齐 齐 哈 尔 大 学

外文翻译（正文）

题 目 基于python的影评数据爬取和分析研究

学 院 计算机与控制工程学院

专业班级 软件152

学生姓名 周世豪

指导教师 吴迪

2019年 5月 15日

# 大数据机器学习：机遇与挑战

# 摘要：机器学习在各种应用中不断释放其力量。近年来，由于大数据的出现，它已被推到了最前沿。在大数据的挑战下，机器学习算法从未得到更好的承诺。大数据使机器学习算法比以往任何时候能够发现更细粒度的模式和做出更及时，准确的预测;另一方面，它提出了机器学习的主要挑战，例如模型可扩展性和分布式计算。在本文中，我们介绍了机器学习大数据框架，以指导其对机遇和挑战的讨论。该框架以机器学习为中心，遵循预处理，学习和评估的各个阶段。此外，该框架还包括四个其他组件，即大数据，用户，域和系统。 机器学习的阶段和机器学习框架的组成部分在许多尚未探索或未探索的研究领域开辟未来的工作提供了别识别相关机会和挑战的方向。

关键词：机器学习；大数据；数据预处理；评估；并行。

## 1.简介

机器学习（ML）技术已经在广泛的应用中产生了巨大的社会影响，例如计算机视觉，语音处理，自然语言理解，神经科学，健康和物联网。大数据时代的到来激发了人们对机器学习的广泛兴趣。机器学习算法在获取对各种业务应用程序和人类行为的新见解时也受到大数据的挑战。一方面，大数据给机器学习算法带来了前所未有的丰富信息，借此可以提取底层模式并构建预测模型;另一方面，传统的机器学习算法面临着诸如可扩展性等关键挑战，以真正释放大数据的隐藏价值。随着大数据不断扩大，机器学习必须发展壮大，才能将大数据转化为可操作的情报。

机器学习解决了如何构建通过经验自动改进的计算机系统的问题[1]。 机器学习问题被称为从某些任务和绩效测量的经验中学习的问题。 机器学习技术使用户能够发现底层结构并从大型数据集中进行预测。 机器学习在高效的学习技术（算法），并在丰富的和不断扩大的数据以及强大的计算环境中茁壮成长。因此，机器学习具有巨大的潜力，并且是大数据分析的重要组成部分[2]。本文重点介绍大数据和现代计算环境中的机器学习技术。具体而言，我们的目标是研究机器学习对大数据的机遇和挑战。大数据为机器学习提供了新的机会。例如，大数据以多种粒度和多样性实现模式学习，从多个视图以固有的并行方式实现。此外，大数据提供了基于序列链进行因果推断的机会。然而，大数据也给机器学习带来了重大挑战，例如高数据维度，模型可扩展性，分布式计算，流数据[3]，适应性和可用性。在本文中，我们介绍了机器学习大数据框架（MLBiD），以指导其对机遇和挑战的讨论。该框架机器学习以为中心，遵循预处理，学习和评估的各个阶段。此外，该框架还包括影响机器学习并受其影响的四个其他组件，即大数据，用户，域和系统。 机器学习大数据框架的组成部分和机器学习的各个阶段为确定机遇和挑战提供了方向，并在许多尚未开发的研究领域开辟了新的道路。

# 大数据机器学习框架

大数据机器学习（MLBiD）的框架如图1所示.MLBiD以机器学习（ML）组件为中心，它与其他四个组件交互，包括大数据，用户，域和系统。 这种互动是双向的。 例如，大数据作为机器学习的输入，后者产生输出，而输出又成为大数据的一部分; 用户可以通过提供领域知识，个人偏好和可用性反馈，以及通过利用学习成果来改进决策来与机器学习进行交互; 域既可以作为指导机器学习的知识来源，也可以作为应用学习模型的背景; 系统架构的会影响学习算法的运行方式以及运行它们的效率，系统架构的协同设计或许能满足机器学习的需求。 接下来，我们将分别介绍大数据机器学习的每个组件。

**2.1 机器学习**

机器学习通常经历数据预处理，学习和评估阶段（参见图1）。数据预处理有助于将原始数据准备为“正确的形式”，以用于后续学习步骤。原始数据可能是非结构化的，嘈杂的，不完整的和不一致的，预处理步骤将这些数据转换为可用作通过数据清理，提取，转换和融合进行学习的输入的形式。学习阶段选择学习算法并调整模型参数以使用预处理的输入数据生成期望的输出。一些学习方法，特别是表征性学习，也可用于数据预处理。评估遵循以确定学习模型的性能。例如，分类器的性能评估涉及数据集选择，性能测量，误差估计和统计测试[4]。评估结果可能导致调整所选学习算法的参数和/或选择不同的算法。

机器学习可以在多个维度上表征：学习反馈的性质，学习任务的目标和数据可用性的时间。因此，我们提出了机器学习的多维分类，如图2所示。

根据学习系统可用反馈的性质，机器学习可分为三种主要类型：监督学习，无监督学习和强化学习[5]。在监督学习中，学习系统提供了输入 - 输出对的示例，目标是学习将输入映射到输出的功能。在无监督学习中，系统没有提供明确的反馈或期望的输出，并且目标是揭示输入中的模式。与无监督学习一样，强化学习系统没有呈现输入 - 输出对。与监督学习一样，强化学习也会对其以前的经验给予反馈。然而，与监督学习不同，强化学习中的反馈是与行动相关的奖励或惩罚，而不是期望的输出或次优动作的明确校正。半监督学习落在有监督和无监督学习之间，其中系统呈现有少量输入 - 输出对和大量未注释的输入。半监督学习的目标与监督学习类似，只是它从注释和未注释的数据中学习。

基于学习目标是使用输入特征还是特征本身的特定任务，机器学习可以分为表征学习和任务学习。特征学习旨在学习新的数据表示，以便在构建分类器或其他预先判断器时更容易提取有用的信息[6]。良好的表现是解开变异的潜在因素的表现。在概率模型的情况下，通常可以捕获观测输出的潜在探索因子的后验分布[6]。表示学习通常与密度估计和维数减少纠缠在一起。密度估计找到随机变量的下位概率密度函数。维度缩减将输入从高维空间映射到低维空间。在特征学习中建立明确的目标或目标通常很困难。相反，任务学习通常具有期望的输出，因此被分类为分类，回归和聚类。在分类中，机器学习技术产生一种模型，该模型将看不见的输入分配给一个或多个预定义的类。回归与分类的不同之处在于其输出是连续的而不是离散的。聚类产生数据组，并且这些组事先不知道，这与分类不同。传统上，分类和回归被称为监督学习，而聚类被称为无监督学习。他们的代表性算法也显示在图2中。

基于使训练数据可用的时间（例如，训练数据是一次全部可用还是一次一个），机器学习可以分为批量学习和在线学习。批量学习通过学习整个训练数据来生成模型，而在线学习基于每个新输入更新模型。批量学习算法假定数据是独立的并且从相同的概率分布中识别分布或抽取，这通常不被实际数据所满足。在线学习通常不对数据做出统计假设[7]。虽然预期批量学习算法可以概括，但是没有关于在线学习的概括的概念，因为该算法仅被期望准确地预测它作为输入接收的示例的标签[7]。当在整个数据集上进行训练或随着时间的推移生成数据并且学习系统需要适应数据中的新模式时，在计算上不可行时使用在线学习。

每个机器学习算法可以分为多个维度。例如，传统的决策树属于监督批量学习算法。

**2.2 大数据**

大数据的特征在于五个维度：数量（数量/数据量），速度（数据生成速度），种类（数据的类型，性质和格式），准确性（捕获数据的可信度/质量）和价值（见解和影响）。我们将五个维度组织成一个堆栈，包括从底部开始的大层，数据层和价值层（参见图3）。大层是最基础的，数据层是大数据的核心，价值方面是大数据实际应用的影响。较低层（例如，体积和速度）更多地依赖于技术进步，而较高层（例如，价值）更倾向于利用大数据的战略力量的应用。为了实现大数据分析的价值并有效地处理大数据，需要调整现有的机器学习范例和算法。

**2.3 其他组件**

2.3.1 用户

机器学习系统有各种各样的利益相关者，如领域专家，最终用户，机器学习研究人员和从业者。传统上，机器学习从业者在应用机器学习时做出大多数决定，从数据收集一直到绩效评估。在此过程中最终用户的参与仅限于提供数据标签，回答与域相关的问题，或者提供有关学习结果的反馈，这通常由从业者调解，导致冗长和异步迭代[8]。但是，最终用户倾向于提供的不仅仅是数据标签。他们重视学习系统设计的透明度，这反过来又有助于他们理解系统并提供更好的标签/反馈。让用户参与机器学习可能会带来更有效的学习系统和更好的用户体验[8]。例如，交互活动机器学习 [8]允许用户以交互方式检查其行为的影响，并调整后续输入以引导机器学习行为以获得所需输出。

2.3.2 域

领域知识有助于机器学习发现有趣的模式，这些模式可能无法单独从数据集中发现。训练数据集可能不够大或代表性不够，不能发现所有模式。获得足够和有代表性的数据也是昂贵的，甚至是不可行的，可能是由于很大的域变化和特定应用的要求。领域知识可以帮助提高数据集引起的模式的通用性和鲁棒性[9]。有几种方法可以将先前的领域知识纳入归纳机器学习 [10] :( 1）准备训练样例; （2）产生假设或假设空间; （3）修改搜索目标; （4）增加搜索。这些学习模式又可用于更新和优化领域知识。

2.3.3 系统

由软件和硬件组成的系统架构或平台创建了一个可以运行机器学习算法的环境。例如，与较简单的对应物相比，具有分布式架构的多核机器有望提高机器学习的效率。已经提出了诸如Hadoop / Spark的新框架和系统架构来解决大数据的挑战。然而，将现有机器学习算法迁移到分布式架构需要修改机器学习算法的实现和部署方式。此外，机器学习的独特需求和价值可能会激发新系统架构的设计和开发。

基于大数据机器学习框架，我们确定了重要的机会和关键挑战。我们在机器学习的三个阶段：预处理，学习和评估中分别讨论它们。

# 数据预处理的机会和挑战

部署机器学习系统的大部分实际工作都用于预处理管道和数据转换的设计，这些转换导致数据的表示可以支持有效的机器学习 [6]。数据预处理旨在解决许多问题，例如数据冗余，不一致性，噪声，异质性，转换，标签（用于（半）监督机器学习），数据不平衡和特征表示/选择。由于需要人工和大量可供选择的选项，数据准备和预处理通常成本很高。另外，一些传统的数据假设不适用于大数据，因此一些预处理方法变得不可行。另一方面，大数据通过直接学习大量，多样化和流数据源，创造了减少人类监督依赖的机会。

**3.1 数据冗余**

当两个或多个数据样本表示同一实体时，会出现重复。数据重复或不一致对机器学习的影响可能很严重。尽管在过去20年中发现了一系列用于识别重复的技术[11]，但成对相似性比较等传统方法对大数据来说已不再可行。此外，与非重复对相比，重复对是少数的传统假设不再成立。为此，动态时间扭曲可以比最先进的欧几里德距离算法快得多[12]。

**3.2 数据噪音**

大数据承诺提供来自不同类型的存储库的多视图数据，不同的格式以及不同的人口样本，因此是高度异构的。这些多视图异构数据（例如，非结构化文本，音频和视频格式[15]）可能对于学习任务具有不同的重要性级别。因此，通过将它们视为同等重要来连接所有特征将不可能导致最佳学习结果。大数据提供了从多个视图并行学习的机会，然后通过了解特征视图对任务的重要性来集成多个结果。预计该方法对数据异常值具有鲁棒性，并且可以解决优化难度和收敛问题[16]。

**3.4 数据离散化**

某些机器学习算法（如决策树和朴素贝叶斯）只能处理离散属性。离散化将定量数据转换为定性数据，从而获得连续域的非重叠划分。属性离散化的目的是将简洁的数据表示作为类别找到，这足以使学习任务尽可能多地保留原始连续属性中的信息。但是，在处理大数据时，大多数现有的离散化方法都不会有效。为了解决大数据挑战，标准离散化方法通过在大数据平台中基于最小描述长度原理开发分布式版本的熵最小化离散器来实现平衡，提高了性能和准确性[17]。在另一项研究[18]中，数据首先根据数值属性的值进行排序，然后分成原始类属性的片段。这些片段通过不同类别的百分比组成来概括，被视为超级实例和离散化的目标。

## 3.5 数据标签

传统的数据注释方法是劳动密集型的。已经提出了几种替代方法来解决大数据的挑战。例如，在线人群生成的存储库可以作为免费注释训练数据的来源，可以在类别数量和类内多样性方面捕获大量变化[19]。此外，通过概率程序归纳可以实现人类概念学习[20]。此外，标记数据的能力被内置到机器学习算法中，例如半监督学习，转移学习和主动学习（例如，[21,22]）。通过使用主动学习作为人群来源数据库中标记任务的优化策略，可以最大限度地减少向人群提出的问题数量，从而允许众包应用程序扩展。然而，为众包数据集设计主动学习算法带来了许多实际挑战，例如通用性，可扩展性和易用性[23]。另一个问题是，这样的数据集可能无法覆盖所有用户特定的上下文，这可能导致性能明显差于以用户为中心的培训[19]。

## 3.6 不平衡的数据

传统的分层随机抽样方法已经解决了数据不平衡的问题。但是，如果涉及子样本生成和错误度量计算的迭代，则该过程可能非常耗时。此外，传统的采样方法无法有效地支持用户指定的数据子集（包括基于值的采样）的数据采样。大数据需要并行数据采样。例如，已经提出了一种并行采样框架，用于基于多个分布式索引文件从原始数据集中生成样本数据集[24]。可以基于数据集大小和可用过程来选择并行级别。

## 3.7 特征表示和选择

机器学习的性能在很大程度上取决于数据表示或特征的选择[6]。 机器学习算法的普遍性取决于数据集，数据集也间接依赖于表示数据集的显着结构的特征。通过识别突出特征，特征选择有助于提高机器学习的性能。它基本上选择不同的特征和数据子集，并以不同的粒度级别聚合它们，这有助于减少大数据量。然而，特征工程需要先前的领域知识和人类的聪明才智，并且通常是劳动密集型的[6]。为了解决当前特征工程算法在处理大数据时的弱点，已经提出了各种解决方案，例如分布式特征选择[25];低秩矩阵近似（例如，标准Nystr-m方法[26]）;表示学习通过学习一般经验[6]使学习算法减少对特征工程的依赖;用于超高维特征选择的自适应特征缩放方案，其迭代地激活一组特征并解决多个核学习子问题的序列[27];基于谱图理论的统一的特征选择框架，能够为监督和非监督特征选择生成算法族[28];分类前的模糊聚类，其中分组以群组为中心实现，然后通过减少数据进行去聚类和分类[29];并且减小了数据维度和体积的大小（例如，随机森林前向选择排名和随机森林后退消除排名[30]，以及具有所选特征的语言对冲神经模糊分类器[31]）。最近，基于深度神经网络的自动编码已被证明在学习视频，音频和文本功能方面非常有效[32,33]。