

大数据环境下机器学习算法趋势研究

张艳敏

(河北软件职业技术学院 071000)

摘要: 数据分析是以统计为基础的,它最初被广泛应用于大量的数据分析和处理。人工智能可以通过电脑等进行某些功能上的处理,从而更好地发挥出人的智能。机器学习是人工智能的一个分支,它的首要目的就是让机器在没有程序的情况下,自主地学习,并处理一些繁琐的数据。大数据分析机器学习有着密切的联系,在相关的研究领域中,如果能将其充分利用,就可以发挥出其独有的作用。利用机器学习技术和大数据处理技术,可以有效地解决实际生活中的各种问题。

关键词: 大数据;机器学习算法;趋势

引言:

大数据是一种利用传统的IT技术和软件、硬件设备,对数据进行感知、获取、管理、处理和存储的一种数据,或者说,它是一种不能被装入电脑内存的数据集,它可以有效地扩展样本之间的距离和区域,减少错误的判断,减少数据的危险,同时,它的支撑矢量则是将一个数据空间转换成另外一个高维空间,然后经过分类,获得更为精确的数据。

1.大数据的特点与分类

1.1 大数据的特点

在大数据发展的进程中,已经由原来的3V型向4V型发展。3V模式在速度、容量和多样性上都有优势,而4V模式在价值、时效性、虚拟性、可变性和有效性等诸多方面表现出了诸多优势。为了更好地理解大数据的这些特点,专家们普遍认为,要把信息技术和人工智能技术相结合,才能确保数据的流畅。而在大数据的分析与处理中,人与机器的智慧是至关重要的。另外,由于对数据和信息的要求越来越高,大数据的收集、传递、处理和应用等技术也在发生着变化,使得部分半结构化、结构化和非结构化数据的价值得以充分利用。

1.2 大数据的分类

1.2.1 支持向量机分类

如果将传统的机器学习方法用于大数据的分类,则会出现大量的计算机会,这对于大规模的数据采集和处理是非常不利的。另一方面,在非参数的空间仿真中,也会出现一些随机的数据。为了解决上述问题,我们需要一种在线的机器学习技术。基于序列原则的在线机器学习方法可以提高运算速度,同时也可以提高数据的采集能力,但是这样做会导致数据的处理效率下降,因此,对于海量数据的分类,可以采用增量式算法和最小二乘支持矢量算法来进行,这样既可以节省存储资源,又可以方便地将海量数据进行分类。

1.2.2 神经网络和极限学习机算法

极限学习机算法(Extreme Learning Machine, ELM),是南洋科技大学的黄广斌教授,提出的一种解决单隐层神经网络的方法。ELM最大优点在于它能较好地提高传统神经网络的学习效率,特别是单隐层前馈神经网络。在传统的神经网络中,它是采用梯度递减法来调节权重,但是它的泛化性差,效率低,运算速度缓慢。如果要解决这个问题,可以采用ELM方法,将网络中的偏离项和输入权重随机分配给网络的权重。与传统的方法比较,ELM方法的运算速度明显提高。

1.2.3 决策树分类

由于传统的决策树算法存在着大量的存储空间,因此,在处理海量数据时,可以采用一种新的大数据分析技术,将其应用于数据挖掘中,从而有效地解决了机器学习算法的局限性,并且在运算速度上也有了较大的提高。此外,采用增量法可以有效地提高决策树算法的效率,而且该方法在保证数据准确性的前提下,能够对大量噪声数据进行处理。

2.在大数据支撑下的机器学习经典算法

在现有的机器学习算法中,不同的样本对同一机器学习算法产生的影响是不同的。由于存在冗余、含噪声的数据,不但会极大地浪费内存,还会影响整体的运算效率,影响学习效果。所以,我们更倾向于利用大数据技术来采集大量的数据,根据相关的规范,选出符合要求的数据,构成计算机所需的样本,然后在该样本空间中构建学习算法,完成学习任务。基于大数据技术的数据采集和处理,机器学习能够更好地实现我们所设置的学习任务,从而达到更好的学习效果,而AlphaGo则是其中最好的例子。

2.1 大数据处理的基本策略

基于大数据的机器学习技术,可以将其划分为“分治”和“大数据采样”两种。分割策略是一种以多分支递推为基础的分支递推算法,其工作原理是把一个复杂的问题分解为若干个类似或类似的子问题,待子问题足够简单时,再把该问题引入到原来的问题中。分割策略就是它的最大优点,即每次数据的处理都会缩减搜索空间,使整个过程的复杂程度与裁剪过程一样,这就是检索的作用。

在机器学习中,大数据取样是一个很重要的环节,取样误差会直接影响到模型的表现,而在训练和学习阶段,保证数据样本能够反映出基本的分布。在使用大数据的过程中,使用适当的取样技术,其代价要高于总体的大数据,降低样本的误差是最重要的。

2.2 大数据分类

大数据分类主要是根据已有的训练数据,判断新观测值归属于哪个子群。它是一种典型的基于监督学习的机器学习方法,它是一种人工控制训练数据集和训练结果的机器学习方法。大数据分类法将数据分为不同的类型,并将其定义为若干可量化的属性,并将其命名为特性或解释变量。这些解释变量可以是真实的,也可以是分类的类别,分类的等级,特定的整数等等。一种特定的分类生成方法,就是由分类器将输入的数据进行映射。在大数据分析中,存在着两种不同的分类方法,其中最常用的是概率法。这是因为概率法在大数据处理中生成概率,这就意味着将概率分类器与大型机器学习任务结合起来,从而减少数据的误差。

2.3 贝叶斯机器学习

贝叶斯方法在二十世纪五十年代以后逐渐形成,是概率理论和数理统计领域的一个主要内容。贝叶斯学习的基本原理是贝叶斯分析法,该方法在一定条件下,根据观测到的数据的概率和观测到的数据本身,给出了一个假定的可能性。贝叶斯学习是将未知参数与样本信息结合起来,利用贝叶斯方程得到后验信息,并在此基础上推导出未知参数。贝叶斯模型需要的参数很少,在属性相关度不高的情况下,其算法简单,具有很低的分类错误率和良好的综合性能。

贝叶斯算法的不足之处是,分类的整体概率分布和各种样本的概率分布往往是未知的,因此,要想得到更准确的整体概率分布,以及各种样本的概率分布,就必须对整体的了解和对样本的要求更高。贝叶斯机器学习是根据过去的事件的发生次数来预测未来的事件,而预测的结果则完全依赖于所收集的数据,而获得的数据越多,

就能得到更好的结果。大数据是大量数据的产生、存储和处理的主要技术，它能够为大叶斯机器学习提供大量的数据。贝叶斯机器学习与大数据技术相结合，在某些领域已有较好的研究与应用。

2.4 K最近邻(KNN)学习

K最近邻法是根据某种规律将相似的数据进行归类，就像是一句俗语：“人以类聚，人以群分”。K最近邻域学习法的基本原理是，要将新的数据进行分类，并将其与原有的数据进行对比，再从中选出K个最接近的样本，以确定哪个类别的数据最多。

K近邻学习法的缺点是，当某个类的样本容量很大时，大量的样本被吞噬。因此，在大数据环境中，如何更好地选择样本就显得尤为紧迫。现有的样本选择算法大多局限于小型数据，而对于大数据环境下的K最近邻学习算法的研究具有很好的应用前景。

2.5 回归学习

回归学习是一种从偶然中发现必然性的经典而又切合实际的预测。回归学习也被称作回归分析，它是一种在未知的随机抽样中求出目标函数的一种近似算法。变量间的关系可以划分为确定性与非确定性，前者具有明显的功能关系，而后的变量间存在着一定的联系，但又不能完全确定，在一定的随机扰动下会出现不能精确表达的统计关系。在此基础上，我们可以利用大量的实验来获取有关的统计资料，并构造出一种能近似此关系的目标函数，称为回归学习。

在回归学习中，需要分析的变量间有一定的相关性，而在适当的实验数据下，回归学习能够精确地反映出各因子的相关性和拟合的程度。回归学习中所用的因素和表达式仅仅是一种猜测，其拟合曲线在不同的资料下会有很大的差别。而在大数据环境中，使用回归算法，则要从海量数据中找出适合的变量和影响变量的因子。在大数据的支持下，如果变量的选取正确，综合变量的影响因子，回归学习的效果更好。

2.6 人工神经网络

ANN是由大量的隐含节点组成的一个数学模型，其特征是大规模并行处理，分布式信息存储，自组织学习。逆向传播算法(BP)是一种新的自适应学习方法。ANN理论上能够近似任何函数，其基本构造依赖于隐藏节点，因而具有较好的非线性映射性能。在不同的条件下，可以灵活地设定神经网络的中间层数、节点数量、节点的初始权重。ANN在医学、生理学、哲学、信息学、计算机科学等诸多方面都有很好的应用前景。

虽然ANN在某些方面的研究成果颇丰，但基于大数据支持的ANN尚处在起步阶段，存在着许多问题。如神经网络层数、节点数的确定、网络训练的效率等问题，尤其是在大量的数据环境中，数据具有较高的维度和不同的类型。而大数据正是解决这一问题的核心技术，它可以将大量的数据进行分布式计算和并行计算，从而为ANN提供更多的信息。

3.大数据环境下机器学习的主要研究方向

在大数据背景下，机器学习的发展主要有两个方面：一是对学习机理的研究，着重于对人的学习机理的仿真；第二，对如何有效地使用信息进行研究，把注意力集中在挖掘具有潜在价值和可被认识的基础知识上。在目前的大数据环境下，机器学习已经成为了各个行业的热点，机器学习可以帮助人们更好的吸收信息，从而掌握更多的机器技术。在当今的大数据时代，机器学习的重要性在于学习方式的选择，而机器学习将会成为学习和服务的一种新技术。在机器学习技术的基础上，如何对海量的数据进行高效的处理，是目前机器学习领域的一个重要课题。

在如今的大数据时代，无论是数据的种类还是数量，都发生了翻天覆地的变化，并且其生成的速度也越来越快。另外，数据类型的推陈出新，也使分析的困难程度不断增加，如文字情感分析、图像检索和理解、图像数据分析等。如此一来，机器学习的方向和学习方式就会变得更加多样化。如何有效地利用半监督学习来提高训练数据的质量，并在不同的知识环境中进行迁移。

为了进一步提高机器学习的效能，需要解决一系列可扩展性问题，也就是大数据问题的求解，这时应该采取平行分析的方式来分析这些问题；可视化分析、数据挖掘算法、预测分析能力、语义引擎以及数据品质与管理。

4.大数据环境下机器学习算法的趋势分析

4.1 提升机器学习的泛化能力

这是机器学习的一个重要方向，也是一个很常见的问题，很多产业都在努力提高它的推广性能。从目前的情况来看，支持向量机是一种将理论和实际相结合的综合型学习方式，它是一种从实践到理论的综合性学习方式。

4.2 提升机器学习的速度

在机器学习的各个领域，人们都在努力提高机器学习的效率，这是一个非常重要的课题。当前，如何正确处理机器学习速度测试和速度训练的矛盾，是当前研究热点。举个例子，K-近邻算法的测试速度一般比较缓慢，但是它的学习速度很快。

4.3 提升机器学习的可理解性

还有很多领域都在致力于提高机器学习的可读性，例如，在医疗领域，患者们想要知道为什么要这么做。目前，在这一领域，机器学习技术比较成熟，包括集成学习、神经网络和支持向量机等。

4.4 提升数据使用能力

过去的机器学习技术大多集中在已标注的数据上，但随着计算机技术的发展和数据的不断完善，越来越多的人面临着诸如垃圾邮件和医疗图像等问题。另外，在许多领域中，都会受到大量的数据不一致、属性缺失、大量的噪声和信息量的干扰，这些不均衡的数据往往会影响图像的正常使用，比如在乳腺癌的诊断和治疗中，患者的样本数量远远超过了健康的样本，这就造成了一个新的问题，就是如何将未经标注的数据和不对称的数据进行有效的处理，以提高数据的利用率。

4.5 提升处理敏感性代价问题的能力

在大数据时代，机器学习的主要目的就是减少错误的概率，但每个行业和专业对于错误的容忍程度都不一样，即使是同一领域，同一专业所需要的成本也是天壤之别。比如，在医疗癌症诊断和治疗方面，当一个患者被误诊为一个健康人或一个健康人而被误诊为癌症时，所付出的代价也是不同的。

以前的机器学习算法都是建立在成本均衡性的基础上，未来的发展方向，就是要提高对这些问题的处理能力，这些年来，很多专家都将医学分析和信号相关的理论应用到了机器学习算法中。

结语

它与诸如模式识别、个性化推荐系统、智能控制等机器学习技术相结合，在淘宝和京东等网上商店中具有典型的应用。从数据的提取、转换、装载等过程中，经过一系列的加工，最后转化为信息或知识，并以此作为判断的依据。在数据量不断增加的情况下，将数据收集、统计、分析的大数据应用到大数据运算中，使得机器学习的深度和广度都提高了数据分析的效率。大数据与机器学习有着密切的关系，两者之间有着密切的关系，大数据分析可以从海量的数据中发现隐藏的、逻辑关系的信息，并且做出正确的判断。大数据分析的大部分计算都来自机器学习，机器学习是对大量数据进行理论上的分析和优化，以达到数据分析的目的。

参考文献：

- [1]刘兴建.基于大数据的机器学习趋势分析[J].信息与电脑:理论版, 2019(13): 121-122, 125.
- [2]姜娜, 顾庆传, 杨海燕, 等.大数据下的机器学习算法[J].电脑与信息技术, 2019, 27(3): 30-33.
- [3]张素芳, 翟俊海, 王聪, 等.大数据与大数据机器学习[J].河北大学学报: 自然科学版, 2018, 38(3): 299-308, 336.
- [4]宋雯博.大数据下的机器学习的应用趋势[J].电脑迷, 2018(9): 158.