Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2017, 6(1), 16-21
Published Online February 2017 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2017.61003

Hans汉斯

The Study of Machine Learning in Big Data Analysis

Qi Hong, Gang Yang, Lishan Hui

School of Mathematics and Computer Science, Shaanxi Sci-Tech University, Hanzhong Shaanxi
Email: hongqi1898@sina.com, 376309490@qq.com

Received: Dec. 29th, 2016; accepted: Jan. 13th, 2017; published: Jan. 18th, 2017

Abstract

Machine learning played a more and more important role in the analysis of large data. The main methods and techniques of machine learning under the background of large data were summarized. Firstly, the basic model and classification of machine learning were introduced. Then, several key technologies of machine learning in large data environment were described. And the article showed the popular four kinds of big data machine learning systems, and analyzed their characteristics. In the end, it pointed out the main research direction and the challenges of the big data machine learning.

Keywords

Big Data, Machine Learning, Semi-Supervised Learning, Machine Learning System in Big Data, Probabilistic Graph Model, R Language

大数据分析中机器学习研究

洪　歧，杨　刚，惠立山

陕西理工大学，数学与计算机科学学院，陕西　汉中
Email: hongqi1898@sina.com, 376309490@qq.com

收稿日期：2016年12月29日；录用日期：2017年1月13日；发布日期：2017年1月18日

摘　要

机器学习在大数据分析中起着越来越重要的作用，本文主要对大数据背景下机器学习方法和技术等进行了归纳和总结。首先对机器学习的基本模型、分类进行简介；然后对大数据环境下的机器学习的几个关键技术进行了叙述；接着展示了目前流行的四种大数据机器学习系统，并分析了其特点；最后指明了大

文章引用: 洪歧, 杨刚, 惠立山. 大数据分析中机器学习研究[J]. 人工智能与机器人研究, 2017, 6(1): 16-21.
https://doi.org/10.12677/airr.2017.61003

数据机器学习的主要研究方向和所遇到的挑战因素等。

关键词

大数据，机器学习，半监督学习，大数据机器学习系统，概率图模型，R语言

1. 引言

在科学领域，包括信息、物理、生物、天文等在内的各个领域的科学发现已经发展到第四阶段，即基于大数据的数据密集型科学发现[1]。大数据分析挖掘处理主要分为简单分析和智能化复杂分析两大类[2]。简单分析常用SQL语句来完成一些统计和查询工作，这些方法与数据库OLAP的处理技术极为相似；而大数据的深度价值通常需要使用基于机器学习和数据挖掘的智能化复杂分析才能实现[3] [4] [5]。

一直以来，机器学习领域的专家和学者们在不断尝试对越来越多的数据进行学习，如今随着大数据时代的到来，对机器学习方法提出了更多新的要求。

2. 相关领域研究现状

机器学习是人工智能的一个核心研究领域。机器学习[2]是一种利用系统本身进行自我改进的过程，在这个过程中计算机程序的性能随着经验的积累而不断提高。专家、学者们不断提出了各种学习任务算法，这些算法大大提高了计算机从大量数据中提取特征并发现隐含规律的能力，数据挖掘和分析中的机器学习方法的应用越来越广泛。研究表明：在很多情况下，机器学习模型的效果会随着所处理的数据规模越大而越好。

近年来大数据机器学习成为机器学习领域的研究热点之一。Kleiner 等人[5]基于集成学习中 Bagging 的思想提出了新型数据采样方法 BLB (Bag of Little Boot-straps)，用来解决 Bootstrap 在遇到大数据时的计算瓶颈问题；Shalev-Shwartz 和 Zhang [6]基于随机学习的思想提出了梯度上升(下降)的改进方法，用来实现大规模模型的快速学习；卓林超等[7]针对大数据中的乱序数据缺少关联规则的问题，提出了一种动态调整的改进型算法，能够获得更多的收敛次数，并能有效地提高收敛率，进而提高整体网络性能；许烁娜等[8]在大数据环境下，应用 L1 准则的稀疏性，提出了一种在线特征提取算法，并用该算法对训练实例进行了分类；Gonzalez 等人提出了基于多机集群的分布式机器学习框架 GraphLab，用以实现基于图的大规模机器学习等。

3. 机器学习概述

3.1. 机器学习的基本模型

机器学习不但是人工智能发展的重要标志，也是计算机获取知识的重要途径，它是一门研究怎样用计算机来模拟或实现人类学习活动的学科。以 H. Simon 的学习定义作为出发点，建立如图 1 的简单学习模型[9]。其中，环境表示外界信息集合；学习环节先从环境获取外部信息，接着将这些信息加工(主要有类比、综合和分析等)成知识并放到知识库中；学习环节得到的知识被存放在知识库中；执行环节利用

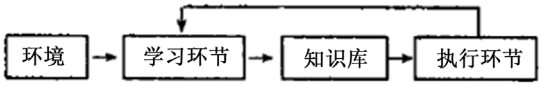

Figure 1. The basic model of machine learning
图 1. 机器学习基本模型

前一阶段的知识来履行某种任务，同时它将本环节中的一些信息反馈给知识库的前一环节从而指导进一步的机器学习过程。

3.2. 机器学习的分类

机器学习按照学习形式可分为以下两类[10] [11] [12]：

1) 监督学习

即在机器学习的过程中做出对错指示。在预测和分类中常常要用到监督学习，在监督学习中一个函数关系式可以从被训练的数据集中总结出来，然后用这个函数关系式来对新的数据进行预测并得到结果。在监督学习中，训练集需要输入，然后可以人为标注训练集中的目标，最后才能得到输出的结果。常见的监督学习算法有统计分类和回归分析。

2) 非监督学习

又称归纳性学习，是一种通过循环和递减运算来减小误差，从而达到分类的目的算法。无监督学习的智能性最高但发展比较缓慢，不是目前研究的主流；监督学习中常常由已知来推断未知，风险较大，有时结果不可靠；因此人们对前两者进行充分研究并发现了半监督学习方法，这种方法目前引起了人们极大的兴趣和关注。

4. 大数据环境下机器学习的关键技术

当前，机器学习中最常用的关键技术有：半监督学习、集成学习、迁移学习、贝叶斯网络、决策树、统计学习理论与支持向量机、隐马尔可夫模型、神经网络、k 近邻方法、序列分析、聚类、粗糙集理论、回归模型等。在大数据分析中，半监督学习、迁移学习、概率图模型和集成学习等技术尤为重要[13]。

1) 半监督学习

在有监督学习中，利用的是已标识数据，而无监督学习中只利用未标识数据。在大数据时代，

已标识数据的数量总是远远小于未标识数据的数量，因此要想利用好这些未标识的数据就应该采用半监督学习法，半监督学习是研究如何综合利用大量未标识数据和少量已标识数据而获得的不但具有良好性能而且具有泛化能力的机器学习方法。半监督学习包括：基于生成式模型的半监督学习、基于低密度划分的半监督学习、基于图的半监督学习以及基于不一致性的半监督学习。

2) 概率图模型

大数据分析的一个重要内容是从具有不确定性的大数据中的获得有价值的知识。概率图模型是图论与概率论相结合的产物，是图形化之后的概率分布形式；概率图模型实际上是一个统一的框架，在这个框架中不但可以为大规模多变量构建一个统计模型，而且可以捕获随机变量之间复杂的依赖关系。概率图模型一方面用图论的语言直观揭示问题的结构，另一方面又按照概率论的原则对问题的结构加以利用，降低推理的计算复杂度。因子分解是概率图模型中的一个核心概念，一个概率图模型是由一组概率分布所构成的。概率图通过图形的方式来捕获并展现所有随机变量的联合分布，通过分解成各因子乘积的方式来实现。

概率图模型主要包括：贝叶斯网络、马尔可夫网络和隐马尔科夫模型 ，其中贝叶斯网络最为流行。贝叶斯网络又称为因果网、概率网或者信念网，变量之间的关系可以用贝叶斯网络来表示 ；贝叶斯网络

可为任何全联合概率分布提供一种有向无环图结构，这种结构具有有效、自然、规范、简明等优点。贝叶斯网络还提供了一系列的算法，这些算法可自动地分析相关信息并得到更多隐含的信息从而指导决策。此外，贝叶斯网络还可以模拟人类的认知过程、学习方式，灵活地对参数和结构进行相应的修正与更新，这种学习机制显得非常灵活。

3) 迁移学习

迁移学习是指在不同情况之间把知识进行迁移转化的能力。提高机器学习能力的一个关键问题就在于要让机器能够继承和发展过去学到的知识，这其中的关键就是让机器学会迁移学习。迁移学习可分为直推迁移学习、归纳迁移学习以及无监督迁移学习。迁移学习试图通过将在一个或多个源任务中学习到的知识进行迁移，将它们用在相关的目标任务中以提高其学习性能。

5. 大数据机器学习系统

大数据机器学习是一个同时涉及机器学习和大数据处理两个主要方面的交叉性研究课题。

5.1. 主要研究问题

大数据机器学习需要重点研究解决大数据场景下所特有的两大技术问题[2]：一是大数据复杂分析时的计算性能问题；二是大数据机器学习系统的可编程性和易用性问题。前者主要是由于在大数据环境下，现有的大多数机器学习算法效果很不理想或常常失效，此时这些算法需要被大幅度地修改或重写。后者是由于大数据处理技术及其平台比较复杂，普通程序员用常规的程序设计方法无法在此环境下编程并进行大数据分析，他们需要提前对大数据处理平台和大数据处理技术进行较系统地学习。

5.2. 典型大数据学习方法和系统简介

1) 基于特定平台的定制式并行化机器学习算法与算法库[2] [14] [15] [16] [17]。

有大量的研究工作致力于基于 Hadoop MapReduce 和 Spark 以及传统的 MPI 并行计算框架，完成各种并行化机器学习和数据挖掘算法的设计。但这些算法或平台的专业技术要求高、工作过程较繁琐，缺乏并行计算和分布式知识背景的普通程序员难以使用。

为了解决这个问题，可以在不同的并行化计算平台上，由专业的机器学习算法设计者实现并行化机器学习算法，提供一个机器学习和数据挖掘工具包以供一般的数据分析和应用开发程序员直接使用，如 Hadoop 下的 Mahout 以及 Spark 环境下的 MLlib 等。

2) 结合传统数据分析平台的大数据机器学习系统

人们已经尝试在 R 中利用分布式并行计算引擎来处理大数据。最早的工作系统 RHadoop，它将统计语言 R 与 Hadoop 结合起来，Hadoop 主要用来存储和处理底层的海量数据，用 R 语言替代 Java 语言完成 MapReduce 算法的设计实现。类似地，2014 出现的 SparkR 也是作为一个 R 的扩展包，它允许用户在 R 的 shell 环境里交互式地向 Spark 集群提交运行作业。但是，目前的 RHadoop 和 SparkR 都存在一个同样的问题：仍要求用户熟悉 MapReduce 或 Spark RDD 的编程框架和程序结构，这给普通程序员带来了很多困难。此外，这些工作都是基于单一平台，无法解决跨平台统一大数据机器学习算法设计的问题。

3) 基于特定平台的大数据机器学习系统

近年来，学术界和工业界开始尝试总结机器学习算法设计的共同特性，结合大规模学习时所需要考虑的底层分布式数据存储和并行化计算等系统问题，专门研究能同时兼顾并支持大数据机器学习和大数据分布并行处理的一体化大数据机器学习系统。

在国内外的学术会议中，已经出现了许多与大数据机器学习系统相关的研究工作，如 Spark MLlib、

Apache Flink、IBM 的 SystemML、Parameter Server、GraphLab、Petuum 等；此外，腾讯、百度等国内著名互联网企业也推出了不同的面向大数据的分布式机器学习系统，如腾讯的 Peacock 和 Mariana 大规模机器学习系统、百度的 ELF 和百度机器学习云平台 BML 等。

4) 跨平台统一大数据机器学习系统 Octopus [2]

南京大学 PASA 大数据实验室设计了一个跨平台大数据机器学习和数据分析的统一编程模型和系统平台。该系统基于矩阵编程计算模型，结合 R 编程语言和编程方法，设计提供一个跨平台的统一编程计算框架，实现了一个跨平台大数据机器学习系统 Octopus (大章鱼)。

Octopus 允许数据分析和大数据应用开发程序员轻松地设计和开发各种大数据机器学习和数据分析算法与应用程序。通过提供基于矩阵的统一编程计算模型，使用基于 R 语言的数据分析程序设计，允许用户方便地编写和运行常规的 R 语言程序，而无需了解底层大数据平台的分布和并行化编程计算知识；底层平台上，通过良好的系统层抽象，可以快速集成 Hadoop 和 Spark 等通用大数据并行计算框架和系统平台，而且程序仅需编写一次，不需要有任何修改即可根据需要选择并平滑运行于任何一个平台，从而实现“Write Once，Run Anywhere”的跨平台特性。

6. 大数据机器学习的研究方向及挑战因素

6.1. 大数据机器学习的研究方向

当前，大数据机器学习主要有两个研究方向[13]：第一个是针对学习机制的研究，主要研究如何使机器具有人类的某些行为特征；第二个是研究如何发现并挖掘大数据中的有价值的信息。当前人们的研究重点主要集中在后者上面。

6.2. 大数据机器学习的挑战因素

在未来几十年内，机器学习领域内会面临下边几方面的挑战[13] [18] [19] [20] [21] [22]：

1) 机器学习泛化能力的提升。这个问题十分普遍，一般地把多个不同对象具有相同的处理能力称作泛化能力，当前支持向量机具有的最强的泛化能力。

2) 速度问题。在机器学习领域，人们一直不断追求的目标就是如何提高机器学习的速度，速度训练和速度测试之间的关系和怎样才能有效削弱二者之间的冲突是人们最关心的问题。

3) 可理解性。大多数情况下，人们常常能通过机器学习算法得到结果，却不知道为什么会得到这样的结果。在未来的大数据分析中，越来越多的人希望利用机器学习算法不但能得到结果，而且能知道结果产生的原因。

4) 数据处理的能力。过去大多数机器学习方法处理的对象是经过标记的数据。但在未来的大数据分析中，机器学习算法不但要处理大量未标记的数据，而且还要受一些不平衡数据、垃圾数据等的干扰和影响。

5) 代价敏感。当前不断追求低误码和降低错误率是人们对机器学习技术的迫切需求，当然人们对错误代价的容忍度是随领域、学科的不同而不同的，未来人们期望能用最小的代价从机器学习技术中得到越来越多的收益。

7. 结语

机器学习方法种类繁杂，在大数据分析中许多机器学习方法需要不断修改和完善才能发挥其作用。目前，大数据机器学习系统尚处在一个初期的探索和研究阶段，许多方面都不太成熟和完善。大数据时代大数据信息分析和挖掘离不开机器学习方法。随着大数据的发展，机器学习方法和机器学习系统也将

不断发展，可以坚信未来人们必将充分利用机器学习方法从大数据中获得越来越多的有用信息。

基金项目

陕西省教育厅科研专项(16JK1163)、陕西理工大学科研基金(SLGQD0903)。

参考文献 (References)

[1] 李武军, 周志华. 大数据哈希学习: 现状与趋势[J]. 科学通报, 2015, 60(5-6): 485-490.

[2] 黄宜华. 大数据机器学习系统研究进展[J]. 大数据, 2015, 1(1): 28-47.

[3] Zhou, Z.H., Chawla, N.V., Jin, Y., *et al*. (2014) Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives. *IEEE Computational Intelligence Magazine*, **9**, 62-74. https://doi.org/10.1109/MCI.2014.2350953

[4] Jordan, M. (2011) Message from the President: The Era of Big Data. *ISBA Bulletin*, **18**, 1-3.

[5] Kleiner A., Talwalkar, A., Sarkar, P., *et al*. (2012) The Big Data Bootstrap. *Proceedings of the* 29*th International Conference on Machine Learning* (ICML), Edinburgh, 27 June-3 July 2012, 1759-1766.

[6] Bryant, R.E. (2011) Data-Intensive Scalable Computing for Scientific Applications. *Computing in Science & Engineering*, **13**, 25-33. https://doi.org/10.1109/MCSE.2011.73

[7] 卓林超, 王堃. 大数据中面向乱序数据的改进型 BP 算法[J]. 系统工程理论与实践, 2014, 34(6): 158-164.

[8] 许烁娜, 曾碧卿, 熊芳敏. 面向大数据的在线特征提取研究[J]. 计算机科学, 2014, 41(9): 239-242.

[9] 田文英. 机器学习与数据挖掘[J]. 石家庄职业技术学院学报, 2004, 16(6): 30-32.

[10] 汪加才, 常青. 面向机器学习与数据挖掘实践教学的自由软件分析[J]. 南京审计学院学报, 2004, 1(3): 91-95.

[11] 王肇国, 易涵, 张为华. 基于机器学习特性的数据中心能耗优化方法[J]. 软件学报, 2014(7): 1432-1447.

[12] 朱军, 胡文波. 贝叶斯机器学习前沿进展综述[J]. 计算机研究与发展, 2015, 52(1): 16-26.

[13] 陈康, 向勇, 喻超. 大数据时代机器学习的新趋势[J]. 电信科学, 2012, 28(12): 88-95.

[14] Zaharia, M., Chowdhury, M., Das, T., *et al*. (2012) Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. *Proceedings of the* 9*th USENIX Conference on Networked Systems Design and Implementation* (*NSDI*), San Jose, 25-27 April 2012, 141-146.

[15] Venkataraman, S., Bodzsar, E., Roy, I., *et al*. (2013) Presto: Distributed Machine Learning and Graph Processing with Sparse Matrices. *Proceedings of the* 8*th ACM European Conference on Computer Systems* (*EuroSys*), Prague, 14-17 April 2013, 197-210.

[16] TeraData (2012) The Threat Beneath The Surface: Big Data Analytics, Big Security and Real-Time Cyber Threat Response For Federal Agencies. TeraData, Miamisburg, 1-35.

[17] Zhang, X., Liu, C., Surya, N., *et al*. (2014) Privacy Preservation over Big Data in Cloud Systems. In: Nepal, S. and Pathan, M., Eds., *Security, Privacy and Trust in Cloud Systems*, Springer, Berlin Heidelberg, 239-257.

[18] 王晓. 大数据环境下机器学习算法趋势研究[J]. 哈尔滨师范大学自然科学学报, 2013, 29(4): 48-50.

[19] 张长水. 机器学习面临的挑战[J]. 中国科学: 信息科学, 2013, 43(12): 1612-1623.

[20] Darwiche, A. (2009) Modeling and Reasoning with Bayesian Networks. Cambridge University Press, Cambridge, 32-35. https://doi.org/10.1017/CBO9780511811357

[21] Pan, J.L. and Yang, Q. (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[22] Bahadori, M.T., Liu, Y. and Zhang, D. (2011) Learning with Minimum Supervision: A General Framework for Transductive Transfer Learning. *IEEE International Conference on Data Mining* (*ICDM*), Vancouver, 11-14 December 2011, 61-70. https://doi.org/10.1109/icdm.2011.92