大数据时代下数据挖掘技术的应用

刘铭,吕丹,安永灿

长春工业大学数学与统计学院,长春 130012

摘要　大数据时代下,数据挖掘技术越来越受到人们的关注。本文介绍了数据挖掘技术的研究背景和研究现状,论述了决策树、支持向量机、神经网络等数据挖掘技术的相关算法,分析了数据挖掘技术在大数据中的相关应用及未来的发展趋势,探讨了在大数据时代数据挖掘技术面临的挑战。

关键词　大数据;数据挖掘;决策树;支持向量机;神经网络

随着社会信息化的迅速发展,无论是数据的变化速率,还是数据的新增种类都在不断更新,数据研究变得越来越复杂,这意味着“大数据时代”到来。2011年,互联网数据中心(internet data center,IDC)将大数据重新定义为:在大数据原有的三维特征——数量、多样、速度基础上,增加了另一新的特征——“价值”。IDC强调:“目前,对于庞大的数据量,通过经济的方式,极速发掘、获取和分析处理的技术,进而提炼获取价值,这是大数据新时代的专属。”“大数据时代”的专属特征被重新定义为:数量(volume)、多样(variety)、速度(velocity)和价值(value),称为“4V”。

随着大数据时代的到来,社会对“挖掘”到的数据要求变得更加严格,每一个精准的结果都具备独自的“价值”,这时,大数据时代的新增属性——“价值”被演绎得有声有色。数据挖掘(data mining, DM)是一门新兴的、汇聚多个学科的交叉性学科,这是一个不平凡的处理过程,即从庞大的数据中,将未知、隐含及具备潜在价值的信息进行提取的过程。1989年8月,在美国底特律市召开的第十一届人工智能联合会议的专题讨论会上,知识发现(knowledge discover in database,KDD)初次被科学家们提出,同时,也有人将知识发现称为数据挖掘,但两者并不完全等同。1995年,KDD这个术语在加拿大蒙特利尔市召开的第一届知识发现和数据挖掘国际学术会议上被人们接受,会议分析了数据挖掘的整个流程。实质上,数据挖掘是知识发现的子过程[1]。

经过了大约20年的发展,数据挖掘研究取得了可观的成绩,渐渐地形成了一套基本的理论基础,主要包括:分类、聚类、模式挖掘和规则提取等[2]。数据挖掘是一种从生活中的海量数据里“挖掘”出潜在的、前所未有的知识的技术。处理大数据需要一个综合、复杂、多方位的系统,系统中的处理模块有很多,而数据挖掘技术以一个独立的身份存在于处理大数据的整个系统之中,与其他模块之间相辅相成、协调发展[3]。在大数据时代中,数据挖掘技术的地位是无可比拟的。

1　数据挖掘的研究现状

数据挖掘将高性能计算、机器学习、人工智能、模式识别、统计学、数据可视化、数据库技术和专家系统等多个范畴的理论和技术融合在一起。大数据时代对数据挖掘而言,既是机遇也是挑战,分析大数据,建立

收稿日期:2017-12-15;修回日期:2018-04-09

基金项目:国家自然科学基金项目(61503150)

作者简介:刘铭,副教授,研究方向为智能计算与数据挖掘,电子信箱:jlcclm@163.com

引用格式: 刘铭, 吕丹, 安永灿. 大数据时代下数据挖掘技术的应用[J]. 科技导报, 2018, 36(9): 73-83; doi: 10.3981/j.issn.1000-7857.2018.09.010

适当的体系,不断地优化,提高决策的准确性,从而更利于掌握并顺应市场的多端变化。在大数据时代下,数据挖掘作为最常用的数据分析手段得到了各个领域的认可,目前国内外学者主要研究数据挖掘中的分类、优化、识别、预测等技术在众多领域中的应用。

1)分类。

伴随着时代的进步和科技的飞速发展,作为人口大国,中国在健康医疗、老龄化社会等方面产生的公共数据呈几何级数进行增长,而基于大数据的挖掘数据所附有的价值问题急需解决。健康医疗数据的结构、规模、范围和复杂度等都在不断扩大,传统的计算方法并不能完全满足分析医疗数据,数据挖掘技术则可以根据医疗数据的一些特点:模式的多态性、信息的缺失性(数据中由于涉及个人隐私问题而导致的缺失值)、时序性、冗余性对健康医疗数据进行分类,从而可以为医生或病人提供准确的辅助决策[4]。

同时,中国正加速进入老龄化社会,而互联网是改善老龄化社会的重要媒介,大数据是评估老龄化社会重要的技术手段。屈芳等[5]提出了"互联网+大数据"模式的养老实现途径,整个养老服务体系是建立在多元异构信息汇聚和数据融合挖掘之上,"互联网+大数据"的养老体系是将多种信息通信技术进行融合,在这里,包括通信技术、数据挖掘技术及人工智能技术等。

2)优化。

道路的交通状况与人们的出行关系密切,随着城市的快速发展、生活水平的改善,机动车的规模也逐渐扩大,带来了交通拥堵等问题。数据挖掘技术可以有效解决交通道路和物流网络之间的优化问题,Pan等[6]提出了一种数据挖掘预测模型,该模型用于"实时预测"短期的交通状况,给陷入交通拥堵的驾驶人员带来极大的帮助。

随着科技的发展,网上购物越来越流行,同时带来了物流运输拥堵及瘫痪等问题。京东——中国最大的在线交易平台之一,在人工智能的优化时代,使用无人机探测道路状况反馈的数据,采用数据挖掘技术精准计算物流网络运输所需要的参数,可以轻松高效地缓解物流运输瘫痪的问题,从而产生了中国第一个机器人快递员,将第一个商品送达至中国人民大学。而随着日后交通网络长度、复杂性等方面的增加,实现无人驾驶的自动化策略难度也大幅增加,只有通过数据挖掘技术才可以快速计算出结果,从而获得从复杂道路信息中产生的高效价值。

3)识别。

自从20世纪50年代数字图像出现以来,数字图像成为人类社会中必不可少的"数据"。在计算机应用中,数据挖掘在图像识别的应用越来越普遍,有代表性应用为人脸识别和指纹识别。人脸识别通过对获得的信息库进行数据挖掘,进一步分析和处理可靠的、潜在的数据,充分准备资料的分析工作和未来的开发工作。Wright等[7]阐述了基于稀疏表示的鲁棒人脸识别,并给出了详细的理论分析与实践总结。

沙亚清等[8]针对目前的电子报税系统中利用用户名和口令的不安全性,提出了一种基于智能卡和指纹识别的身份认证方案,并结合指纹技术,构建新的口令参数,从而使得安全性明显提高。随着数据挖掘技术的不断发展,大数据识别人脸和指纹的精确度会越来越高。

4)预测。

预测问题是各领域中研究最多的问题,其目的是通过历史数据预测出未来的数据值或发展趋势。大部分历史数据是时间序列数据,即指按照时间的顺序排列,得到了一系列观测值。由于信息技术的不断进步,时间序列的数据也日益剧增,如气象预报、石油勘探、金融等。时间序列数据挖掘的最终目标就是通过分析时间序列的历史数据,预测未来一段时间的变化趋势及其带来的影响。

"气象"与地球的生态平衡和人们的正常生活息息相关,因此,气象的准确预报显得格外重要。周磊等[9]总结了目前的气象监测模型,基于遥感数据的干旱方面,将目前的遥感监测方法进行分类,对于外界的环境条件(温度、湿度等)进行分类讨论,提出解决复杂问题的新方法。

石油作为一种不可再生资源,目前全球储量日益减少,从而使得石油勘探变得越来越重要。在石油勘探管理中,所采集的数据具有数据量大、计算量大、采集来源单一及数据处理流程复杂的特点[10],用数据挖掘技术对其采集的大数据集进行高性能并行计算和分析,才可以保证结果的有效性和准确性。

在大数据时代下,银行、证券公司、保险公司等每天的业务都将生成海量数据,采用当前的数据库系统可以高效地实现数据的录入、查询和统计等功能,目前,从简单的查询提升到利用数据挖掘技术挖掘知识、

提供决策支持的层次显得格外重要。数据挖掘技术在金融行业应用具有可行性,将理论基础应用到相关的实例包括预测股票指数、发现金融时间序列中的隐含模式、信用风险管理及汇率预测等。

2　数据挖掘主要方法

数据挖掘是一门交叉性的新兴学科,它将数据可视化、数据库技术、高性能计算机、统计学、机器学习、模式识别、人工智能等多个范畴的理论和技术融合在一起。数据挖掘的主要方法概括为:预测模型方法、数据分割方法、关联分析法和偏离分析法(图1)。解决实际问题时,将已知的数据库蕴含的复杂信息转换成数学的语言,建立数学模型,运用相应的处理方法结果会更加有效。

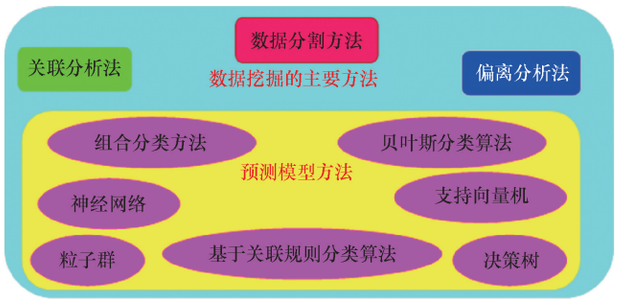

图1　数据挖掘的主要方法

Fig. 1　Overview of main methods of data mining

2.1　预测模型方法

预测模型方法是数据挖掘主要方法中分支较为复杂的一类,包括神经网络与决策树等相关人工智能算法、进化算法及支持向量机等算法。

2.1.1　神经网络与决策树等相关人工智能算法

在预测模型方法中,神经网络算法、决策树算法、贝叶斯分类算法、基于关联规则分类算法等都是经典的人工智能算法。

1943年,心理学家McCulloch和数理逻辑学家Pitts建立了神经网络和数学模型,称为MP模型,证明了单个神经元能够执行逻辑功能,从而开创了人工神经网络研究的新时代[11-13]。通过仿真和模拟生物的神经系统而获得非线性处理能力的一种新的算法——人工神经网络算法(artificial neural network,ANN)。

现有的决策树的分类算法有ID3、C4.5[14]等。1986年,Quinlan提出了著名的ID3算法,在ID3的基础上,1993年Quinlan又提出了C4.5算法[15-17]。决策树(decision tree,DT)分类算法是一种以决策树形式表示的分类规则,它能够根据一定的规则将众多的数据分类,从中挖掘出那些有价值的、潜在的信息。决策树[18]的主要优点在于处理大数据的能力强,适合分类及处理预测模型的任务,结论易于解释和理解。

目前的主要研究有3种:CBA、CMAR[19]和CPAR[20]。自1993年Agrawal提出数据库中的关联规则挖掘后,基于关联规则分类算法[21](classification base of association,CBA)及应用得到迅速发展。1997年,Ali等提出了使用分类关联规则进行部分分类的思想。1998年,Liu等[22]提出了基于分类关联规则的关联分类算法CBA,从此揭开了关联分类的序幕。基于关联规则分析的分类算法搜索频繁模式与类标号之间的强关联,有效避免了决策树归纳一次只考虑一个属性的限制,使其比一些传统的分类算法更为准确。

贝叶斯(Bayes)[23-25]分类算法是一种算法相对比较简单、分类精度相对较高的分类算法。在分类的性能方面,决策树算法、贝叶斯分类算法及神经网络算法之间关系十分紧密。现有的贝叶斯分类算法包括朴素贝叶斯算法、动态贝叶斯算法等。常见组合分类方法有随机森林方法、bagging方法及boosting方法。其中,随机森林方法是将多个决策树分类器组合在一起的方法,在boosting算法中最常见的一种是AdaBoost算法。在准确度上,二者不相上下,但是,在运行速度上,随机森林方法更占优势。朱凌云等[26]提出了一种新的技术并在医学中的应用,体现了数据的处理、多属性信息的融合、挖掘算法的高效性和鲁棒性。由于神经网络系统具有高度的抗干扰能力,所以,在各个领域内神经网络算法应用广泛,例如数据挖掘、信号处理、自动控制、模式识别及图像处理等多个范畴。

2.1.2　进化算法

进化算法,又称“演化算法”(evolutionary algorithms,EAs),其代表性算法为遗传算法。1969年,Holland[27]提出了一种随机搜索的最优化方法,它是模拟自然界中的遗传机制和生物进化论而成的,称为遗传算法(genetic algorithms,GA)。它将利用自然界中的“优胜劣汰,适者生存”的生物进化原理改变优化参数,根据适应度函数的选取,最终形成编码串联到群体中。遗传算法的基本步骤:选择、交叉和变异。遗传算法的

主要目的是留下适应度值好的个体,淘汰适应度值差的个体,继续循环选择、交叉和变异步骤。

近几年,又演化出新的进化算法,如粒子群算法、蚁群算法以及灰狼优化算法等。粒子群算法(particle swarm optimization,PSO)是由Eberhart等[28]开发的一种新的进化算法。与模拟退火算法相似,PSO算法也是从随机解出发,通过迭代进而寻找最优解,与上述的"遗传算法"相比而言,规则更为简单,它没有遗传算法基本步骤中的"交叉"和"变异",而是通过追随当前搜索获得的最优值来寻找全局的最优解。粒子群算法以实现简便、精度高、收敛快等优点引起了学术界的重视,并且在解决实际问题中展示了其优越性。

2.1.3 支持向量机

1995年,Corinna和Vapnik等首先提出了支持向量机(support vector machine,SVM)[29-31],它是一种具备较强的分类能力和泛化能力的分类算法,主要解决小样本、非线性、高维模式识别及函数拟合等其他机器学习问题。支持向量机主要分为以下3种情况。

1)线性可分情况。

针对线性可分的情况,现实生活中存在大量的实例,例如,在一组医疗数据中,通过支持向量机可以将患者和正常人进行分类(即二分类),判断哪些是患者,哪些是正常人;在一组由民歌和古筝演奏的音乐辨别中进行有效的分类,判断哪些是民歌,哪些是古筝。

2)线性不可分情况。

解决线性不可分问题时,构建核函数,这是支持向量机的优势所在。但是,对于数据集训练的"复杂度"最终还是取决于它的规模,在处理大规模数据时,模型局部受限,泛化能力有时也会有所消耗或损失。

3)非线性可分情况。

支持向量机利用结构风险最小化替代经验风险最小化原则,较好地解决了小样本情况下的学习问题。针对非线性问题与线性问题是怎样建立起联系的,它们之间是如何进行转化的,"核函数的思想"提供了新的思路。

2.2 数据分割方法

数据分割是将数据依据某些属性将其聚类,使之具有一定的意义。由于数据的类型、数据的复杂度和聚类的数目等特点,聚类算法有很多,如划分方法、基于网络的方法、基于密度的方法、层次方法等。

肖娟等[32]针对传统的算法处理多层次的复杂建筑物中涉及的困难,提出了一种新的算法,对建筑物进行分割,对几何基元进行提取。

2.3 关联分析法

关联分析法是寻找数据间的关联,但从大数据集中寻找关联可能会导致效率降低,找到的关联也可能毫无意义。在研究过程中存在"支持度"和"置信度","支持度"可以有根据地将那些毫无意义的数据删除,而"置信度"可以衡量设置规则的可能性。关联分析法的主要算法有Apriori算法、DHP算法和DIC算法等。

Chen等[33]在现有的分析方法基础上,积累了海量的数据,利用数据挖掘技术,提出了一种新的算法,即通过关联分析法建立相关模式挖掘方法,借助多种新型优化技术,可以有效且高效地减少搜索空间。此外,将该算法应用于现实世界的数据集中,展示了相关模式挖掘的实用性。

2.4 偏离分析法

偏差包括潜在的信息量,例如设定模式中的特例、分类中的异样实例以及分析实验得到的最终结果与实验前设定的期望之间的偏差等。观察比较最终的结果与参照量之间的偏差是偏离分析法的核心所在。

在企业的预警或是危机解决的过程中,专业的管理者对突发的意外规则更感兴趣,在异常信息的发现、识别、观察、分析、挖掘、评价和预警等方面,挖掘意外规则的应用价值备受关注。

3 大数据时代下数据挖掘的应用

在大数据时代下,数据挖掘已经广泛地应用到生活中各种各样的领域中,成为当今高科技发展的热点问题。无论在软件开发、医疗卫生方面,还是在金融、教育等方面都可以随处看到数据挖掘的影子,可以使用数据挖掘技术发现大数据的内在的巨大价值。

3.1 恶意软件的智能检测

在大数据时代下,在恶意软件检测中数据挖掘技术得到广泛的应用。恶意软件严重损害到网络和计算机,恶意软件的检查依赖于签名数据库(signature database,SD)[34-36],通过SD,对文件进行比较和检查,如果字节数相等,则可疑文件将被识别为恶意文件。有些基于有标签的恶意软件检测的主题,集中在一个模糊的环境下,进而,无法进行恶意软件行为的动态修改,无法识别隐藏的恶意软件。相反地,基于行为的恶意软

件检测就可以找到恶意文件的真实行为[37-39]。而如果采用基于数据挖掘技术的分类方法,就可以根据每个恶意软件的特征和行为进行检测,从而检测到恶意软件的存在。

3.2　生物信息学中的广泛应用

生物信息学是一门交叉学科,融合了生命科学、计算机科学、信息科学和数学等众多学科。随着科技的快速发展、技术的提升及结果的优化,将高科技信息技术拓展到生物研究领域。但是,单纯凭借原有的计算机技术是远远不够的,需要以计算机科学做辅助,将生命科学、信息科学和数学等交叉学科融合在一起,通过数据挖掘技术进行处理,仔细分析生物数据之间的内在联系,挖掘生物数据内部的潜在信息。生物信息数据的特点有很多,孙勤红[40]总结了当前生物信息数据的特点,包括数量大、种类多、维度高、形式广及序列性等。当前生物信息学的热点包括[41]:从以序列分析为代表的组成分析向功能分析的转变;从单个生物分析的研究到基因调控的转变;对基因组数据进行整体分析等。人类目前在生物基因组计划中的研究,仅仅是冰山的一角,未来在差异基因表达、癌症基因检测、蛋白质和RNA基因的编码等生物基因方面的研究工作都与数据挖掘技术密不可分,只有更好地利用数据挖掘技术,才可以挖掘出生物基因组中的非凡价值。

3.3　信用卡的违约预测

如今,随着科技的高速发展,信息量急剧增加,内容变得越来越丰富,信用卡在人们的生活中具有不可忽视的地位。众所周知,信用卡是由银行发放,银行需要对申请人的个人信息进行核实,确认无误后再进行发放信用卡,Chen等[42]针对商业银行贷款行为提出了一种关于信用率的模糊算法。信用卡在办理之前,银行首先需要对申请人进行细致调查,根据申请人的实际情况判断是否有能力来偿还所贷金额,刘铭等[43]在传统的神经网络基础上,采用灰狼优化算法计算神经网络的初始权值和阈值,并提出了一种改进的模糊神经网络的算法,通过建立的信用卡客户的违约预测模型,与目前其他的预测方法进行比较,得到较好的预测结果,进一步,验证了模糊神经网络在信用卡客户的预测上具有较好的鲁棒性、准确性和高效性。采用有效的数据挖掘技术,针对信用卡客户属性和消费行为的海量数据进行分析,可以更好的维护优质客户,消除违约客户的风险行为,为信用卡等金融业务价值的提升提供了技术上的保障。

3.4　疾病的智能诊断

1)宫颈癌的诊断。

宫颈癌是国际上最普遍的妇科恶性肿瘤之一。2012年统计数字显示,宫颈癌在全球的新发病例数为52.8万,死亡数26.6万,居女性生殖道恶性肿瘤发病率的首位。按照有关数据统计,发展中国家占83%,其中死亡病例占85%,由于宫颈癌的筛查工作不够完善,导致高发病率和高死亡率。相反地,在发达国家,很大程度上宫颈癌的低发病率源于有效的筛查和诊断。为了减少来自每个专家的标签数据量,Fernandes等[44]提出一种基于正则化的转移学习策略,鼓励源模型和目标模型共享相同的系数符号。

2)乳腺癌的诊断。

乳腺肿瘤是女性恶性肿瘤中最常见的肿瘤,影响妇女的身体和精神健康,甚至威胁生命。20世纪以来,全世界范围内乳腺癌的患病率均有所增加,特别是欧洲和北美地区,分别占欧洲和北美女性恶性肿瘤发病率的第一和第二位。目前,世界女性乳腺癌在癌症中的发病率最高,据美国疾病预防中心统计,早期乳腺癌的治愈率可高达97%,进展期的治愈率仅为40%[42]。因此,越早发现乳腺癌,治愈效果越好,即“早发现,早治疗”。

在大数据时代下,医疗方面的数据呈现出数量大、类型多、处理方法复杂等特点,数据挖掘技术对这些问题的处理起到了至关重要的作用。威斯康星大学医院Wolberg提供的乳腺肿瘤分析结果显示,乳腺肿瘤的特征可以由9个参数来表示[44]。基于改进的BP神经网络,刘铭建立了乳腺肿瘤的模拟模型[45],对传统的BP神经网络进行改进和发展,当Levenberg-Marquardt(L-M)迭代替代了梯度下降算法时,网络收敛速度得到了明显的提高。

使用Matlab2010a进行求解,采用L-M迭代后,目标误差为0.1,得到结果。通过图2[46]可知,神经网络在第7代达到收敛。测试数据有83个样本。其中良性54例,恶性29例。采用检测资料进行检测,诊断结果为良性54例,良性发生率100%,恶性28例,恶性发生率96.6%,所以平均诊断发病率为98.8%,结果良好。

3)冠心病的诊断。

近年来,心血管疾病已成为威胁人类的最严重疾病之一,冠心病是心血管疾病中常见的疾病。因此,研

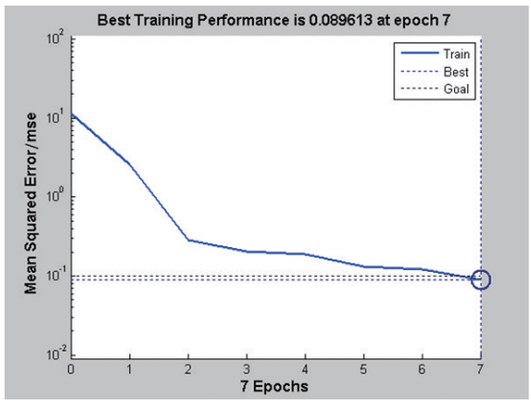

图2 神经网络训练性能

Fig. 2 Neural network training performance

究冠心病的有效诊断方法是必要的,有助于进一步采取预防措施和及时治疗。目前,冠状动脉造影是观察冠状动脉形态的唯一直接途径,被医学界称为“金标准”[47]。然而,这是一项创伤性诊断,需要高水平的医疗条件,否则不慎操作会引起严重并发症甚至死亡,这限制了诊断技术的发展。因此,许多专家专注于研究国内外冠心病的有效和非创伤性诊断[48-49]。经对Cleveland诊所基金会提供的冠心病病例分析后,刘铭得出了反映冠心病特征的14个参数,采用BP算法,通过使用L-M算法的迭代对BP算法进行改进和开发,提高了网络收敛速度,在改进的BP算法的基础上,建立了智能诊断的仿真模型。随着该方法的应用,诊断率可达99.3%[50]。

针对疾病的智能诊断,数据挖掘具有4个应用角度:在医院信息系统中的应用、在疾病辅助诊断中的应用、在药物开发中的应用、在遗传学方面的应用。

3.5 地质灾害的风险评估

地质灾害研究具有悠久的历史,地质灾害风险评估是一个新兴的研究领域。近年来,在某些领域已经开发出更准确的预测和分析的方法,这些领域涉及到坍塌、地震、山体滑坡和泥石流等地质灾害。

刘铭提出了一种新颖的智能计算方法,将数据挖掘技术与地质灾害风险实际问题融合在一起,这种混合计算方法促进了对地质灾害风险的准确评估。混合智能算法包括粒子群优化、遗传算法和反向传播神经网络。反向传播神经网络和粒子群算法优化了网络连接权重,阈值的初始化采用遗传算法,同时,在迭代过程中更新连接权重和阈值。这项地质灾害预测研究是在吉林灾害监测数据的基础上,模拟中国东北地区,通过混合智能算法获得的准确度远高于BP神经网络方法带来的准确度。随着地质灾害风险评估在国际风险评估机构中得到肯定,混合方式得到更广泛的应用,如混合智能算法将促进更有效的应急响应、环境管理、土地利用和开发规划[51]。

3.6 污水的成因分析

在大数据时代的背景下,当研究水环境和污水处理时,生物膜的组成和活性是两个非常重要的参数。而处理污水问题时,面对的数据海量,单一的传统数学方法解决效果不够理想,引入数据挖掘技术进行分析,问题优化的结果将会更令人满意[52]。

研究水环境的重点在于对污水处理、运行和控制方面的实际需要,通过数据挖掘技术可以准确找到生物膜的表征和活性,并进行估计,进而对于参数不足以描述生物膜活性的问题得以解决。

在给定的限度内,随着生物膜的厚度增加,生物膜的活性也随之增强。测量或估计生物膜厚度和活性的方法是评估生物膜废水处理效率的重要因素,然而目前用于预测生物膜厚度和空间分布适应性的工具较差。对此林山松等[53]基于碳-氮-磷浓度的空间分布生物膜厚度和活性,提出了支持向量回归模型,用以预测反应器中的生物膜的厚度和活力。

采用共聚焦激光扫描显微镜方法对12个样点的4个随机位置上形成的成熟生物膜的厚度进行估算,并将其平均值作为每个载体的最终厚度。图3[53]为共聚焦激光扫描显微镜的微图,展示了在运行100天后载体上的典型生物膜的厚度,其中Z轴上的数字(30.6 μm)是由激光共聚焦显微镜测量的生物膜厚度。得到的数据作为观测值来估计反应器中未被采样点的生物膜厚

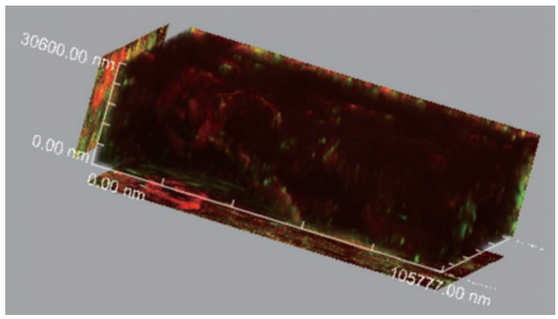

图3 用于检测生物膜厚度的激光共聚焦显微镜显微照片的例子

Fig. 3 Example of a laser confocal microscope photomicrograph for detecting biofilm thickness

度,这些未被采样的点的生物膜厚度通过使用Kriging插值得到。

基于实际值的Kriging插法和距离反应器底部垂直35 cm处的生物膜厚度和生物膜活性的支持向量回归模型预测值进行了比较。图4[53]比较了使用支持向量回归模型的生物膜厚度和生物膜的活性的实际值和预测值。结果表明较高的系数R^2=(0.996,0.997),并且通过支持向量回归基于碳-氮-磷值在碎石球状骨料反应器中预测生物膜厚度和生物膜活性的高度可行性,同时根据实际值验证Kriging插值的准确性。

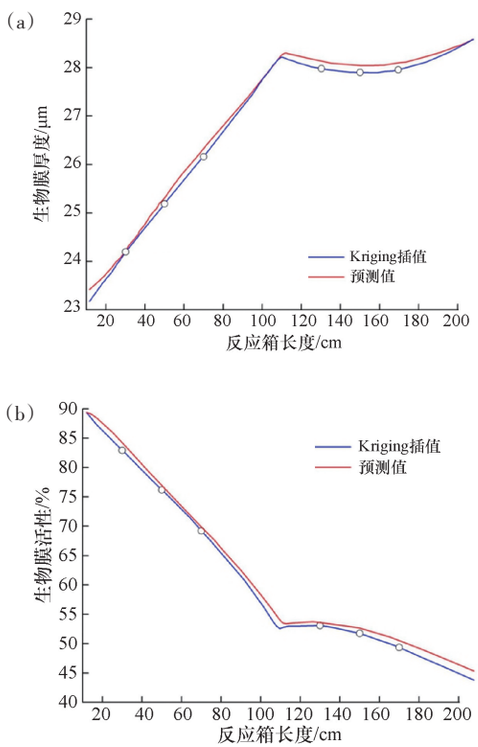

图4　生物膜厚度(a)和生物膜活性(b)实际值与预测值

Fig. 4　Actual and observed values of biofilm thickness and biofilm viability

利用Kriging插值法分析组合共聚焦激光扫描显微镜和流式细胞术显示,生物膜厚度从22 μm到31 μm,生物膜活性在反应器的流动方向上从80%降至30%。同时,证实了化学需氧量,总氮量和总磷酸盐去除特征与生物膜厚度和生物膜活性的水分分布之间存在明显的相关性[53]。

3.7　教育大数据的挖掘

教育是国家发展的根本,在大数据时代,教育大数据的挖掘是教育数据价值的体现。根据教育部的数据显示,截至2013年,中国高校贫困学生数目已经高达500余万,中国高校的贫困学生比例已经高达20%,其中,特困学生的比例已经超过了总在校人数的5%。全国各个高校都对贫困学生都有各种资助政策,尽量不让每个学生因为贫困而放弃学业。传统的资助形式都是大学生进行申请,并递交相关贫困证明材料,但部分学生因为较强的自尊心,不想让同学发现自己的特殊性而放弃申请,从而导致贫困助学金并不能准确地发放到每个贫困学生的手中。2015年3月2日,南京理工大学的"暖心饭卡工程"受到来自各界的关注。南京理工大学教育发展基金会工作人员对学生在日常生活中的数据进行了调查和数据的采集,该项调查涉及的共有16000余名南京理工大学当前在校学习的本科生,采集的数据为在2014年9月中旬至11月中旬期间学生的饭卡刷卡记录,将每个月平均在食堂消费60次以上,消费总额不足420元的学生确立为补助对象,不需要学生申报,直接将补助打入学生的饭卡。这次针对学生生活行为的数据挖掘,不仅在教育大数据的基础上实现了"精准扶贫",而且对学生真正做到了"人文关怀",体现出了数据的价值性。

3.8　国内图书情报的研究

目前,数据挖掘技术在图书情报领域的研究可分为6个方面:数字图书馆及个性化服务;WEB和信息服务;信息资源及参考咨询;图书馆及信息检索;高校图书馆及图书馆采购;情报学领域等。

大数据时代下,数据挖掘技术在中国图书情报领域中,基于中国知网数据库中图书情报领域的相关研究论文,郭婷等[54]分别利用了共词分析法和文献分析法对文献的增长规律和期刊的分布情况进行分析,在中国图书情报领域中,对数据挖掘的研究现状进行研讨,进一步强调了数据挖掘技术在图书情报领域研究的热点和重点。而且中国知网等在线图书机构采用数据挖掘技术研发的"学术不端文献检测系统"有效地避免了学术舞弊行为,保证了中国科研工作的正常发展。

4　大数据时代下数据挖掘的发展趋势

无论是研究领域,还是商业应用,数据挖掘都是热点问题,得到越来越多的人们关注,人们逐渐了解、学习并加以运用,相关领域日益成熟。在利用数据挖掘技术处理和解决实际问题时,王光宏等[55]提出了3个值得注意的角度:用数据挖掘技术解决问题的类型、解决

数据挖掘的数据准备工作及数据挖掘的理论基础。在大数据时代下,数据挖掘的发展趋势将会围绕数据价值的挖掘体现在以下5个层面。

1)多媒体数据挖掘。

大数据时代下,视频、音频、图像等都属于多媒体的范畴,随着时代的发展,海量的数据结构变得复杂化和动态化,而通过单独的传统数学方法去管理现实生活中的问题,得到的效果往往不能满足人们的期待。无人机和无人车的实际应用、公安天网工程的展开、智慧医疗项目的全面发展都会要求对多媒体数据进行快速处理,为了得到更理想的效果,得到的效果变得最优化,需要开发和设计数据挖掘的新智能算法。

2)金融领域潜在数据的挖掘。

在信用卡业务中,违约预测的数据挖掘具有预言性、有效性、实用性的优势。在信用卡交易的过程中,数据挖掘的应用类型也比较多,如在信用卡异常行为检测、高端信用客户的维护和信用卡风险控制等方面,均可以展开深入研究。

3)数据挖掘算法的改进和可视化。

当采用数据挖掘的算法分析和处理海量数据时,算法的改进主要取决于算法的精度和速度,即算法的准确度和效率。如今,学术研究主要集中在精度和效率之间设定适当的临界值和对数据挖掘的结果进行可视化两个方面。针对数据挖掘算法中的新贵——RNN、CNN、DNN、Capsule等一系列深度学习算法的研究,将成为引领大数据研究方法的风向标。

4)数据挖掘和隐私保护。

在解决实际问题时,难免会涉及隐私的数据,例如在研究信用卡和用户之间的关系时,数据中难免会有用户的个人信息;在研究宫颈癌(危险因素)与人的年龄、怀孕次数、性伴侣数等关系时,会有部分隐私信息不便透漏外界。在进行数据挖掘过程中,不泄露用户的个人隐私问题,对数据进行脱敏处理,将成为人们研究数据挖掘的另一个重要方面。

5)数据挖掘技术与其他系统的集成。

数据挖掘是一个完整的过程,而不是单纯的某一个算法或者其中的几个算法简单混合就可以的。将数据挖掘应用到实战演练的过程中,还是需要将数据挖掘与其他领域和系统有条理地集成,而不能理解成单独的一个算法就足以解决一个问题,进而最大化地体现了数据挖掘的优势。

5 结论

在大数据时代下,当运用传统的数学方法遇到困难时,熟练地应用数据挖掘技术显得格外重要。本文通过对国内外的研究现状进行剖析,分析了数据挖掘技术的主要方法,介绍了数据挖掘技术的应用领域,总结了在大数据时代下数据挖掘技术未来的发展趋势。

无论是在金融、医疗方面,还是在电信、教育等社会各个领域,每一时刻都会产生海量数据,由于社会存在过多的不确定性因素,导致处理的数据类型越来越繁杂,即便是采用计算机辅助,对于传统的处理方法、解决实际问题依然能力局限,但是通过数据挖掘技术,解决大数据问题,则开辟了另一个途径。未来的时代是“数据为王”,数据挖掘技术会面对更加严峻的挑战,利用数据挖掘的相关算法,处理实际问题和分析数据的能力将会更加显著。

参考文献(References)

[1] 吉根林, 赵斌. 面向大数据的时空数据挖掘综述[J]. 南京师大学报(自然科学版), 2014, 37(1): 1-7.
Jin Genlin, Zhao Bin. A Review of spatio- temporal data mining for big data[J]. Journal of Nanjing Normal University (Science & Technology Edition), 2014, 37(1): 1-7

[2] 王刚, 黄丽华, 张成洪, 等. 数据挖掘分类算法研究综述[J]. 科技导报, 2006, 24(12): 73-76.
Wang Gang, Huang Lihua, Zhang Chenghong, et al. Research summary of data mining classification algorithm[J]. Science Technology Review, 2006, 24(12): 73-76.

[3] 李海林, 梁叶, 王少春. 时间序列数据挖掘中的动态时间弯曲研究综述[J]. 控制与决策, 2017, doi: 10.13195/j.kzyjc.2017.1037.
Li Hailin, Liang Ye, Wang Shaochun. Review of dynamic time bending in time series data mining[J]. Control and Decision Making, 2017, doi: 10.13195/j.kzyjc.2017.1037.

[4] 龚著琳, 陈瑛, 苏懿, 等. 数据挖掘在生物医学数据分析中的应用[J]. 上海交通大学学报(医学版), 2010, 30(11): 1420-1423.
Gong Zhulin, Chen Ying, Su Yi, et al. Application of data mining in biomedical data analysis[J]. Journal of Shanghai Jiaotong University(Medical Science), 2010, 30(11): 1420-1423.

[5] 屈芳, 郭骅. “互联网+大数据”养老的实现路径[J]. 科技导报, 2017, 35(16): 84-90.
Qu Fang, Guo Hua. "Internet + big data" pension path to achieve[J]. Science & Technology Review, 2017, 35(16): 84-90.

[6] Pan T L, Sumalee A, Zhong R X, et al. Short-term traffic state prediction based on temporal-spatial correlation[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(3): 1242-1254.

[7] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2008, 31(2): 210-227.

[8] 沙亚清, 孙宏伟, 顾明. 基于智能卡和指纹识别的电子报税认证系统[J]. 计算机工程, 2006, 32(14): 133-135.
Sha Yaqing, Sun Hongwei, Gu Ming, et al. Electronic tax certification system based on smart card and fingerprint identification[J]. Computer Engineering, 2006, 32(14): 133-135.

[9] 周磊, 武建军, 张洁. 以遥感为基础的干旱监测方法研究进展[J]. 地理科学, 2015, 35(5): 630-636.
Zhou Lei, Wu Jianjun, Zhang Jie. Research progress of remote sensing based drought monitoring methods[J]. Geography Science, 2015, 35(5): 630-636.

[10] 谢玮, 刘斌, 刘鑫 ,等. 大数据时代的石油地震勘探系统与软件平台[J]. 科技导报, 2017, 35(15): 57-62.
Xie Wei, Liu Bin, Liu Xin, et al. Petroleum seismic exploration system and software platform in big data era[J]. Science & Technology Review, 2017, 35(29): 172-174.

[11] Bishop C M. Neural networks for pattern recognition[M]. New York: Oxford University Press, 1995.

[12] Kistler, Werner M. Spiking neuron models[M]. Cambridge: Cambridge University Press, 2002.

[13] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[14] Quinlan J R. C4.5: Programs for machine learning[M]. Cambridge: Morgan Kaufmann Publishers Inc., 1992.

[15] 万赟. 从图灵测试到深度学习: 人工智能60年[J]. 科技导报, 2016, 34(7): 26-33.
Wan Yun. From Turing test to in-depth learning: 60 years of artificial intelligence[J]. Science & Technology Review, 2016, 34(7): 26-33.

[16] Quinlan J R. Introduction of decision trees[J]. Machine Learning, 1986(1): 81-106.

[17] Guo H, Gelfand S B. Classification trees with neural network feature extraction[J]. IEEE Transactions on Neural Networks, 1992, 3(6): 923-33.

[18] 何禹德. 基于数据挖掘技术的糖尿病临床数据分析[D]. 长春: 长春工业大学, 2016.
He Yude. Clinical data analysis of diabetes based on data mining technology[D]. Changchun: Changchun University of Technology, 2016.

[19] Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules[C]//Proceedings 2001 IEEE International Conference on Data Mining. Piscataway NJ: IEEE, 2001, 28(6): 369-376.

[20] Han J, Yin X. CPAR: Classification based on predictive association rules[J]. Lecture Notes of the Institute for Computer Sciences Social Informatics & Telecommunications Engineering, 2003, 24: 236-255.

[21] 唐晓东. 基于关联规则映射的生物信息网络多维数据挖掘算法[J]. 计算机应用研究, 2015, 32(6): 1614-1616.
Tang Xiaodong. Multidimensional data mining algorithm for bioinformatics network based on association rule mapping[J]. Application Research of Computers, 2015, 32(6): 1614-1616.

[22] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining[C]//Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining California AAAI, 1998, 1711: 80-86.

[23] Bishop C M. Pattern recognition and machine learning (Information science and statistics) [M]. New York: Springer-Verlag New York, Inc., 2006.

[24] Friedman N, Dan G, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2/3): 131-163.

[25] Sahami M. Learning limited dependence Bayesian classifiers [C]//International Conference of Knowledge Discovery and Data Mining. California: AAAI,1996: 335-338.

[26] 朱凌云, 吴宝明. 医学数据挖掘的技术、方法及应用[J]. 生物医学工程学杂志, 2003(3): 559-562.
Zhu Lingyun, Wu Baoming. Techniques, methods and applications of medical data mining[J]. Biomedical Engineering Journal, 2003(3): 559-562.

[27] Holland J H. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence[M]. Cambridge: The MIT Press, 1975.

[28] Eberhart R, Kennedy J. A new optimizer using particle swarm theory[C]//Proceedings of the Sixth International Symposium on International Symposium on MICRO Machine and Human Science, 1995. Piscataway NJ: IEEE, 2002: 39-43.

[29] 邓乃扬, 田英杰. 支持向量机: 理论、算法与拓展[M]. 北京: 科学出版社, 2009.
Deng Naiyang, Tian Yingjie. Support vector machines: Theory, algorithms, and extensions[M]. Beijing: Science Press, 2009.

[30] Drucker H, Burges C J C, Kaufman L, et al. Support vector regression machines[J]. Advances in Neural Information Processing Systems, 1996, 28(7): 779-784.

[31] Hsieh C J, Chang K W, Lin C J, et al. A dual coordinate descent method for large-scale linear SVM[C]//Proceeding of International Conference on Machine Learning. New York: ACM, 2008: 408-415.

[32] 肖娟. 数据挖掘在物流业的应用综述[J]. 统计与决策, 2013

(11): 95-97.
Xiao Juan. Application of data mining in logistics industry[J]. Statistics and Decision, 2013(11): 95-97.

[33] Chen Y C, Chen C C, Peng W C, et al. Mining correlation patterns among appliances in smart home environment[J]. Lecture Notes in Computer Science, 2014, 8444: 222-233..

[34] Ollmann G. The evolution of commercial malware development kits and colour-by-numberscustom malware[J]. Computer Fraud & Security, 2008(9): 4-7.

[35] Ghiasi M, Sami A, Salehi Z. Dynamic VSA: A framework for malware detection based on register contents[J]. Engineering Applications of Artificial Intelligence, 2015, 44: 111-122.

[36] Bruschi D, Martignoni L, Monga M. Detecting self-mutating malware using control-flow graph matching[C]//International Conference on Detection of Intrusions and Malware & Vulnerability Assessment. Verlag: Springer-Verlag, 2006: 129-143.

[37] Kuzurin N, Shokurov A, Varnovsky N, et al. On the concept of software obfuscation in computer security[C]//International Conference on Information Security. Verlag: Springer-Verlag, 2007: 281-298.

[38] Christodorescu M, Jha S. Testing malware detectors[C]//ACM Sigsoft International Symposium on Software Testing and Analysis. New York: ACM, 2004: 34-44.

[39] Norouzi M, Souri A, Zamini M S. A data mining classification approach for behavioral malware detection[M]. Cairo: Hindawi Publishing Corp., 2016.

[40] 孙勤红. 基于梯度采样局部收敛的生物信息大数据挖掘[J]. 科技通报, 2015, 31(10): 214-216.
Sun Qinhong. Bioinformatics big data mining based on gradient sample local convergence[J]. Bulletin of Science and Technology, 2015, 31(10): 214-216.

[41] 朱佳俊, 郑建国, 李金兵. 基于粗糙分类的不确定可拓群决策数据挖掘及应用[J]. 控制与决策, 2012, 27(6): 850-854.
Zhu Jiajun, Zheng Jianguo, Li Jinbing. Uncertain extension computer aided decision data mining based on rough classification[J]. Control and Decision Making, 2012, 27(6): 850-854.

[42] Chen L H, Chiou T W. A fuzzy credit-rating approach for commercial loans: A Taiwan case[J]. Omega, 1999, 27(4): 407-419.

[43] 刘铭, 张双全, 何禹德. 基于改进型模糊神经网络的信用卡客户违约预测[J]. 模糊系统与数学, 2017(1): 143-148.
Liu Ming, Zhang Shuangquan, He Yude. Credit card customer default prediction based on improved fuzzy neural network [J]. Fuzzy Systems and Mathematics, 2017(1): 143-148.

[44] Fernandes K, Cardoso J S, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening [C]//Iberian Conference on Pattern Recognition and Image Analysis. Berlin: Springer, 2017: 243-250.

[45] Mangasarian O L, Street W N, Wolberg W H. Breast cancer diagnosis and prognosis via linear programming[J]. Operations Research, 1995, 43(4): 570-577.

[46] Liu M, Dong X G. The application of improved BP neural network in the diagnosis of breast tumors[C]//International Conference on Systems and Informatics. Piscataway NJ: IEEE, 2012: 1239-1242.

[47] Zheng C H, LI D W. The value of coronary arteriography in diagnosing coronary heart disease[J]. Shandong Medical Journal, 2005, 45(32): 42.

[48] Karimi M, Amirfattahi R, Sadri S, et al. Noninvasive detection and classification of coronary artery occlusions using wavelet analysis of heart sounds with neural networks[C]//London: Medical Applications of Signal Processing, the 3rd IEEE International Seminal. Piscataway NJ: IEEE, 2005: 117-120.

[49] Yang W, Fang P. New developments of resting ECG in detecting ventricular function in coronary artery disease[J]. Chinese Journal of Medicine, 2006, 41(1): 13-16.

[50] Liu M, Wang Y, Dong X G, et al. Improved BP algorithm and its application to intelligent diagnosis of coronary heart disease[C]//International Conference on Electronic Measurement & Instruments. Piscataway NJ: IEEE, 2011: 204-207.

[51] Liu M, He Y D, Wang J, et al. Hybrid intelligent algorithm and its application in geological hazard risk assessment[J]. Neurocomputing, 2015, 149(PB): 847-853.

[52] Lazarova V, Manem J. Biofilm characterization and activity analysis in water and wastewater treatment[J]. Water Research, 1995, 29(10): 2227-2245.

[53] Lin S, Wang X, Chao Y, et al. Predicting biofilm thickness and biofilm viability based on the concentration of carbon-nitrogen-phosphorus by support vector regression[J]. Environmental Science & Pollution Research, 2015, 23(1): 418-425.

[54] 郭婷, 郑颖. 数据挖掘在国内图书情报领域的应用现状分析——基于文献计量分析和共词分析[J]. 情报科学, 2015, 33(10): 91-98.
Guo Ting, Zheng Ying. Application of data mining in library and information service in china—Based on bibliometric analysis and co-word analysis[J]. Information Science, 2015, 33(10): 91-98.

[55] 王光宏, 蒋平. 数据挖掘综述[J]. 同济大学学报(自然科学版), 2004, 32(2): 246-252.
Wang guanghong, Jiang Ping. Data mining overview[J]. Journal of Tongji University(Science & Technology Edition), 2004, 32(2): 264-252.

Applications research of data mining technology in big data era

LIU Ming, LÜ Dan, AN Yongcan

School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China

Abstract In the era of big data, data mining technology has received more and more attention. This paper introduces the research background and status of data mining technology, followed by detailed description of its relevant algorithms such as decision tree, support vector machine, neural networks in detail. It then analyzes the data mining related applications and future development trend. Finally, it summarizes the challenges data mining technology will be faced with in the era of big data.

Keywords big data; data mining; decision tree; support vector machine; neural network

（责任编辑　刘志远）