大数据背景下数据挖掘技术探析

赵波 高超 王晓菡
青岛海关技术中心，中国·山东 青岛 266114

摘 要：随着信息技术的飞速发展，大数据背景下数据挖掘技术成为研究热点。数据挖掘旨在从海量数据中提取有价值的信息和知识，其定义与目标是发现数据中的模式和关联。主要任务包括分类、聚类、关联规则挖掘等。在大数据环境下，新兴的数据挖掘技术、分布式数据处理、NoSQL 数据库、机器学习与深度学习技术以及实时数据分析技术等，共同推动了数据挖掘领域的发展。论文将对这些技术进行探析，以期为相关领域的研究和应用提供参考。
关键词：数据挖掘；大数据；分布式处理

Analysis of Data Mining Techniques in the Context of Big Data

Bo Zhao Chao Gao Xiaohan Wang
Qingdao Customs Technology Center, Qingdao, Shandong, 266114, China

Abstract: With the rapid development of information technology, data mining technology has become a research hotspot in the context of big data. Data mining aims to extract valuable information and knowledge from massive amounts of data, with the definition and goal of discovering patterns and associations in the data. The main tasks include classification, clustering, association rule mining, etc. In the big data environment, emerging data mining technologies, distributed data processing, NoSQL databases, machine learning and deep learning technologies, as well as real-time data analysis technologies, have jointly promoted the development of the field of data mining. The paper will explore these technologies in order to provide reference for research and application in related fields.
Keywords: data mining; big data; distributed processing

0 前言

在信息时代，数据已成为新的生产要素，其价值日益凸显。数据挖掘作为从大量数据中提取知识和信息的重要手段，对于商业智能、科学研究等领域具有重要意义。随着数据量的爆炸性增长，传统的数据挖掘技术面临挑战，促使了大数据环境下新兴技术的发展。论文将概述数据挖掘的基本概念、主要任务和流程，并深入探讨在大数据背景下，分布式处理、NoSQL 数据库、机器学习与深度学习技术以及实时数据分析技术的应用与发展，旨在为数据挖掘领域的研究者和实践者提供全面的视角。

1 数据挖掘技术概述

1.1 数据挖掘定义与目标

数据挖掘，就是从海量数据中，通过具体的算法与技术来抽取隐含，潜在有用信息与知识。目的是找出隐藏在数据之中的规律，趋势与关系，从而对决策起到强有力的辅助作用。在信息爆炸式发展的今天，各领域已经累积了大量数据，而数据挖掘像一把钥匙可以开启这些数据宝库。通过数据挖掘有助于企业深入理解顾客的需求、优化产品与服务、增强市场竞争力；在科学研究领域中，数据挖掘能够加快新发现的出现并揭示出自然现象背后所隐藏的法则；医疗领域帮助疾病诊断及治疗方案等。数据挖掘旨在把数据变成宝贵的知识并最大限度地发挥其作用。

1.2 数据挖掘的主要任务

数据挖掘主要工作有分类，聚类，关联规则挖掘和回归分析。分类任务为基于已知类别标签预测新数据。以垃圾邮件过滤为例，利用已经标注好的垃圾邮件与正常邮件的学习来判断新接收邮件的类别。聚类的核心思想是将数据对象分为多个不同的组或簇，这样可以确保同一簇内的对象具有高度的相似度，但在不同的簇中，对象之间的差异则更为显著。例如在进行市场细分时，依据顾客的种种特点把顾客划分为不同人群，从而使企业能够对不同人群制定个性化营销策略。关联规则挖掘的目的是在数据中找出不同项目间的关联，在超市购物篮进行分析时，找出哪些物品往往是共同购买的。回归分析被应用于预测数值型的变量，如基于房屋的面积和位置来预估房价。这些工作互相配合共同服务于从中发掘宝贵信息。

1.3 数据挖掘流程

数据挖掘通常由如下过程组成：数据的采集，从多种数据源中得到原始数据，其中可能有数据库，文件和网络。接下来是数据的预处理阶段，这一阶段至关重要，涵盖了数据的清理、去除含有噪声的数据以及处理数据中的缺失值等环节；数据集成是指从不同数据源中集成数据；为了使后续的数据挖掘算法能更有效地处理数据，我们进行了数据的转

换，包括数据的标准化和离散化等步骤。然后就是选择适合的数据挖掘算法进行研究，针对具体问题以及数据的特点选择了分类，聚类以及关联规则挖掘算法。算法执行时参数需不断调整优化才能得到较好的挖掘结果。对挖掘结果进行了评价与说明，并判断其是否有效可靠。若效果不满，需回到上一步调整。

2 大数据环境下的数据挖掘技术

2.1 大数据背景下的新兴数据挖掘技术

大数据时代下，传统数据挖掘技术受到了很多挑战，数据规模庞大，数据类型复杂多变以及对处理速度有较高要求等。所以新兴数据挖掘技术层出不穷。深度学习是其中的一种，其通过建立一个有多个级别的神经网络模型来自动地从大量数据中学到复杂特征的表示方法，并在图像识别，语音处理和自然语言理解方面获得了令人惊叹的研究成果。以图像识别为例，深度学习算法能够对各类对象，场景进行精确识别，显著提升图像分类与目标检测准确率。流数据挖掘是一个新兴的技术领域。随着物联网等技术的进步，大量的实时数据以流的方式不断产生。流数据挖掘可以对这些连续的数据流进行实时的分析和处理，从而及时发现异常情况和重大事件。在网络安全监测方面，数据挖掘能够迅速地发现网络上的攻击行为并对保护网络安全做出及时预警。也有图数据挖掘技术在社交网络和生物信息学中，数据通常都是以图的方式出现，图数据挖掘能够挖掘出图中节点之间的关系、社区结构和其他信息为了解复杂网络系统提供了强有力的支撑。这些新出现的数据挖掘技术，对大数据进行分析与处理，提供了有力的手段。

2.2 分布式数据处理技术

在大数据爆炸式发展的今天，一台计算机已不能满足人们对数据处理的要求，分布式数据处理技术由此产生。分布式计算框架，如 Hadoop、Spark，已经成为处理海量数据的主流手段。Hadoop 是一个开源的分布式系统基础架构，它包括分布式文件系统 HDFS 和分布式计算框架 MapReduce。HDFS 能够把大规模数据存储到多个节点中，达到了高可靠性、高可用性。MapReduce 的方法是将计算任务拆分为多个小任务，并在多个节点上并行执行，这极大地提升了数据处理的效率。Spark 作为大数据处理的快速普适性引擎，在内存计算上有较大优势，能够更快处理海量数据。分布式数据处理技术核心是对数据进行分布式存储与并行计算，该技术通过把数据离散存储到若干节点中，利用若干节点计算资源进行同步计算，能够对海量数据进行高效处理。

2.3 NoSQL 数据库技术

大数据环境中，传统关系型数据库逐渐暴露出其处理海量数据，高并发读写和灵活数据结构的局限性，NoSQL 数据库技术随之产生。NoSQL，也被称为“Not Only SQL”，并不是完全替代关系型数据库，而是为特定的大数据环境提供了一个创新的解决策略。NoSQL 数据库的种类很多，有键值存储数据库，列族数据库，文档数据库以及图形数据库。键值存储数据库采用键值对方式进行数据存储，读写性能极高，可扩展性强，非常适用于缓存、快速存储等简便数据结构。以 Redis 为例，其是目前应用较为广泛的键值存储数据库之一，该数据库能够将数据保存到内存中以达到对数据进行快速访问的目的，并且也支持对数据进行持久化处理以保证安全性。列族数据库以列为单位存储数据，适用于大规模分布式数据处理。它能够根据数据结构的变动，灵活地增加或减少列的数量。HBase 作为基于 Hadoop 分布式文件系统搭建的典型列族数据库可以处理大量结构化数据并具有高可靠性与高可扩展性。文档数据库是将数据存储在文档中，每一个文档都可含有不同字段与结构，十分适用于半结构化数据的存储。MongoDB 是一种流行的文档数据库，它支持丰富的查询语言和索引机制，可以方便地进行复杂的数据查询和分析。图形数据库主要致力于图形数据的储存和处理，例如社交网络中的人与人之间的关系、知识结构等。Neo4j 作为一个功能强大的图形数据库能够有效地实现图形遍历与查询，并能帮助用户在数据中找到复杂的关系。

2.4 机器学习和深度学习技术

大数据环境中，机器学习与深度学习技术正在扮演着日益重要角色。机器学习就是使计算机能够自动地从数据中学到规律与模式，可分为有监督，无监督与强化三类。监督学习就是通过使用带标签数据来学习从而对未知数据中的标签做出预测。以图像分类为例，对海量图像进行类别标签来使计算机学会如何依据图像特点来判断它们所属类别。常用监督学习算法包括决策树，支持向量机和随机森林。这些算法通过在大数据环境中不断地优化与完善，可以对大规模数据集进行处理，并且获得了很好的准确率。无监督学习的目的是在无标签的数据集中寻找可能的结构和模式。聚类作为无监督学习的典型方法之一，能够把数据分成不同簇，从而使相同簇内的数据相似。在顾客细分时，可通过聚类算法把顾客按照行为特征划分为不同人群，便于企业精准营销。像主成分分析和关联规则挖掘这样的方法也被认为是无监督学习中的关键技术。强化学习通过智能体和环境之间的相互作用，学习最优策略。智能体行动于环境之中，根据所获回报调整战略，最大限度地达到长远目标。强化学习已经在机器人控制，游戏和其他方面取得显著成效。深度学习作为机器学习中的重要分支之一，通过构造一个拥有多个级别的神经网络模型来实现对数据中复杂特征表示的自动化学习。深度学习已经在图像识别，语音处理和自然语言处理中获得突破。以卷积神经网络为例，它在图像识别领域展现出了卓越的性能，能够精确地鉴别各种不同的物体和场景。在处理连续数据时，循环神经网络展现出了明显的优越性，如在语音识别和自然语言处理中生成文本的能力。机器学习与深度

学习技术应用于大数据环境中的优势是其能自动地从大量数据中发掘出宝贵的信息与知识以辅助决策。技术同样存在一定的挑战。大数据规模大、复杂程度高，增加了模型训练与优化的难度，耗费了大量计算资源与时间。数据质量与标注问题同样影响模型性能。

2.5 实时数据分析技术

实时数据分析技术的重要性正在逐渐凸显出来，也是大数据背景下。在数据生成速度越来越快的今天，企业及组织要求能对其进行及时的处理与分析，从而迅速地进行决策、捕捉转瞬即逝的时机或者处理突发情况。实时数据分析技术，首先要求数据采集方法具有高效性。传统数据采集方式不一定能满足实时性需求，于是产生了流数据采集工具等一系列新型技术。这些工具能够实时从各种数据源（传感器、日志文件、网络流量）获取数据，并将其以流的形式传输到分析系统中。例如，在物联网的应用过程中，海量传感器会源源不断地生成数据，而流数据采集工具则能够实时地对这些数据进行采集，从而为之后的分析工作提供依据。在数据存储中，要支持实时分析就必须使用能快速阅读和写入的存储系统。内存数据库是人们经常使用的方案，这种方式把数据保存到内存之中，极大提高了访问数据的速度。部分分布式文件系统还进行了优化，可以支持快速随机读写操作以适应实时分析中数据存储需求。以 Apache Kafka 为例，它不但是消息队列系统而且还可作为数据存储介质来保存实时生成的流数据以进行后续分析处理。从数据分析算法上看，实时数据分析技术一般使用增量式算法与近似算法相结合。增量式算法能够在新数据抵达后无需对整个数据集进行再处理就能迅速地对分析结果进行更新。以在线机器学习为例，增量式学习算法能够依据新数据持续调节模型参数以达到实时预测与分类。近似计算法则旨在确保一定的计算精度的同时，通过牺牲一部分准确度来实现更高的计算速度。又如，在对大范围数据进行聚类分析时，可采用近似聚类算法快速获得近似聚类结果以适应实时分析需要。实时数据分析技术广泛应用于各个领域。在金融领域中，对股票市场数据进行实时分析，有助于投资者及时识别交易机会并作出迅速投资决策。再如，通过实时分析股票价格，成交量和其他数据，就能预测出股票价格变化趋势，从而为制定交易策略奠定基础。电子商务领域中，对用户行为数据进行实时分析能够实现个性化推荐以及精准营销。用户对网站中的物品进行浏览后，该系统能够实时地对用户行为特征进行分析，并向用户推荐与其感兴趣的物品，从而提高了用户购买的转化率。智能交通领域中，交通流量数据的实时分析能够对交通信号的控制进行优化，从而缓解交通拥堵问题。本实用新型通过设置在路面上传感器与摄像头对交通流量进行实时数据采集，分析系统能够根据交通状况对信号灯时间进行实时调节，从而提高路面通行效率。但是实时数据分析技术同样面临着一定的挑战。数据质量与精度是关键问题。鉴于数据需要实时收集和处理，可能会出现如噪声、数据缺失等问题，因此有必要采用高效的数据清理和预处理手段。实时分析系统要有高可靠性、高可用性才能保证各种条件下的正常工作。为了实现业务流程的自动化和智能化，实时数据分析技术还需与其他如数据采集、存储和决策支持等系统进行整合。大数据环境中实时数据分析技术有着巨大的应用价值与前景。伴随着科技的进步，实时数据分析技术会在更多的领域中扮演更加重要的角色，给企业与组织都带来了更大的价值。

3 结语

大数据时代的到来，为数据挖掘技术的发展带来了前所未有的机遇与挑战。通过论文的探讨，我们了解到新兴技术如何应对大数据环境下的挑战，并推动数据挖掘技术的创新与进步。未来，随着技术的不断成熟和应用领域的拓展，数据挖掘将在促进知识发现、优化决策过程等方面发挥更加关键的作用。持续关注和研究这些技术的发展，对于把握大数据时代的发展脉络具有重要意义。

参考文献：

[1] 彭三益.大数据挖掘技术背景下隐私权的特殊保护[J].求索,2023(3):170-178.

[2] 徐梦馨,高德立,柳景.大数据背景下数据挖掘技术在档案管理系统中的运用[J].信息与电脑(理论版),2021,33(2):28-30.

[3] 杨焱烨,梁晓庆.大数据背景下数据挖掘技术的应用[J].中国新通信,2020,22(6):105.

[4] 赫然,黄今慧.大数据背景下数据挖掘技术的算法[J].电子技术与软件工程,2019(20):141-142.

[5] 廖彦飞.大数据背景下档案数据挖掘技术应用探讨[J].城建档案,2019(8):30-31.