基于大数据的神经网络算法分析与研究

赵 洋
(国家计算机网络应急技术处理协调中心黑龙江分中心,黑龙江 哈尔滨 150000)

摘 要:随着互联网时代逐渐成为当下全球发展主流,大数据技术已逐步渗透到人们生活的各个方面,并促进了科学,经济和人文科学的快速发展。神经网络算法可以模拟如人脑的生物神经系统,是一种仿生学计算方法,是大数据关键技术的核心,也是将数据转换成知识、实现价值的重要举措。它已被广泛应用于医学医疗,智能语音识别和计算机视觉等领域,并且已经取得了一定的研究成果。该文主要基于大数据的特点和技术,与神经网络相结合进行分析和研究,探讨神经网络算法在大数据分析中的实际应用和核心科学问题。

关键词:大数据分析;神经网络算法;人工智能;智能语音识别;计算机视觉
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2021)0052-04

Analysis and Research of Neural Network Algorithm Based on Big Data

ZHAO Yang
(Heilongjiang Branch of National Computer Network Emergency Technology Coordination Center, Harbin 150000, China)

Abstract: As the Internet age has gradually become the mainstream of current global development, big data technology has gradually penetrated into all aspects of people's lives and promoted the rapid development of science, economy and humanities. The neural network algorithm can simulate the biological nervous system like the human brain. It is a bionic calculation method, the core of the key technology of big data, and an important measure to transform data into knowledge and realize value. It has been widely used in medical care, intelligent speech recognition and computer vision and other fields, and has achieved certain research results. Mainly based on the characteristics and technology of big data<combined with neural network for analysis and research, the practical application and core scientific issues of neural network algorithms in big data analysis is discussed.

Key words: big data analysis; neural network algorithm; artificial intelligence; intelligent speech recognition; computer vision

0 引 言

近年来,基于大数据时代的人工智能技术的发展迅速,数据每时每刻都在快速变化,容量也不断扩大,这使得互联网行业拥有更多的发展前景和挑战。大数据分析的关键是大数据技术的发展,数据在信息系统中的生命周期主要包含了数据的准备、存储、运算、分析和共享五个环节。大数据分析是将收集到的各种来源的原始数据进行价值转换的关键。其中数据分析是最重要的一个环节,需要在海量的复杂数据中提取出有价值的知识,它体现了数据可以创造价值的概念。传统的人工建模方法已经无法满足大数据分析的需求,需要进一步地研究新的分析方法。神经网络算法是数据分析中的核心,也是应用最广泛的技术。下面就神经网络算法在大数据分析中的相关基础理论和实际应用进行详细地探讨,进一步提出两者相结合所产生的一些核心问题。

1 大数据相关理论

1.1 概 念

大数据在 1997 年被科学家提出,是社会信息化快速发展的历史产物,大数据一般是指无数个体系包含在一起的数据集合,完全可以从大数据的四个主要特点入手,对其基本概念进行全面的了解[1]。这四个特点又称为“4V 特性”:

第一,数据海量,大数据最基本的特征就是拥有最完整和超高维的样本数据体系,少量的数据样本的集合不能称之为大数据,具有一定的整体代表性。

第二,大数据所包含的数据大多都是来自各种不

收稿日期:2020-10-15
作者简介:赵 洋(1979-),男,研究生,通信工程师,研究方向为网络通信、信息安全、网络安全。

同的渠道和传播途径,具有广泛性和多样化的特点,数据一般都处于集成化和模式非结构化的状态,例如图像、视频和语音等,不易于实现数据的分类和整合。

第三,数据每分每秒都在毫不停歇的生产、变化,这也就导致了大数据的容量也随之快速扩大且指数型增长,数据库每天至少生产至少是 TB 级别的用户数据和交易流水。

第四,大数据虽然拥有体量浩大的数据量,但是其中有价值的数据占比非常小,数据与数据之间的相关性强,这就意味着单个数据无法实现有用的价值。

1.2　大数据核心技术

1.2.1　数据平台

数据平台的主要研究工作是收集,标记,存储和整合各种数据,为数据分析平台提供基础,并确保大数据的计算和分析实现数据分析的可用价值。大数据的收集过程中应具有收集样品完整数据的能力,而不是样本数据的单个或少数几个,这样就可以保证将误差和准确率控制在想要的范围,保证数据分析的准确性。在大数据信息时代,应该如何实现“只存储技术知识而不是原始数据”的机制,可以大大提高存储管理效率和数据分析平台的工作学习效率。

1.2.2　分析平台

分析平台计算和分析处理后的海量数据。分析平台的主要工作是以平台的计算为基础,建设计算资源,设计相关的分析算法,这也是将原始数据转化为知识,使有效价值具体化的关键。传统的大数据分析方法是建立人工建模的方式进行分析数据,但是这种方法成本高、耗时、不易实现,还有很多应用缺陷,已经不再适用现在的数据分析发展要求了。另外一种方法是运用现在非常流行的人工智能技术分析大数据,其中最成功的技术就是神经网络算法。

1.2.3　展示平台

数据经过数据平台和分析平台的相应处理后,就由原始数据转化为知识,这就需要展示平台将产品知识推广出去,其中包括数据的研究规律和研究价值。在整个展示过程中,平台通过流量渠道对数据进行分析、标记、提取和推广。经过相应的分析过程后,数据主要表现为两种知识状态,即直接和间接知识;其中直接知识是在数据中发现特定的客观规律,使知识具象化。那么,所谓间接知识就是指可计算的分析模型,它完全可以应用到数据知识的获取过程中,最为典型的就是“举一反三”,比如在掌握一种技巧后,就可以将这种技巧进行分散化操作,应用到更多的事物中[2]。然而,重点是如何能够十分清楚地辨别这两种知识形态并科学的展示出去。传播和推广分析后的数据知识,有助于产品的形成,以进一步推动科学技术飞速发展,这也是在这一阶段要深入研究的问题。

2　神经网络算法相关理论

人工神经网络(ANN)也称为神经网络,它是一种模拟生物神经系统处理数据的特征的数学模型,它可以执行数据处理,分类和归纳,并降低容错率。ANN 具有学习和建模非线性和复杂关系的能力,这实际上是实时的,因为输入和输出之间的许多关系都是非线性和复杂的。一旦了解了初始输入和输出的关系,就可以在看不见的数据上推断出看不见的关系,使模型可以对未知数据执行预测。对于不完整、模糊、非线性等相关数据或知识,具有应用大数据技术的优势,效果显著,可以最大程度地解决与大数据分析相关联的问题。

神经网络是根据某些规则连接的许多神经元,最基本的是全连接神经网络,以下是其规则[3]:

(1)神经元从左到右按层次排列,包括输入层、隐藏层和输出层。输入层负责接收输入数据,输出层输出计算出的数据。中间的隐藏层对外部不可见。

(2)同一层中的神经元之间没有连接。

(3) N 层中的每个神经元都应与 $N-1$ 层中的所有神经元相连, $N-1$ 层中神经元的输出是 N 层中神经元的输入。

(4)相邻两层的神经元之间的连接具有一定连接权重。

2.1　神经网络算法模型

神经网络是智能技术开发领域的研究热点。如图 1 所示,其研究内容主要包括:模拟大脑神经网络结构,构建神经网络结构模型;模拟大脑神经网络的记忆机制,发展学习算法[4]。目前,研究实现的神经网络算法模型主要有:前馈神经网络、回复式神经网络、时序记忆神经网络[5]。

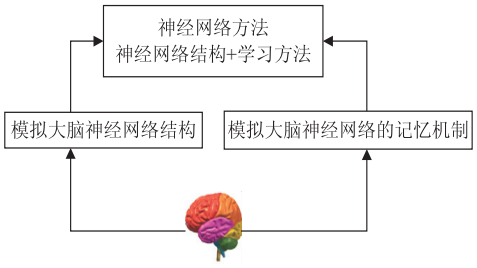

图 1　神经网络研究框架

前馈神经网络最主要的特点是学习能力强,可以训练大规模的数据集合,因此,前馈神经网络可在有效读取数据空间结构特征实现数据合理分类及综合计算方面发挥积极作用[6]。文中主要是对这种模型进行研究。目前被广泛应用于各种智能识别中,例如语音,图像等。感知机和深度自动编码机都是运用前馈型的神

经网络模型,被应用在手语识别技术和人脸图像的监测与识别,感知器和深度自动编码器等最著名的实现方法为计算机视觉、人工智能和医学医疗等领域的发展做出了杰出贡献。

前馈神经网络的数据传输是进行单向传播的,具有表达能力强的特点,这是由于其结构为多层神经元组织叠加在一起,而每一层组织又由多个神经元通过突触连接在一起,根据这个分布特点可以精确地逼近复杂函数。

2.2 深度神经网络

深度神经网络,一般是指具有一定深度且没有时间参数限制的前馈神经网络,拥有一定的层级限制,可以处理静态数据,学习和记忆模拟神经网络的机制,用数学方法组合,可以建立一个有效的神经网络的学习算法。现在最典型的两种神经网络学习算法是:Hebb 规则和反向传播算法。

2.2.1 Hebb 规则

Hebb 规则的主要思想是神经元之间的突触连接将随着突触两端神经元的同时激活而增强,否则被减弱。Hebb 规则不需要目标输出的任何相关信息,与“条件反射”作用机理相同,是一种无监督的学习规则。Hebb 学习率可以表示为:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha y_i(t) y_j(t)$$

其中,$w_{ij}(t)$ 可以表示神经元 j 到神经元 i 的连接权重,$y_i(t)$ 与 $y_j(t)$ 表示两个输出神经元。α 表示学习速度的常数。如果 $y_i(t)$ 与 $y_j(t)$ 同时被激活,也就是说,$y_i(t)$ 与 $y_j(t)$ 同时为正,则连接权重 $w_{ij}(t+1)$ 将增强。

2.2.2 反向传播算法

反向传播算法是一种迭代算法,以能够解决线性不可分问题被广泛研究和应用。

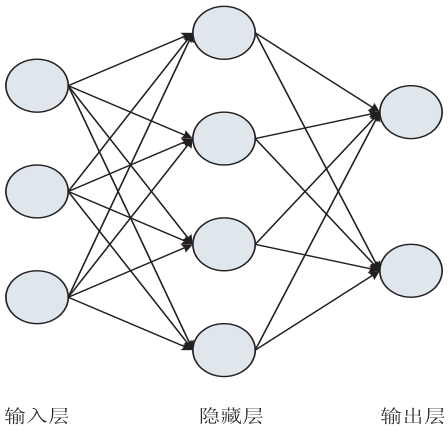

图 2 三层感知器实例

如图 2 所示,从数学优化理论出发进行神经网络训练,在多层神经网络研究领域中被广泛应用[7]。基本思想是:第一步,计算从第一层到最后一层的所有的状态量和激活值,因为信号是向前传播的;第二步,进行每一层的误差计算,这个过程是从最后一层进行计算的,这就是反向传播算法叫法的起源;第三步,更新参数,目的是减小误差;对第一和二步进行迭代操作,若相邻两次迭代之间的误差达到所需的要求时,迭代结束。BP 算法进行大数据分析时,用性能函数来体现网络性能,为高效率地优化网络性能和进一步学习网络,更新网络连接权重运用的是梯度下降法。在计算深度神经网络时,性能函数对相邻两个神经元连接权重的梯度计算比较复杂。

3 神经网络+大数据的实际应用

大数据的神经网络处理速度随着计算能力和硬件技术的发展有了显著效果。神经网络在大数据分析中的应用已经在当代人工智能、大数据、认知发展科学、神经科学等各个学科领域刮起研究热潮。

3.1 医学医疗

从预防到康复以及从政府到市场,以人为中心,观察人的健康已被认为是高质量的医疗服务,是最具挑战性的行业。健康的应用包括药品,医疗工具,初级产品,包括医疗保健服务,养老金服务,动态医疗保健,保健房地产领域(即养老金,医疗保健和健康金融部门)在内的健康服务领域,以及医疗保险以及其他各种经济服务,因此医疗技术应用的广泛、严谨和特殊性一直以来都是科学研究的先决条件。神经网络在大数据分析中的发展前景对促进医疗技术的发展有很大的裨益。充分利用大数据分析可以解决医疗方面的许多问题,改善医疗条件,进而提高病人治愈效率,在医学医疗领域获得了新的突破,并引起了全社会的关注。2018 年医学研究机构开发了一个医疗影像分析平台,其主要就是运用了卷积神经网络方法,在临床治疗过程中,可以应用新型的医疗影像中。

3.2 智能语音识别

语音识别是神经网络算法的最先获得显著成果的应用领域。在这一阶段,语音识别以声学和深度学习为基础,主要针对与识别语言相对应的文本进行开发,以提高语音识别的错误率。语音识别技术的基础在于人机交互,其目的是解决机器如何获取人类语言的内容,人工智能技术发展目前应用最广的就是智能语音识别。语音识别目前主要应用在 Siri 语音助手、微信、智能家居系统、音乐索检、车载智能语音交互系统和同声翻译等,大大地改善了居民的生活质量。

3.3 计算机视觉

随着现代化人工智能技术的逐步演变和应用,“深度学习”是计算机视觉研究的标准配置。借助神经网络算法,计算机视觉研究可以图像形式或模型形

式呈现,方便公众学习,识别,总结,区分。换句话说,计算机视觉研究的最重要的结合点是能够根据事物的不同特征实现计算机视觉研究的边缘化和规模一致性,并收集大量的图像,并使用特定的图像进行存储以表达各个方向的事物特征。使用以上功能,可以完成图像的分类、聚类、分割和目标检测、跟踪等计算机视觉任务,使用这些功能可以完成特定的计算机视觉任务。传统的图像特征一般是相对直观的初级特征,其具有较低的抽象度和较弱的表达自己能力,它主要是依赖于人工设计。神经网络是通过体量浩大的图像数据来表现学习特征,这个过程完全是自动化的,不受人为的操控。为使抽象程度大幅度提高,神经网络每一层的特征都划分成边缘、线条、轮廓、形状和对象等的层次。在日常生活中,旅客可以凭借自己的身份证通过智能识别过机场和高铁站等安检;警察在追捕犯人时,可以依据嫌疑人的指纹等相关信息与从大量人群所产生的数据库进行智能比对。大数据分析的神经网络算法在计算机视觉领域的应用已经深入的人们是日常生活中。

4　神经网络在大数据的核心科学问题

基于先前的概念和对神经网络算法以及对当前状况相关的大数据的分析,该算法在大数据分析中的应用存在一定的问题,限制了其应用。具体来说,神经网络与大数据分析相结合的应用所存在的核心问题主要包括三个方面:

4.1　数据的表达

表达效果是大数据分析的关键,因而数据如何很好地进行表达是第一个核心技术问题。原始数据是一个密集的集合,是由各种原始数据的详细信息组成。各种不同模式的数据信息都是十分杂乱的,其中内部主要包含数据噪声和数据缺失两个部分。在大数据分析的过程中,一个优秀的数据展示平台可以对数据的表达提供基础技术支持。神经科学表明,系统中的神经元在激活状态是稀疏的,因而,为了建立高维稀疏表达空间,神经网络方法需要注重学习训练样本数据的相关性。

4.2　数据的存储

在认知科学的研究领域中,人们主张生物神经系统只能对知识进行记忆,并不能存储各种最原始的信号,这也就意味着神经网络只能客观地记忆原始数据进行一系列转化后的知识,并不能存储原始数据,基于此,就引发数据存储问题,这也是需要解决的第二个核心科学问题。对于神经网络算法中的相关记忆机制,即内存的形式和学习的方法包括都是未知的,这需要进行重点研究。一般情况下,神经网络中相邻两个神经元突触连接强度会形成特定的网络记忆形式,即知识存储是通过神经网络连接权限的方法来实现功能的,但是在这个过程中连接权限需要随时调整。

4.3　数据的预测

大数据分析的预测是第三个的核心科学问题,是预测和展望神经网络算法与大数据相结合的应用前景。促进神经网络和大数据的发展为目标,逐步加强对认知计算原理的发展,深入研究基于神经网络算法的大数据预测。所以,在后期的方法开发中,加强对认知计算原理的相关应用,进一步开发基于神经网络算法的大数据预测,并且还要不断投入对该研究方面的成本投入,从而为大数据分析提供最为系统的软硬件支撑[3]。

5　结束语

随着大数据的发展,也促进了神经网络研究的发展趋势和完善,基于大数据的神经网络分析方法是一个比较系统化的工程,促进互联网技术、人工智能技术等科技领域的飞速发展。文中主要是对大数据的和神经网络算法的相关理论、实际应用和核心问题进行探讨,大数据借助神经网络强大的分析功能,可以实现数据转成知识,反过来神经网络借助大数据所提供足够的数据资源充分学习并发挥其功能。大数据和神经网络相结合的模式已经在许多应用领域中获得显著成果,从而推动了大数据分析的进程和革新。

参考文献:

[1] WANG Jun, LIU Mingzhe, TUO Xianguo, et al. A genetic-algorithm-based neural network approach for EDXRF analysis[J]. Nuclear Science and Techniques, 2014, 25(3): 18-21.

[2] 方　芳. 基于神经网络算法的大数据分析方法研究[J]. 软件工程, 2018, 21(9): 34-36.

[3] SHEN Junjie. Data processing and artificial neural network under big data[J]. Journal of Electronic Research and Application, 2020, 4(2): 10-15.

[4] 章　毅, 郭　泉, 王建勇. 大数据分析的神经网络方法[J]. 工程科学与技术, 2017, 49(1): 9-18.

[5] 周林腾. 基于神经网络算法的大数据分析方法研究[J]. 电子设计工程, 2018, 26(9): 19-22.

[6] 郭家超. 大数据分析的神经网络方法[J]. 科学技术创新, 2019(21): 67-68.

[7] 张　蕾, 章　毅. 大数据分析的无限深度神经网络方法[J]. 计算机研究与发展, 2016, 53(1): 68-79.