度学习算法问题，包括高维度、流数据分析、深度学习模型的可扩展性、改进的数据抽象形式、分布式计算、语义索引、数据标记、信息检索、提取良好数据表示和领域适应的标准。 未来的工作应该集中解决大数据中经常出现的这些问题之一或多个问题，从而为深度学习和大数据分析研究做出贡献。

第6章
结论

在本书中，我们全面介绍了物联网和雾计算的各个方面，包括定义、架构、雾应用管理、雾分析、大数据分析中的深度学习、雾安全与隐私以及雾和移动雾环境中的研究方向，以及大数据分析中的深度学习。

© Springer Nature Singapore Pte Ltd. 2019

C. S. R. Prabhu,雾计算、深度学习与大数据分析-研究方向, https://doi.org/10.1007/978-981-13-3209-8_6

参考文献

1. J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, 物联网(IoT)：一个愿景，架构要素和未来方向. Future Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
2. H. Gupta, A. Vahid Dastjerdi, S.K. Ghosh, R. Buyya, iFogSim:用于建模和模拟物联网、边缘和雾计算环境中资源管理技术的工具包. *ArXiv*预印本 *arXiv*: 1606, 02007, 技术报告 *CLOUDS-TR*-2016-2, 云计算和分布式系统实验室, 墨尔本大学 (2016)
3. N.K. Giang, M. Blackstock, R. Lea, V.C.M. Leung, 在雾中开发物联网应用程序：一种分布式数据flow方法，在*2015*年第五届国际物联网会议*(*IOT*)(*IEEE, *2015)*，第*155–162*页
4. Y. Kang, Z. Zheng, M.R. Lyu, 一种云基服务的延迟感知共部署机制，在*2012*年*IEEE*第五届云计算国际会议，第*630*–637页(2012). http://dx.doi.org/10.1109/CLOUD.2012.90
5. T. Nishio, R. Shinkuma, T. Takahashi, N.B. Mandayam, 用于优化移动云服务延迟的面向服务的异构资源共享，在第一届移动云计算网络研讨会*(*MobileCloud *'13)(*ACM, 纽约，纽约，美国，2013)，第19– 26页. http://dx.doi.org/10.1145/2492348.2492354
6. B. Ottenwalder, B. Koldehofe, K. Rothermel, U. Ramachandran, MigCEP: 迁移操作员用于移动驱动的分布式复杂事件处理，在第*7*届*ACM*国际分布式事件系统会议（*DEBS'13*）(ACM,纽约，纽约，美国，20*13*年)，第*183–19*4页。 http://dx.doi.org/*10.11*45/2488222.24882657. Takouna, R. Rojas-Cessa, K. *S*achs, C. Meinel, 面向通信和能源高效调度虚拟化数据中心中的并行应用程序，在20*13*年*I*EEE/ACM第6届实用和云计算国际会议*(*20*13*年*)*，第25*1*–255页。http://dx.doi.org/10.1109/UCC.2013.50
8. Md. Redowan Mahmud, M. Afrin, Md. Abdur Razzaque, M. Mehedi Hassan, A. Alelaiwi,M. Alrubaian, 通过上下文感知的移动应用程序调度在云端基础设施中最大化体验质量。软件实践经验 **46**(11), 1525–1545 (2016). http://dx.doi.org/10.1002/spe.2392
9. L.M. Vaquero, L. Rodero-Merino, 在雾计算中找到你的方向：走向全面的雾计算定义。SIGCOMM计算机通信评论 **44**(5), 27–32 (2014). http://dx.doi.org/10.1145/2677046.2677052

© Springer Nature Singapore Pte Ltd. 2019

C. S. R. Prabhu,雾计算、深度学习和大数据分析-研究方向, https://doi.org/10.1007/978-981-13-3209-8

10. K. Hong, D. Lillethun, U. Ramchandran, B. Ottenwalder, B. Koldehofe, 移动雾: 互联网物联网大规模应用的编程模型, 在第二届*ACM SIGCOMM*移动云计算研讨会上 (*ACM*, 2013), pp. 15–20

11. M. Satyanarayanan, G. Lewis, E. Morris, S. Simanta, J. Boleng, K. Ha, 云雾在恶劣环境中的作用. IEEE Pervasive Comput. **12**(4), 40–49 (2013)12. F. Bonomi, R. Milito, J. Zhu , S. Addepalli, 雾计算及其在物联网中的作用, 在第一届MCC移动云计算研讨会上 (ACM, 2012), pp. 13–16

13. A.V. Dastjerdi, H. Gupta, R. Calheiros, S. Ghosh, R. Buyya, 第4章—雾计算原理、架构和应用, 在物联网: 原理与范例*Morgan Kaufmann*由 R. Buyya, A.V. Dastjerdi (2016), 第61–75页
14. C. Dsouza, G.J. Ahn, M. Taguinod, 基于策略的雾计算安全管理:初步框架和案例研究, 在*2014 IEEE*: 第15届国际信息重用与集成会议 *(IRI)(2014*年8月*)*, 第16–23页

15. S. Yi, C. Li, Q. Li, 雾计算综述: 概念、应用和问题, 在 *2015*年*ACM*移动大数据研讨会论文集, 杭州, 中国 (2015年), 第37–42页
16. Y. Cao, S. Chen, P. Hou, D. Brown, FAST: 一种辅助分布式分析的雾计算系统，用于监测中风的缓解，在*IEEE*国际会议论文集上（*N*AS），美国波士顿（2015年），第2-11页17. L.M. Vaquero, Rodero-Merino L fi在雾中找到你的方向：朝着雾计算的全面定义。ACM S*I*GCOMM Comp. Commun. Rev. **44***(5)*，27–32（2014年）18. K. Hong, D. Lillethun, *U*. Ramachandran, B. Otten Walder, 移动雾：一种用于物联网大规模应用的编程模型，在第二届ACM S*I*GCOMM移动云计算研讨会上（ACM，2013年）19. X. Hou, Y. Li , M. Chen, D. Wu, D. Jin, S. Chen, 车载雾计算：将车辆视为基础设施的观点。*IEEE* Trans. Veh. Technol.（2016年6月）20. D. Cai, X. He, J. Han, Srda：一种用于大规模判别分析的高效算法。

Knowl. Data Eng. IEEE Trans. **20**(1), 1–12 (2008)
21. W. Shi, Y.-F. Guo, C. Jin, X. Xue, 一种改进的广义判别分析方法，用于大规模数据集，在 *ICMLA'08*. 第七届机器学习和应用国际会议，*2008*(*I*EEE, *2008*)，页码769–772

22. B.-H. Park, H. Kargupta,分布式数据挖掘：算法、系统和应用 (2002)
23. O. Rana, D. Walker, M. Li, S. Lynden, M. Ward, Paddmas: 并行和分布式数据挖掘应用套件，在第*14*届国际并行与分布式处理研讨会论文集，*2000. IPDPS 2000*(IEEE, 2000)，页码387–392
24. T. Kraska, A. Talwalkar, J.C. Duchi, R. Griffith, M.J. Franklin, M.I. Jordan, Mlbase: 一个分布式机器学习系统，在 *CIDR* (2013)
25. Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, J.M. Hellerstein, 分布式graphlab: 云中机器学习和数据挖掘的框架. VLDB Endowment **5**(8), 716–727 (2012)
26. J. Dean, S. Ghemawat, Mapreduce: 大规模集群上简化的数据处理, in *OSDI \'04* (2005), pp. 137–150
27. G. Malewicz, M.H. Austern, A.J. Bik, J.C. Dehnert, I. Horn, N. Leiser, G. Czajkowski, Pregel: 大规模图处理系统, in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (ACM, 2010), pp. 135–146
28. K. Shvachko, H. Kuang, S. Radia, R. Chansler, Hadoop分布式 file系统, *2010*年 *IEEE*第*26*届大规模存储系统和技术研讨会*(MSST)*(IEEE, 2010), pp 1–10

29. M. Ovsiannikov, S. Rus, D. Reeves, P. Sutter, S. Rao, J. Kelly, Quantcast file系统. Proc. VLDB Endowment **6**(11), 1092–1101 (2013)
30. S. Owen, R. Anil, T. Dunning, E. Friedman, Mahout实战. Manning (2011)
31. J. Dean, S. Ghemawat, Mapreduce: 简化大规模集群上的数据处理. Commun. ACM **51**(1), 107–113 (2008)
32. J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, G. Fox, Twister: 一个用于迭代式*M*apReduce的运行时在第*19*届*ACM*国际高性能分布式计算研讨会上(ACM, 2010), pp. 810–818
33. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S.Shenker, I. Stoica, 弹性分布式数据集: 一种内存中容错抽象用于第*9*届*USENIX*网络系统设计与实现会议(*USENIX*协会, 2012), p. 2

34. Y. Low, J.E. Gonzalez, A. Kyrola, D. Bickson, C.E. Guestrin, J. Hellerstein, Graphlab: 一个新的并行机器学习框架, arXiv预印本 arXiv:1408.2041 (2014)35. W. Gropp, E. Lusk , N. Doss, A. Skjellum, 一个高性能、可移植的*M*PI消息传递接口标准实现. Parallel Comput. **22**(6), 789–828 (1996)36. C.M. Bishop et al.,模式识别与机器学习, vol. 4, no. 4 (Springer,纽约, 2006)

37. D.K. Bhattacharyya, J.K. Kalita, 网络异常检测 *:* 机器学习视角(CRC出版社, 2013)
38. L. Floridi, 大数据及其认识论挑战. 哲学与技术. **25**(4), 435–437 (2012)
39. S.C. Hoi, J. Wang, P. Zhao, R. Jin, 用于挖掘大数据的在线特征选择" in 第*1*届大数据、流数据和异构源挖掘国际研讨会*:* 算法、系统、编程模型和应用 (ACM, 2012), 第93–100页
40. M. Lopez, G. Still, 半finite规划. 欧洲运筹学杂志. **180**(2), 491–518 (2007)
41. M. Tan, I.W. Tsang, L. Wang, 面向超高维特征选择的大数据方法. J. Mach. Learn. Res. **15**(1), 1371–1429 (2014)
42. M. Bhagyamathi, H.H. Inbarani, 一种新的混合粗糙集和改进的和谐搜索基于蛋白质序列分类的特征选择方法，在复杂系统的大数据中（*S*pringer，2015），pp. 173–204

43. A. Barbu, Y. She, L. Ding, G. Gramajo, 特征选择与模拟退火的大数据学习,"arXiv预印本 arXiv:1310.2880 (2013)
44. A. Zeng, T. Li, D. Liu, J. Zhang, H. Chen, 一种用于混合信息系统增量特征选择的模糊粗糙集方法。模糊集合系统 **258**, 39–60 (2015)45. T.M. Mitchell,机器学习*1997*, vol.
45 (McGraw Hill, Burr Ridge, IL, 1997)
46. R.O. Duda, P.E. Hart, D.G. Stork，模式分类（W*iley*，2012）
47. C.M. Bishop et al.，模式识别与机器学习，第4卷，第4期（Sp*ringer*，纽约，2006）

48. M. Mohri, A. Rostamizadeh, A. Talwalkar,机器学习基础(MIT出版社, 2012).
49. N. Djuric, 大数据可视化和监督学习的算法. 博士论文, 天普大学 (2014)
50. C.-J. Hsieh, S. Si, I.S. Dhillon, 一种用于核支持向量机的分而治之求解器. arXiv预印本 arXiv:1311.0914 (2013)
51. F. Nei, Y. Huang, X. Wang, H. Huang, 具有线性计算成本的新原始SV*M*求解器，用于大数据分类，第*31*届国际机器学习大会(J*ML*R, 2014)，第1–9页

52. S. Haller, S. Badoud, D. Nguyen, V. Garibotto, K. Lovblad, P. Burkhard, 使用支持向量机分析扩散张量成像数据的帕金森病患者个体检测: 初步结果. Am. J. Neuroradiol. **33**(11), 2123–2128 (2012)

53. D. Giveki, H. Salimi, G. Bahmanyar, Y. Khademian, 基于互信息和修改的布谷鸟搜索的特征加权支持向量机自动检测糖尿病诊断。 arXiv预印本 arXiv:1201.2173 (2012)

54. S. Bhatia, P. Prakash, G. Pillai, 基于支持向量机的心脏病分类决策支持系统，使用整数编码的遗传算法选择关键特征，在世界工程与计算机科学大会(*WCECS*)论文集(2008)，pp. 22–2455. Y.-J. Son, H.-G. Kim, *E.*-H. Kim, S. *C*hoi, S.-K. Lee, 应用支持向量机预测心力衰竭患者的用药依从性。Healthc. Inf. Res.**16**(4), 253–259 (2010)

56. J. Ye, J.-H. Chow, J. Chen, Z. Zheng, 随机梯度提升分布式决策树, 在第*18*届*ACM*信息与知识管理会议上 (ACM, 2009), pp. 2061–2064
57. D. Borthakur, Hadoop分布式 file系统: 架构与设计。 Hadoop项目网站 **11**(2007), 21 (2007)
58. R. Calaway, L. Edlefsen, L. Gong, S. Fast, 大数据决策树与R语言。革命
59. L.O. Hall, N. Chawla, K.W. Bowyer, 决策树学习在非常大的数据集上, 在 *1998*年 *IEEE*国际系统、人类和控制论大会上*, 1998*, vol. 3 (IEEE, 1998), pp. 2579–2584
60. C.C. Aggarwal, C.K. Reddy,数据聚类：算法与应用(*CRC*出版社，2013年)

61. P.N. Tan, K. Steinbach, V. Kumar,数据挖掘聚类分析：基本概念和算法 (2006年)
62. Y. Cheng, G.M. Church, 表达数据的双聚类，出现在Ismb，第8卷(2000年)，第93–103页
63. H. Ahmed, P. Mahanta, D. Bhattacharyya, J. Kalita, A. Ghosh, 交叉共表达子立方体挖掘器：一种有效的三聚类算法，出现在*2011*年世界信息与通信技术大会*(WICT)*(IEEE, *2011*年)，第846–85*1*页64. A.K. Jain, M.N. Murty, P.J. Flynn, 数据聚类：一项综述。A*C*M计算机调查*(C*SUR)**31***(3)*，*2*64–323页*(1*999年*)*

65. L. Kaufman, P. Rousseeuw,通过中心点进行聚类(North-Holland, 1987)
66. L. Kaufman, P.J. Rousseeuw, 数据中的群组发现: 聚类分析导论, vol. 344 (Wiley, 2009)
67. R.T. Ng, J. Han, Clarans: 一种用于空间数据挖掘的对象聚类方法. 知识. 数据工程. IEEE Trans. **14**(5), 1003–1016 (2002)
68. P. Berkhin, 数据挖掘技术聚类方法综述, in多维数据分组(Springer, 2006), pp. 25–71

69. S. Guha, R. Rastogi, K. Shim, Cure: 一种用于大型数据库的高效聚类算法, in *ACM SIGMOD Record*, vol. 27, no. 2 (ACM, 1998), pp. 73 –84
70. H.-P. Kriegel, P. Kroger, J. Sander, A. Zimek, 基于密度的聚类，在 *W*iley跨学科评论：数据挖掘和知识发现，卷1，第3期，（2011年），pp. 231–240

71. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, 一种基于密度的算法用于发现大型空间数据库中的聚类与噪声。 Kdd **96**(34)，226–231（1996年）
72. A. Hinneburg, D.A. Keim, 一种在大型多媒体数据库中进行聚类的高效方法带有噪声，在 *KDD*，卷98（1998年），pp. 58–65
73. L.J. Hubert, 图论在聚类中的一些应用。 Psychometrika **39**(3)，283–309（1974年）
74. G. Karypis, E.-H. Han, V. Kumar, Chameleon: 使用动态建模的分层聚类。 计算机 **32**(8), 68–75 (1999)
75. F. Hoppner,模糊聚类分析：分类、数据分析和图像识别的方法 (Wiley, 1999)
76. T. Kohonen, 自组织映射。IEEE会议录 **78**(9), 1464–1480 (1990)
77. A. Ben-Dor, B. Chor, R. Karp, Z. Yakhini, 发现基因表达数据中的局部结构：保序子矩阵问题。计算生物学杂志 **10**, 373–384 (2003)

78. A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L.Thiele, E. Zitzler, 基因表达数据的双聚类方法的系统比较和评估，发表于生物信息学, vol. 22, no. 9, (2006), pp. 1122–112979. Y. Kluger, R. *B*asri, J. Chang, M. Gerstein, 微阵列数据的光谱双聚类：基因和条件的共聚。基因组研究 **13**(4), 703–716 (2003)80. A. Tanay, R. Sharan, M. Kupiec, R. Shamir, 通过整合高度异质的全基因组数据揭示酵母分子网络的模块化和组织。

Proc. Nat. Acad. Sci. U. S. A. **101**(9), 2981–2986 (2004)

81. J. 杨, H. 王, W. 王, P. 于, 增强的表达数据双聚类, 在第三届*I*EEE生物信息学和生物工程研讨会论文集中, (2003),pp. 321–327

82. S. Bergmann, J. Ihmels, N. Barkai, 迭代签名算法用于大规模基因表达数据分析. Phys. Rev. E, **67**, 031 902–031 919 (2003)83. B. Pontes, R. Giraldez, J. Aguilar-Ruiz, 测量双聚类中移位和缩放模式的质量. Pattern Recogn. Bioinf. **6282**, 242–252 (2010)

84. F. Divina, B. Pontes, R. Giraldez, J.S. Aguilar-Ruiz, 一种评估双聚类质量的有效度量。计算机生物学与医学 **42**(2)，245–256 (2011)

85. W.-H. Yang, D.-Q. Dai, H. Yan, 从基因表达数据中找到相关的双聚类。知识. 数据工程. IEEE交易. **23**(4), 568–584 (2011)

86. H. Ahmed, P. Mahanta, D. Bhattacharyya, J. Kalita, 基于移位和缩放相关性的双聚类算法。计算机生物学. 生物信息学. IEEE/ACM交易. **11**(6), 1239–1252 (2014)87. D. Jiang, J. Pei, A. Zhang, “从基因-样本-时间微阵列数据中挖掘一致的基因簇，于第*10*届ACM SI*G*KDD会议论文集*(*KDD*'04)* (2004)88. L. Zhao, M.J. Zaki, 三聚类:一种在*3*D微阵列数据中挖掘一致聚类的有效算法(ACM, 2005), pp. 694–705

89. H. Jiang, S. Zhou, J. Guan, Y. Zheng, gtricluster: 一种更通用和有效的三维聚类算法，用于基因-样本时间微阵列数据，在 *BioDM'06*(2006)，第48–59页90. Z. Huang，对具有分类值的大数据集进行聚类的k-means算法的扩展。数据挖掘知识发现。 **2**(3)，283–304 (1998)

91. C. Ordonez, E. Omiecinski，用于关系数据库的高效基于磁盘的k-means聚类。知识数据工程。IEEE Trans. **16**(8), 909–921 (2004)

92. P.S. Bradley, U.M. Fayyad, C. Reina等，将聚类算法扩展到大型数据库中，在 *KDD*(1998)，第9–15页

93. G. Sheikholeslami, S. Chatterjee, A. Zhang, Wavecluster: 基于小波的空间数据聚类方法，适用于非常大的数据库。VLDB J. **8**(3–4), 289–304 (2000)94. X. Li, Z. Fang, 并行聚类算法。Parallel Comput. **11**(3), 275–290 (1989)

95. W. Zhao, H. Ma, Q. He, 基于MapReduce的并行k-means聚类，在云计算(Springer, 2009)，pp. 674–679

96. X. Xu, J. Jager, H.-P. Kriegel, 一种用于大型空间数据库的快速并行聚类算法，在高性能数据挖掘(Springer, 2002)，pp. 263–29097. *D*. Judd, P.K. *M*cKinley, A.K. Jain, 大规模并行数据聚类，在第*1*3届国际模式识别会议论文集，*1*996, vol. 4 (*I*EEE, *1*996)，pp. 488–493

98. A. Garg, A. Mangla, N. Gupta, V. Bhatnagar, Pbirch: 一个可扩展的并行增量数据聚类算法，在第*10*届国际数据库工程和应用研讨会上，*2006*年。 *IDEAS'06*(*IEEE*, *2006*)，第3*15*–3*16*页99. *S*. Chakraborty, N. Nagwani, 增量k-means聚类算法的分析和研究，在高性能架构和网格计算(*S*pringer, *2011*)，第338–34*1*页

100. D.H. Widyantoro, T.R. Ioerger, J. Yen, 一种构建集群层次的增量方法，在*2002*年I*EEE*国际数据挖掘会议上，*2002*年。*ICDM 2003*(IEEE, 2002)，第705–708页

101. 陈宁，陈爱珍，周立新，一种增量网格密度聚类算法。
J. 软件 **13**(1), 1–7 (2002)
102. 凯林，克里格尔，普里亚金，舒伯特，带噪声的多重表示对象聚类，在知识发现和数据挖掘进展（Springer，2004），pp. 394–403

103. 曾华杰，陈忠，马文勇，一种用于聚类异构网络对象的统一框架，在第三届国际网络信息系统工程会议论文集，*2002*年。W*ISE 2002*(*IEEE*，*2002*)，pp. 161–170

104. Chaudhuri，Kakade，Livescu，Sridharan，通过规范相关分析进行多视图聚类，在第*26*届国际机器学习年会论文集（*A*CM，*2009*），pp. 129–136

105. A. Kumar, H. Daume, 一种多视角谱聚类的协同训练方法，发表于第*28*届国际机器学习大会*(ICML*-11*)(*2011*)*，pp. 393–400

106. G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V.Vanhoucke, P. Nguyen, T.N. Sainath等，深度神经网络在语音识别中的声学建模: 四个研究小组的共同观点。IEEE信号处理杂志 **29**(6)，82–97 (2012)

107. G.E. Hinton, R.R. Salakhutdinov，用神经网络降低数据的维度。 Science **313**(5786)，504–507 (2006)
108. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng，多模态深度学习，发表于第*28*届国际机器学习大会*(IC*ML-11*)(*2011*)*，pp. 689–696

109. M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, 深度学习在大数据分析中的应用和挑战。J. Big Data **2**(1), 1–21 (2015)
110. P.B. Madhamshettiwar, S.R. Maetschke, M.J. Davis, A. Reverter, M.A. Ragan, 基因调控网络推断：评估和应用于卵巢癌的药物靶点优先级。Genome Med. **4**(5), 1–16 (2012)

111. H. Bolouri, 用大数据建模基因组调控网络。Trends Genet. **30**(5), 182–191 (2014)
112. S.A. Thomas, Y. Jin, 重建生物基因调控网络：优化与大数据相遇的地方。Evol. Intell. **7**(1), 29–47 (2014)
113. H. Lee, A. Hsu, J. Sajdak, J. Qin, P. Pavlidis, 人类基因在许多微阵列数据集中的共表达分析。基因组研究。 **14**(6), 1085–1094 (2004)
114. N. Friedman, M. Linial, I. Nachman, D. Pe'er, 使用贝叶斯网络分析表达数据。计算生物学杂志。 **7**(3–4), 601–620 (2000)
115. M. Davidich, S. Bornholdt, 布尔网络模型预测分裂酵母的细胞周期序列。 PLoS One **3**(2), e1672 (2008)
116. J. Faith, B. Hayete, J. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. Collins,T. Gardner, 从表达谱的综合中大规模映射和验证大肠杆菌转录调控。PLoS生物学。 **5**(1), e8 (2007)117. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.
Favera, A. Califano,Aracne: 一种在哺乳动物细胞环境中重建基因调控网络的算法。BMC生物信息学。 **7**(Suppl 1), S7 (2006)

118. P. Meyer, K. Kontos, F. Lafitte, G. Bontempi, 通过信息论推断大型转录调控网络。EURASIP J. Bioinf. Syst. Biol. **2007** (2007)119. S. Roy, D.K. Bhattacharyya, J.K. Kalita,
利用局部表达模式从微阵列数据重建基因共表达网络。BMC Bioinf. **15**(Suppl 7), S10 (2014)
120. R. Agrawal, T. Imielinski, A. Swami, 在大型数据库中挖掘项集之间的关联规则，
ACM SIGM*O*D Record, vol. 22, no. 2 (ACM, 1993), pp. 207–216

121. R. Agrawal, R. Srikant 等人，用于挖掘关联规则的快速算法，*Proceed*ings *of 20th international conference on very large data bases, VLDB* , vol. 1215 (1994), pp. 487–499
122. M. Houtsma, A. Swami, 面向关联规则的集合导向挖掘在关系数据库中, 在第十一届国际数据工程会议论文集中, 1*995* (IEEE,1995), pp. 25–33

123. J.S. Park, M.-S. Chen, P.S. Yu,一种有效的基于哈希的关联规则挖掘算法, vol. 24, no. 2 (ACM, 1995)
124. A. Savasere, E.R. Omiecinski, S.B. Navathe,一种高效的大规模数据库关联规则挖掘算法 (1995)
125. H. Toivonen et al., 用于关联规则的大型数据库采样, 在 *VLDB*, vol. 96 (1996), pp. 134–145
126. S. Brin, R. Motwani, J.D. Ullman, S. Tsur, 动态项集计数和推理规则用于市场篮子数据, 在 *ACM SIGMOD Record*, vol. 26, no. 2 (ACM, 1997), pp. 255–264
127. J. Han, J. Pei, 通过模式增长挖掘频繁模式: 方法和影响. ACM SIGKDD Explor. Newsl. **2**(2), 14–20 (2000)
128. S. Roy, D.K. Bhattacharyya, Opam: 一种高效的一次性关联挖掘技术无需候选生成。 J. Convergence Inf. Technol. **3**(3) (2008)
129. R. Srikant, R. Agrawal, 在大型关系表中挖掘定量关联规则，在*ACM SIGMOD Record*, vol. 25, no. 2 (ACM, 1996), pp. 1–12
130. B.-C. Chien, Z.-L. Lin, T.-P. Hong, 一种用于挖掘模糊定量关联规则的高效聚类算法，在*IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th*, vol. 3 (IEEE, 2001), pp. 1306 –1311
131. R. Agrawal, J.C. Shafer, 并行挖掘关联规则. IEEE Trans. Knowl. Data Eng. **8**(6), 962–969 (1996)
132. J.S. Park, M.-S. Chen, P.S. Yu, 高效并行数据挖掘关联规则, in *Proceedings of the Fourth International Conference on Information and Knowledge Management* (ACM, 1995), pp. 31–36
133. D.W. Cheung, J. Han, V.T. Ng, A.W. Fu, Y. Fu, 一种快速分布式挖掘关联规则的算法, in *Fourth International Conference on Parallel and DistributedInformation Systems, 1996*(I EEE, 1996), pp. 31–42
134. D.W. Cheung, Y. Xiao, 数据倾斜对并行挖掘关联规则的影响, in *Research and Development in Knowledge Discovery and Data Mining* (Springer, 1998), pp. 48–60
135. S. Moens, E. Aksehirli, B. Goethals, 针对大数据的频繁项集挖掘，在*2013*年*IEEE*国际大数据会议(IEEE, 2013)，第111–118页
136. S. Thomas, S. Bodagala, K. Alsabti, S. Ranka, 一种用于大型数据库中关联规则增量更新的高效算法，在 *KDD*(1997)，第263–266页137. Y. Aumann, R. Feldman, O. Lip
shtat, H. Manilla, 边界: 一种用于动态数据库中关联生成的高效算法。智能信息系统杂志 **12** (1)，第61–73页 (1999)138. H. Mannila, H. Toivonen, 水平搜索和知识发现理论边界。数据挖掘与知识发现 **1**(3)，第241–258页 (1997)

139. 张三，吴小，张杰，张超，一种在动态数据库中维护频繁项集的递减算法，在数据仓库和知识发现（Springer，2005），305–314页

140. 思科，思科视觉网络指数：全球移动数据流量预测更新，2014–*2019*（思科公共信息，2015）
141. A. Choudhury，P. B. Nair，A. J. Keane等，用于大规模高斯过程建模的数据并行方法。在*SDM*（*SIAM*，2002），95–111页
142. A.E. Raftery，T. Gneiting，F. Balabdaoui，M. Polakowski，使用贝叶斯模型平均校准预测集合。月度天气评论。 **133**（5），1155–1174页（2005年）

143. R. Wright, Z. Yang, 隐私保护的贝叶斯网络结构计算在分布式异构数据上, 在第十届*ACM SIGKDD*国际知识发现与数据挖掘会议(*ACM*, 2004), pp. 713–718144. J. *G*oecks, *A*. Nekrutenko, J. *T*ay*l*or等, *Gal*axy: 一种支持可访问、可重现和透明的生命科学计算研究的综合方法. *G*enome Bio*l*. **11**(8), R86 (2010)

145. A. Matsunaga, M. Tsugawa, J. Fortes, Cloudblast: 结合分布式资源上的MapReduce和虚拟化的生物信息学应用, 在第四届*IEEE*国际e*S*cience会议, *2008*. e*S*cience *'08*(IEEE, 2008), pp. 222–229146. T.H. Stokes, R.A. Moffitt, J.H. Phan, M.D. Wang, Chip artifact CORRECTion(caCORR*E*CT): 一种用于基因组学和蛋白质组学阵列数据质量保证的生物信息学系统. Ann. Biom. *E*ng. **35**(6), 1068–1080 (2007)

147. J.H. Phan, A.N. Young, M.D. Wang, omniBiomarker: 一个基于知识驱动的生物标志物识别的网络应用。生物医学工程。IEEE Trans。 **60**(12), 3364–3367(2013)

148. M. Liang, F. Zhang, G. Jin, J. Zhu, FastGCN: 一种用于快速基因共表达网络的GPU加速工具。 PloS one **10**(1), e0 116 776–e0 116 776 (2014)149. A.S. Arefin, R. Berretta, P. Moscato, 一种基于GPU的计算基因表达网络特征向量中心性的方法，在第十一届澳大利亚并行与分布式计算研讨会的论文集中，第140卷（澳大利亚计算机学会，2013年），第3–11页

150. D.G. McArt, P. Bankhead, P.D. Dunne, M. Salto-Tellez, P. Hamilton, S.-D. Zhang, cudaMap: 一个用于基因表达连接映射的GPU加速程序。BMC Bioinf. **14**(1), 305 (2013)

151. A. Day, J. Dong, V.A. Funari, B. Harry, S.P. Strom, D.H. Cohn, S.F. Nelson, 通过大规模共表达分析进行疾病基因特征化。 PLoS One **4**(12), e8491 (2009)

152. A. Day, M.R. Carlson, J. Dong, B.D. O　'Connor, S.F. Nelson, Celsius: 一个社区资源，用于Affymetrix微阵列数据。Genome Biol. **8**(6), R112 (2007)

153. P. Langfelder, S. Horvath, WGCNA: 一个用于加权相关网络分析的R包。BMC Bioinf. **9**(1), 559 (2008)

154. C.G. Rivera, R. Vakil, J.S. Bader, NeMo: Cytoscape中的网络模块识别. BMC Bioinf. **11**(Suppl 1), S61 (2010)

155. G.D. Bader, C.W. Hogue, 一种在大型蛋白质相互作用网络中 finding分子复合物的自动方法. BMC Bioinf. **4**(1), 2 (2003)

156. T. Nepusz, H. Yu, A. Paccanaro, 在蛋白质相互作用网络中检测重叠蛋白质复合物. Nat. Methods **9**(5), 471–472 (2012)157. B.P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, T. Ideker, PathBLAST: 一种用于蛋白质相互作用网络对齐的工具. Nucleic Acids Res. **32**(suppl 2), W83–W88(2004)

158. H. Nordberg, K. Bhatia, K. Wang, Z. Wang, BioPig：基于Hadoop的大规模序列数据分析工具包。 生物信息学 **29**(23), 3014–3019 (2013)

159. A. Schumacher, L. Pireddu, M. Niemenmaa, A. Kallio, E. Korpelainen, G. Zanetti, K.Heljanko, SeqPig：用于Hadoop中大规模测序数据集的简单可扩展脚本。生物信息学 **30**(1), 119–120 (2014)

160. B. Langmead, M.C. Schatz, J. Lin, M. Pop, S.L. Salzberg, 使用云计算搜索SNP。 基因组生物学 **10**(11), R134 (2009)

161. B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg等，快速且内存高效的将短DNA序列对齐到人类基因组。基因组生物学。 **10**(3), R25 (2009)162. R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, J. Wang, SNP检测用于大规模全基因组重测序。基因组研究。 **19**(6), 1124–1132 (2009)163. S. Zhao, K. Prenger, L. Smith, Stormbow：用于大规模RNA-Seq研究中的reads映射和表达定量的基于云计算的工具。国际学术研究通知。 **2013** (2013)

164. S.V. Angiuoli, M. Matalka, A. Gussman, K. Galens, M. Vangala, D.R. Riley, C. Arze, J.R.White, O. White, W.F. Fricke, CloVR：一种用于从桌面使用云计算进行自动化和便携式序列分析的虚拟机。BMC Bioinf. **12**(1), 356 (2011)165. S. Zhao, K. Prenger, L. Smith, T. Messina, H. Fan, E. Jaeger, S. Stephens, Rainbow：一种使用云计算进行大规模全基因组测序数据分析的工具。 BMC基因组学 **14**(1), 425 (2013)

166. S. Kurtz, vmatch大规模序列分析软件，出版物类型：计算机程序(2003), pp. 4–12

167. A.C. Zambon, S. Gaj, I. Ho, K. Hanspers, K. Vranizan, C.T. Evelo, B.R. Conklin, A.R.Pico, N. Salomonis, GO-Elite：一种灵活的通路和本体超表达解决方案。 生物信息学 **28**(16), 2209–2210 (2012)

168. M.P. van Iersel, T. Kelder, A.R. Pico, K. Hanspers, S. Coort, B.R. Conklin, C. Evelo, 使用PathVisio展示和探索生物通路。BMC Bioinf. **9**(1), 399 (2008)

169. P. Yang, E. Patrick, S.-X. Tan, D.J. Fazakerley, J. Burchfield, C. Gribben, M.J. Prior, D.E.James, Y.H. Yang, 大规模蛋白质组学数据的方向通路分析揭示了胰岛素作用通路的新特征。 生物信息学 **30**(6), 808–814 (2014)170. P. Grosu, J.P. Townsend, D.L. Hartl, D. Cavalieri, Pathway processor: 一种将整个基因组表达结果整合到代谢网络中的工具。Genome Res. **12**(7), 1121–1126(2002)

171. Y.S. Park, M. Schmidt, E.R. Martin, M.A. Pericak-Vance, R.-H. Chung, Pathway-PDT: 一个灵活的家庭核心分析工具. BMC Bioinf. **14**(1), 267 (2013)

172. W. Luo, C. Brouwer, Pathview: 一个 R/Bioconductor 包用于基于路径的数据集成和可视化. 生物信息学 **29**(14), 1830–1831 (2013)

173. S. Kumar, M. Nei, J. Dudley, K. Tamura, MEGA: 一个面向生物学家的软件，用于 DNA 和蛋白质序列的进化分析. Briefings Bioinf. **9**(4), 299–306 (2008)

174. M.S. Barker, K.M. Dlugosch, L. Dinh, R.S. Challa, N.C. Kane, M.G. King, L.H. Rieseberg, EvoPipes. net: 生态和进化基因组学的生物信息学工具. Evol. Bioinf. (online) **6**, 143 (2010)

175. J.R. Frederich, F.W. Samuel, D. Maithaias, 物联网和大数据分析导论, 在迷你研讨会上, *2015*年夏威夷国际系统科学会议; A.A. Faquha, M. Mohammedi, M. Aledhari, M. Ayyash, 物联网: 关于启用技术、协议和应用的调查. IEEE通信.调查.教程 **17**(4)(2015)

176. X. Xu, S. Huang, Y. Chen, K. Browny, I. Halilovicy, W. Lu, TSAaaS: 基于物联网的时间服务分析作为一种服务, 在I*EEE* ICWS会议论文集中(2014), pp. 249–256177. J. Gantz, D. Revisel, 2020年的数字宇宙: 大数据、更大的数字阴影和远东地区的最大增长. IDC, iView: IDC分析.未来 **2007**, 1–16 (2012)178. N. Komninos, *E. P*hillipon, A. *P*ilsilli*d*es, 智能电网和家庭安全调查: 问题、挑战和对策. I*EEE*通信.调查. **16**(4), 1933–1954 (2014年第四季度)179. C. Tsai, C. Lai, M. Chiang, L.T. Yong, 物联网的数据挖掘: 一项调查. I*EEE* 通信.调查.**16**(1), 77–97 (2014年第一部分)

180. http://www.intel.in/contain/lan/www/program/embedded/internet-of-things/blueprints/iot-building-intelligent-transport-system

181. P. Domingos, 关于机器学习的几个有用的知识. Commun. ACM **55**(10) (2012)

182. N. Dalal, B. Triggs, 用于人体检测的梯度方向直方图，在 *IEEE* 计算机视觉与模式识别会议上，*2005*年。*CVPR 2005*, vol. 1 (IEEE, 2005), pp. 886 –893

183. D.G. Lowe, 从局部尺度不变特征进行物体识别，在第七届*IEEE*国际计算机视觉会议论文集，*1999*年, vol. 2 (IEEE计算机学会, *1999*), *pp*. 1150–1157

184. Y. Bengio, Y. LeCun, 尺度学习算法的扩展， 在大规模核机器, vol. 34 由 L. Bottou, O. C hapelle, D. DeCoste, J. Weston J 编写 (*M*IT Press,Cambridge, MA), pp. 321–360 (2007). ht tp://www.iro.umontreal.ca/～lisa/pointeurs/bengio+lecun_chapter2007.pdf

185. Y. Bengio, A. Courville, P. Vincent, 表示学习：综述与新视角。模式分析。机器智能。 IEEE Trans. **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
186. I. Arel, D.C. Rose, T.P. Karnowski, 深度机器学习-人工智能研究的新前沿[研究前沿]。I EEE计算智能。 **5**, 13–18 (2010)187. G.E. Hinton, S. Osindero, Y.-W. Teh, 一种用于深度置信网络的快速学习算法。神经计算。 **18**(7), 1527–1554 (2006)

188. Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle,贪婪逐层训练深网络, vol. 19 (2007)
189. H. Larochelle, Y. Bengio, J. Louradour, P. Lamblin, 探索深度神经网络训练策略。J. Mach. Learn. Res. **10**, 1–40 (2009)
190. R. Salakhutdinov, G.E. Hinton, 深度玻尔兹曼机，在国际会议上，人工智能和统计(2009) JMLR.org. pp. 448–455
191. I. Goodfellow, H. Lee, Q.V. Le, A. Saxe, A.Y. Ng, 测量深度网络中的不变性，在神经信息处理系统进展(Curran Associates, Inc.,2009), pp. 646–654

192. G. Dahl, M. Ranzato, A.-R. Mohamed, G.E. Hinton, 使用均值-协方差受限玻尔兹曼机进行电话识别，在神经信息处理系统进展(Curran Associates, Inc., 2010), pp. 469–477

193. G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke,P. Nguyen, T. Sainath, G. Dahl, B. Kingsbury, 深度神经网络用于语音识别中的声学建模：四个研究小组的共同观点。信号处理杂志 IEEE**29**(6), 82–97 (2012)

194. F. Seide, G. Li, D. Yu, 使用上下文相关的深度神经网络进行会话语音转录，在*INTERSP EECH*(ISCA, 2011), pp. 437–440
195. A.-R. Mohamed, G.E. Dahl, G. Hinton, 使用深度置信网络进行声学建模。音频语音语言处理 IEEE Trans. **20**(1), 14–22 (2012)
196. G.E. Dahl, D. Yu, L. Deng, A. Acero, 上下文相关的预训练深度神经网络用于大词汇量语音识别。音频语音语言处理 IEEE Trans. **20**(1),30–42 (2012)

197. A. Krizhevsky, I. Sutskever, G. Hinton, 使用深度卷积神经网络进行Imagenet分类（Curran
Associates, Inc., 2012年），第1106–1114页
198. T. Mikolov, A. Deoras, S. Kombrink, L. Burget, J. Cernocky, 经验评估和组合先进的语言建模技术（*I*SCA, 2011年），第605–608页

199. R. Socher, E.H. Huang, J. Pennin, C.D. Manning, A. Ng, 用于近义词检测的动态池化和展开递归自编码器（Curran Associates, *I*nc., 2011年），第801–809页

200. A. Bordes, X. Glorot, J. Weston, Y. Bengio, 联合学习词汇和意义表示以进行开放文本语义解析，在国际人工智能和统计学会议上. JMLR.org. (2012), pp. 127–135

201. 国家研究委员会,大规模数据分析的前沿(国家学院出版社, 华盛顿特区, 2013). http://www.nap.edu/openbook.php?record_id=18374
202. E. Dumbill, 什么是大数据? 大数据景观简介, 在*Strata 2012: Making Data Work* (O'Reilly, 圣克拉拉, 加利福尼亚 O'Reilly, 2012)
203. T.M. Khoshgoftaar, 克服大数据挑战. 在第 *25*届国际软件工程和知识工程大会上的论文集, 波士顿, 马萨诸塞州(ICSE. 特邀主题演讲嘉宾 (2013)

204. Y. Bengio,学习深度架构用于人工智能(现在出版社，汉诺威，马萨诸塞州，美国，2009年)
205. Y. Bengio, 表示的深度学习：展望未来，在第一届国际统计语言和语音处理会议论文集。*SLSP'13*(斯普林格，塔拉戈纳，西班牙，2013年)，第1–37页。 http://dx.doi.org/10.1007/978-3-642-39593-2_1

206. G.E. Hinton, R.R. Salakhutdinov, 用神经网络降低数据的维度。 科学 **313**(5786)，504–507
207. G.E. Hinton, R.S. Zemel, 自编码器，最小描述长度和Helmholtz自由能。Adv. Neural Inform. Process Syst. **6**，3–10 (1994年)
208. P. Smolensky，动力系统中的信息处理：和谐理论的基础，在并行分布式处理：认知微结构的探索，卷1（*M*IT出版社，1986年），第194–281页

209. G.E. Hinton，通过最小化对比散度训练专家产品。神经计算 **14**（8），1771–1800（2002年）
210. L.M. Garshol，大数据/机器学习简介。在线幻灯片展示（2013年）。 http://www.slideshare.net/larsga/introduction-to-big-datamachine-learning. http://www.slideshare.net/larsga/introduction-to-big-datamachinelearning
211. M. Grobelnik，大数据教程。欧洲数据论坛（2013年）。 http://www.slideshare.net/EUDataForum/edf2013-bigdatatutorialmarkogrobelnik?related=1
212. G. Salton, C. Buckley, 在自动文本检索中的术语加权方法. 信息处理与管理 **24**(5), 513–523 (1988)
213. S.E. Robertson, S. Walker, 对于概率加权检索的2-泊松模型的一些简单有效的近似方法，在第17届年度国际*ACM SIGIR*信息检索研究与开发会议论文集(Springer, 纽约, Inc, 1994), pp. 232–241

214. G. Hinton, R. Salakhutdinov, 通过学习深度生成模型来发现文档的二进制编码. 认知科学专题 **3**(1), 74–91 (2011)
215. R. Salakhutdinov, G. Hinton, 语义哈希. 近似推理国际期刊 **50**(7), 969–978 (2009)
216. M. Ranzato, M. Szummer, 基于深度网络的紧凑文档表示的半监督学习，在第25届国际机器学习大会(ACM, 2008)上，第792–799页

217. T. Mikolov, K. Chen, J. Dean, 高效估计向量空间中的词表示。CoRR: 计算机研究存储库 1 –12. abs/1301.3781 (2013)
218. J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A.Senior, P. Tucker, K. Yang, A. Ng, 大规模分布式深度网络，在:神经信息处理系统进展, 编者 P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L.Bottou, K.Q. Weinberger, 第25卷，第1232–1240页 (2012). http://books.nips.cc/papers/files/nips25/NIPS2012_0598.pdf

219. T. Mikolov, Q.V. Le, I. Sutskever, 利用语言之间的相似性进行机器翻译。CoRR: Comput. Res. Repository 1 –10. abs/1309.4168 (2013)
220. G. Li, H. Zhu, G. Cheng, K. Thambiratnam, B. Chitsaz, D. Yu, F. Seide, 基于上下文的深度神经网络用于实际数据的音频索引，在 *Spoken Language Technology Workshop (SLT), 2012 IEEE* (IEEE, 2012), pp. 143 –148
221. A. Zipern,在网络上快速搜索图像的方法. 纽约时报. 新闻观察文章 (2001). http://www.nytimes.com/2001/07/12/technology/news-watch-a-quick-way-to-search-for-images-on-the-web.html
222. M.A. Cusumano，Google：它是什么，它不是什么。Commun ACM Med. Image Moeling **48**(2)，15–17 (2005)。 https://doi.org/10.1145/1042091.1042107
223. H. Lee, A. Battle, R. Raina, A. Ng, Ef ficient sparse coding algorithms, in *Advances in Neural Information Processing Systems* (MIT Press, 2006), pp. 801 –808

224. Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng, Building high-level features using large scale unsupervised learning, in *Proceeding of the 29th International Conference in Machine Learning, Edingburgh, Scotland* (2012)
225. A. Freytag, E. Rodner, P. Bodesheim, J. Denzler, Labeling examples that matter: relevance-based active learning with Gaussian processes, *35th German Conference on Pattern Recognition (GCPR)* (Saarland University and Max-Planck-Institute for Informatics, Germany, 2013), pp. 282 –291
226. R. Socher, C.C. Lin, A. Ng, C. Manning, 使用递归神经网络解析自然场景和自然语言，在第*28*届国际机器学习大会上。*Om*ni*p*ress(2011), pp. 129–136

227. R. Kumar, J.O. Talton, S. Ahmad, S.R. Klemmer, 数据驱动的网页设计，在第*29*届国际机器学习大会上。 i*c*ml.*cc*/Omnipress (2012)228. Q.V. *Le*, W.Y. Zou, S.Y. Yeung, A.Y. Ng, 使用独立子空间分析学习层次不变的时空特征进行动作识别，在2011年*IEEE*计算机视觉与模式识别大会上(IEEE, 2011), pp. 3361–3368

229. G. Zhou, K. Sohn, H. Lee, 使用去噪自编码器进行在线增量特征学习，在国际人工智能与统计学大会上。JMLR.org. pp. 1453–1461 (2012)

230. P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, 通过去噪自编码器提取和组合稳健特征，在第*25*届国际机器学习大会上(A*C*M, 2008)，第1096–1103页

231. R. Calandra, T. Raiko, M.P. Deisenroth, F.M. Pouzols, 从非平稳流中学习深度置信网络，人工神经网络和机器学习-*I*C*ANN* 2*0*12(Springer, Berlin, Heidelberg, 2*0*12)，第379–386页
232. M. Chen, Z.E. Xu, K.Q. Weinberger, F. Sha, 边缘化去噪自编码器用于领域适应，在第*29*届国际机器学习大会上，苏格兰爱丁堡(2012)

233. A. Coates, A. Ng, 编码与稀疏编码和向量量化的训练的重要性，在第*28*届国际机器学习大会上的论文集。*O*mnipress。pp. 921–928 (2011)

234. G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, 通过防止特征检测器的共适应性来改进神经网络。CoRR: Comput. Res.Repository 1–18. abs/1207.0580 (2012)
235. I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout网络，在第*30*届国际机器学习大会上，亚特兰大，佐治亚州 (2013)236. A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, N. Andrew, 使用商用*hpc*系统进行深度学习，在第*30*届国际机器学习大会上的论文集(*20*13)，pp. 1337–1345

237. X. Glorot, A. Bordes, Y. Bengio, 面向大规模情感分类的领域自适应: 一种深度学习方法, 在第*28*届国际机器学习大会*(ICML-11)*上, 第*51*3–520页

238. S. Chopra, S. Balakrishnan, R. Gopalan, Dlid: 通过在领域之间插值进行领域自适应的深度学习, 在第*30*届国际机器学习大会上, 亚特兰大, 佐治亚

(2013)
239. S. Suthaharan, 大数据分类: 在网络入侵预测中的问题和挑战, 在*ACM* Sigmetrics: 大数据分析研讨会上, 匹兹堡, 宾夕法尼亚

240. W. Wang, D. Lu, X. Zhou, B. Zhang, J. Mu, 基于统计小波的大数据异常检测与压缩感知. EURASIP J. 无线通信网络. 20**13**, 269页 http://www.bibsonomy.org/bibtex/25e4**3**2dc72**3**0087ab**1**cdc65925be6d4cb/dblp

241. G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P.Nguyen, T. Sainath, G. Dahl, B. Kingsbury, 深度神经网络在语音识别中的声学建模。 四个研究小组的共同观点。信号处理杂志。IEEE**29**(6), 82–97 (2012)

242. A. Kizhevsky, I. Sutskever, G. Hinton, 使用深度卷积神经网络进行Imagenet分类在神经信息处理系统中的进展, vol. 25 (Curran Associates, Inc., 2012), pp. 1106 –1114
243. 国家研究委员会,大规模数据分析的前沿(国家学院出版社, 华盛顿特区, 2013). http://www.nap.edu/openbookphp?record_id=18374
244. E. Dumbill, 什么是大数据? 大数据景观简介, 在*Strat 2012* (Making Data Work O'Reilly, 圣克拉拉, 加利福尼亚 O'Reilly, 2012)
245. C.S.R. Prabhu, T. Bhaskara Reddy, 通过OKM扩展语义电子治理网格，一个用于在网页内容上启用语义/知识搜索的框架带宽知识表示框架，国际电子政务会议（ICEG 2010），2010年4月22日–24日，印度班加罗尔。论文编号6（2010年）。

246. L. Deng, 三类深度学习架构及其应用：教程调查。 APSIPA信号与信息处理交易。 （2012年）。
247. Goodfellow, I. J., Shlens J., Szegedy, C（2014年）解释和利用。 Advesarial Examples. ArXiv e-prints.

网站

248. https://www.researchgate.net/publication/319442626_Security_and_Privacy_in_Fog_Computing_Challenges
249. http://www.engpaper.com/fog-computing-2016.htm
250. http://ijrise.org/asset/archive/CSE_UG510.pdf
251. www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing
252. https://www.gsma.com/iot/gsma-iot-security-guidelines-complete-document-set/