大数据时代数据挖掘与分析应用实践

刘子琪
中国人民大学统计学院　北京　100000

摘　要：在大数据时代下，以Hadoop为典型的NoSQL技术和以陈列式数据处理为典型的，以MPP NewSQL技术开始出现，推动了半结构化数据、非结构化数据的产生，以互联网企业为代表，在各种新的商业模式的试点下，促使人们逐渐进入到了大数据应用时代。在此时代下，人们对数据的应用需求越来越多，为了满足自身的发展需求，很多企业都进行了数据挖掘技术、挖掘算法、挖掘工具的研究和应用，并完善了系统化的数据挖掘技术方法。为了更好的应用研究成果，需要通过实际案例分析来学习、应用数据挖掘技术，对此本文主要浅谈大数据时代数据挖掘与分析应用实践。

关键词：大数据时代；数据挖掘；分析应用

Data mining and analysis and application practice in the era of big data

Ziqi Liu
School of Statistics, Renmin University of China, Beijing 100,000

Abstract: in the era of big data, with Hadoop as a typical NoSQL technology and display data processing as the typical, with MPP NewSQL technology began to appear, promote the generation of semi-structured data, unstructured data, represented by Internet enterprises, under the pilot of various new business model, prompting people gradually into the era of big data application. In this era, people have more and more needs for data application. In order to meet their own development needs, many enterprises have carried out the research and application of data mining technology, mining algorithms and mining tools, and improved the systematic data mining technology and methods. In order to better apply the research results, it is necessary to learn and apply the data mining technology through the actual case analysis, and this paper mainly discusses the application practice of data mining and analysis in the era of big data.

Keywords: big data era; data mining; analysis and application

引言：

随着大数据”的概念的出现，为很多企业都带来了一定的技术条件，可以高效化的处理海量数据信息，但是很多企业的员工，仍然习惯于使用原始低效的人工统计和分析方法，不仅浪费时间和资源，导致成本增加，也影响最终的结果。大数据时代要求企业的每个员工能够高效的处理日常的工作任务，并能系统的分析工作中存在的问题，不断改进工作，达成更佳的绩效。对此，企业需要把握大数据的知识背景，重视如何将数据分析的技能应用于企业日常的工作当中，了解大数据的概念，大数据在企业各项工作中的应用，学习数据挖掘方法，把握数据挖掘对营销带来的巨大价值和作用。

一、大数据时代下数据挖掘技术概述

数据挖掘技术是在上世纪出现的，其主要集中在人工智能产品研发和应用活动中，该技术还没有形成完整的体系，技术手段不成熟，相关设施不完善，相比于现代化的人工智能技术，有很大的不足。在大数据时代下，人工智能技术和数据挖掘技术之间的紧密对接的，需要在先进技术的支持下，通过机器深度学习算法对复杂、非结构化、不可控、无规律的数据信息进行深度挖掘和分析，以此获得深层的价值信息。从技术层面来看，数

作者简介：刘子琪（1991-05），女，汉族，北京，本科学历，主要负责概率论与数理统计专业的相关研究。

据挖掘技术可以深度挖掘数据内部价值信息，为企业提供更多的产品信息，推动企业的健康发展，当前数据挖掘技术主要处理一些随意性、信息内容模糊、无法识别的数据信息，通过高效、深度处理，进行更为精准、先进的数据价值分析和利用。总之，数据挖掘过程复杂、环节多、流程繁琐，但是在实践研究下也形成了多种方法，比如统计分析法、遗传算法、神经网络法等，这些方法各有有缺点，因此需要科学选择[1]。

二、大数据时代下数据挖掘方法

（一）统计分析法

在数据库字段中有以下两种逻辑关系：通过函数公式表示数据之间关联性的关系，对此需要根据数据库中的字段项关系特点，将关系明确的部分可以通过特定的函数公式定义Wie函数关系，对于不明确的和明确的之前没有相关关系。在确定相关认定标准后就需要通过统计学分析方法对关系之间的数据信息进行系统化的分析，一般可以借用统计学工具对数据的总和、最大值、最小值、平均值进行计算。并采用回归方程来表达数据之间之间的数量关系，在该统计学方法下可以深度挖掘数据价值，将数据信息作为样本信息，通过统计学计算，把握数据差异和共性特点，最终得出深层次的数据信息。

（二）遗传算法

该方法也是一种常见的数据挖掘方法，该方法随着基础数据挖掘工具的应用而不断发展成熟，数据挖掘攻击可以彰显数据实用价值，对此遗传算法也是一种基于自然选择和遗传因素下的大范围搜索方法，其融合性高，在后期实践应用下可以和神经网络算法、集合技术进行融合，成为最为广泛的一种算法。

（三）神经网络算法

对不同的数据挖掘方法进行对比可以发现，神经网络算法主要应用于数据挖掘中，其可以解决数据挖掘难题，应用广泛也非常广，该方法具有自适应性、高容错性的特点，可以有效进行处理、数据运行，对此该方法也是未来大数据应用领域的研究重点。

（四）粗集合方法

粗集合方法是在集合论的基础上形成的，是在数学理论的指导下产生的数据挖掘方法，因为数学理论的影响，该方法可以用来处理连续性数据，一般在获取信息表中的连续属性时需要结合其他方法共同使用，以此得到精准的内容。虽然该方法优势显著，但是其也具备一定的局限性，在采用该方法进行数据挖掘时不需要获取额外的信息，可以简单生成需要的数据信息，有效解决数据空间。总之该方法操作简单，应用效率高，也可以作为一种常用的数据挖掘方法应用[2]。

（五）决策树方法

该方法主要通过绘制决策树来描述数据，该方法可以有效对数据进行分类，操作简单，可以处理海量化的数据信息，在该方法出现后也推动了ID3算法的应用和发展，后在一系列实践探索下，在此基础上还形成了新的递增式的学习算法，该方法可以有效弥补决策树算法的不足，进一步彰显了该算法的优势。

（六）聚类分析算法

聚类分析算法是指在数据信息挖掘时需要根据不同的数据类型、特点进行划分，形成不同的组，后对组进行深度分析，该方法可以提高不同组别数据之间的关联程度，可以有效应用在客户信息挖掘中，当前该方法已经应用在了心理学、医学领域中。

（七）关联分析

人们可以从物体之间的关系性入手进行数据挖掘，不同的数据群之间有不同的分类标准，对此可以通过不同数据群数据之间的关联性分析数据元素和集合之间的因果关系，以此发现其中存在的问题和缺陷。当前该方法主要应用在企业产品精细加工中，可以通过挖掘消费者和产品之间的关系，帮助企业改变产品特点和功能，以此提高效益[3]。

三、大数据时代数据挖掘技术的应用

（一）MPP和Hadoop的结合应用

在具体应用大数据时需要采用多种架构，通过业务支撑系统有效发挥大数据技术优势，我国一些企业就将MPP和Hadoop进行结合，形成了混合式架构，并连接原有数据仓系统，有效体提高了大数据应用水平。在此过程中可以将传统数据库作为高价值数据，通过结构化数据的加工，实现长期结构化数据的存储和自助分析，Hadoop技术可以应用于非结构化数据处理、挖掘和历史存储。对于MPP技术，其是将传统分布式数据库的理论应用在相关产品中的实践，在列存、内存和副本优化后可以直接替代传统DW，但在大数据时代，该目标的实现还有待研究。总之，MPP技术可以精确地查明数据分布的原因，但是其可扩展性和高可用较低，根据CAP理论，系统需要不断优化升级才可以不断发展，我国较大的MPP集群有几十个节点，但是在国际上分有更多节点的集群。根据数据处理需求，需要不断增加节点，当前大数据的应用不单一是通过IT开发的，是一种自动更新和提升的行为，对此人们需要需要在MPP中要接入沙盒，依靠业务部门或第三方能自助地分析和开发。在此过程中不需要在每一个沙盒中接入物理的MPP集群，以

此确保后期检修维护和安装的便捷性，避免造成数据重复问题，对此需要借鉴云计算技术，为其提供按需服务，实现虚拟化目标。

（二）在移动领域中的应用

当在移动领域中的大数据应用主要表现在以下三个方面：一是让移动现有商业模式更加有竞争力；二是发掘新的商业模式，让行业运转更顺畅；三是承担社会责任，发挥大数据社会价值。当前移动DW/BI系统运转时间已经非常久，主要用于客户洞察、市场营销、客户服务和运营管理，在大数据应用过程中需要接入非结构化数据，以此深度应用，比如在客户洞察中，多种类型的数据，比如，消费、通话、位置、浏览、使用等，需要采用不同的算法，比如分类、聚类、标签、RFM、Pagerank，通过分析后可以形成客户视图，在不同的联系记录下自动形成社交网络，便于深入到客户交际中。也可以通过影响力分析，寻找关键人员，识别家庭和政企客户，以此明确重入网客户，把握客户各种情况，在移动销售多部终端后，促使TD-SCDMA芯片进入到了主流行列中，但是也导致企业面临着一系列压力。一般情况下，想要实现该目标，就需要通过贴营销成本来进行，但是这种方法也会随着企业利润的降低不被重视，对于这种问题，需要通过分析用户的终端偏好和消费能力，捉住终端机生命周期到期、合约机期满等时机，以此在不消耗营销成本的前提下顺利完成定制机销售任务[4]。

（三）在广告公司的应用

想要推动大数据应用的发展就是要转变观念，根据各个领域中的应用经验，需要遵循以下几个原则：利用大数据技术，收集整理数据，尤其是各种关联数据，并保存数据，将数据视作企业的核心资产；挖掘大数据价值，确保企业当前的商业模式更加具有竞争力；大力发掘新的商务模式，将数据转化成为潜在性价值。此外，也可以通过大数据技术对各个投放渠道的目标受众进行用户画像、整合媒体资源，以此创作出新的短视频，通过数据分析，促使内容与品牌的完美契合，最终精准广告投放，更精准有效的把握目标受众，最终帮助企业降低广告成本。

（四）在科研领域中的应用

随着科研活动的进行，很多科研成果都需要通过实践、研究、总结才可以获得，尤其是实验类科研活动，需要反复进行实验验证才可以得出有价值的关键数据信息，对于科研领域来说，数据价值巨大，数据内容非常重要。其中具体包括原始数据、失败数据、实验数据、新发现数据等，这些数据处理的好坏直接关系着科研活动的顺利进行，对此在进行数据处理时需要采用统计学方法进行分析。也可以采用数据挖掘技术根据科研项目基础数据，在对比后进行深度挖掘，以此降低成本，节省时间，提高处理效率[5]。

（五）在教育领域中的应用

随着教育事业的发展，各种大数据挖掘技术也被应用到了教育系统中，通过数据挖掘准确把握学生个人情况，在教学平台、学生电子档案下都可以通过数据挖掘技术把握学生各项素质，将不同学业阶段的成绩、表现、能力、素质等信息纳入到系统中。在数据处理和对比后可以为教师教学和学生管理提供数据依据和参考。

（六）在制造业领域中的应用

随着社会经济的发展，人们对于产品的质量和性能、用途也提出了新的要求，对于制造企业而言，想要健康发展，就需要把握消费者的个性需求，通过数据挖掘技术精准获取产品数据信息，通过分类整理后进行深度分析，把握市场产品的优缺点，从而不断优化产品。

四、结束语

总之，在大数据时代下，数据挖掘技术应用价值巨大，应用优势显著，应用广泛，技术方法较多，其在一系列理论研究和实践应用下已经发展成熟，应用效果显著，在未来数据技术的不断发展下，大数据挖掘技术会成为多个行业的主要转型方向，对此人们需要把握大数据挖掘技术的内涵和用途，应用原则，以此推动各个行业的发展。

参考文献：

[1]刘晓丹，张娜，王磊.大数据时代数据挖掘与分析应用实践——评《数据挖掘概念与技术》[J].科技管理研究，2021，41（20）：1.

[2]李涛.数据挖掘的应用与实践：大数据时代的案例分析[M].厦门大学出版社，2013.

[3]马堃，原博超，宫林娟.大数据时代妇科门诊临床科研云平台构建和数据挖掘的创新与实践[J].统计学与应用，2020，9（6）：15.

[4]张泽平.大数据时代的数据挖掘及应用问题分析[J].数字化用户，2018.

[5]李祥歌，王奇奇，郭轶博.基于大数据时代的数据挖掘及分析[J].电子制作，2015（3Z）：1.

[6]韩宣伟，蒋文军.基于灰关联分析的商业地产数据挖掘与批量评估技术应用[C]//挑战与展望——大数据时代房地产估价和经纪行业发展论文集.2013.