# 浅析数据挖掘在化学化工中的研究进展论文

来源：网络 作者：浅唱梦痕 更新时间：2024-01-01

*随着数据库技术的不断发展，各行各业所积累的数据也越来越多，人们已不能满足数据表面所提供的信息，并且逐渐开始尝试挖掘出大量数据背后所隐藏的更加有用的信息，数据挖掘技术便由此应运而生。随后，数据挖掘技术便在信息产业和商业服务等领域产生了重大影响...*

随着数据库技术的不断发展，各行各业所积累的数据也越来越多，人们已不能满足数据表面所提供的信息，并且逐渐开始尝试挖掘出大量数据背后所隐藏的更加有用的信息，数据挖掘技术便由此应运而生。随后，数据挖掘技术便在信息产业和商业服务等领域产生了重大影响，人们利用数据挖掘技术从所积累的数据库中得到了过去所不可能获取的知识。随着数据挖掘技术取得成效之后，许多不同领域的研究人员也显示出了极大的兴趣。

自计算机诞生以来，信息技术便不断地推动着化工行业的发展。信息化时代的到来，促进了信息技术和化学工业的进一步结合，为数据挖掘技术在化工领域的应用打下了坚实的基础。随着大数据时代的来临，数据挖掘技术势必会对化学与化工领域产生更加深远、更加全面的影响。近年来，很多研究者针对以往积累的数据，将数据挖掘技术运用到自己的研究课题中，取得了很多极具应用价值的研究成果。本文通过整合近年来数据挖掘技术在化学化工领域所取得的突破性进展，评述了数据挖掘技术在化工建模、工艺条件的优化、数据库平台的创建和安全生产等方面的应用。

>1 数据挖掘技术

数据挖掘技术是智能系统理论的重要内容，同样也是数据库处理的重要的技术手段。它是利用关联规则、分类、聚类等分析方法从大量无序的、存在噪声的、冗余的或缺损的数据中提取出有用的潜在信息，并为进一步的研究提供数据支撑的数据库处理技术。数据挖掘技术的操作流程一般为: 信息收集、数据集成、数据预处理、数据挖掘过程、模式评估和知识表示几个阶段。其中数据挖掘过程为该流程的技术核心，运用的主要方法有: 统计方法、决策树、规则推理、模糊集、遗传算法、神经网络、主元分析、偏最小二乘法等。

>2 数据挖掘在化学化工中的应用

2. 1 化工建模的应用

化工过程建模一直是化工领域的难点。因为在实际过程中，化工过程系统的很多条件因素具有交互性和严重的非线性，因此化工过程建模具有很大的困难。然而数据挖掘技术具有丰富的算法，神经网络算法能够以任意精度逼近任意非线性映射，对处理含噪声和非线性数据具有较大的优势; 遗传算法在解决非线性问题时不依赖于问题模型的特性，不仅具有鲁棒性和全局最优性，还具有高效率和可并行性等特点; 主元分析在解决非线性数据时能够做到数据简化、数据压缩、建模等; 偏最小二乘法对多因变量对多自变量的回归建模具有较为明显的优势。

孔德根等对18 种酚类化合物的量子化学参数进行了优化计算，采用多元线性回归分析和人工神经网络反传误差算法，研究了酚类化合物对抑制黑曲霉毒性的定量构效关系，并采用去一法通过对模型的预测能力和稳健性的验证筛选出了最佳的计算模型。分析表明酚类化合物的分子体积和苯环上的净电荷增量与其对黑曲霉的抑制毒性呈正相关，而最低空轨道能与黑曲霉的抑制毒性呈负相关。该文章的建模方法为评估和预测其他化合物对抑制黑曲霉的毒性提供了参考。

林凌等提出了一种针对体表组织内进行的高谱图技术。该技术通过对组织内的光谱信息和图像信息等多元信息进行关联，然后采用数据挖掘技术，对数据和病症进行分析建模。通过该模型挖掘出组织的图像、光谱以及二者交互项与疾病之间的关系，大大提高了临床皮肤病的诊断精度和准确度。还能够有效地改善光谱技术和光学相干层析技术对组织成分和位置信息探测不足和缺乏系统性、多元性等局限性，为皮肤病的早期诊断提供了技术支撑，并且对人体内更深层组织以及器官的检测提供了思路。

Kusiak 等通过收集202\_ ～ 202\_ 年间甲烷生产的工业数据，对甲烷生产的参数进行了探究，采用自适应神经模糊推理系统( ANFIS) 对甲烷的生产预测进行建模。并通过百分比误差、偏差因子、均方根误差、归一化均方根误差和一致性指数等五个指标对该模型进行了评估。结果表明，利用自适应神经模糊推理系统所建立的模型性能优异，所预测的数据与实测值显示出了高度的一致性。与采用神经网络、支持向量机、随机森林等算法所建立的甲烷生产预测模型比较也显示出了更好的预测精度。该模型不仅能够对甲烷的生产过程进行更准确的模拟和预测，而且能够指导并优化甲烷的生产过程。

刘天玲等采用主成分分析、局部线性嵌入、支持向量机、偏最小二乘和小波变换五种模式识别方法，利用NIR 技术建立了玉米种子的分类识别模型。该模型已成功运用于108 玉米杂交种和母本178 种子的识别，而且具有较高的识别精度和分类精度。Rommel 等提出了一种基于机器学习算法和电感耦合等离子体质谱法测定有机甘蔗样品真实性的识别模型。该模型可以采用Random Forest 和Naive Bayes 算法对样品进行分析识别，RandomFores 算法通过检测32 个化学元素作为参考，准确率约为90%; Naive Bayes 算法只需检测8 个化学元素作为参考，但是准确率却提升到了95. 4%。此模型能够方便且高效的区分出样品中的有机甘蔗。Sabanci 等则使用UCI 机器学习数据库分别运用KNN、Naive Bayes、J48 和多层感知器的数据集等数据挖掘算法建立了种子物种分类的数学模型。并成功运用于小麦种子的分类，当神经元数目为7 时，分类成功的准确率约为97. 174 9%，MAE 误差和均方根误差分别为0. 029 8 和0. 118 1，显示出较好的分类识别效果。

2. 2 优化工艺条件的应用

数据挖掘能从海量的交互、高维、无序的自变量数据库中发现潜在的且能够导向因变量的有用信息。并且能够对特定的因变量借助分类决策树算法、聚类算法和分类树算法等对多维无序的自变量进行择优。利用数据挖掘的这一特点，可以对化工工艺条件进行优化。将多维、非线性的工艺条件作为自变量，特定的工艺结果作为因变量采用不同的算法进行优化，直至找到最佳的工艺条件。姜求宇等运用数据挖掘技术实现了对固相合成纳米氧化镍工艺参数的优化。通过均匀设计方法分析了多因素对纳米氧化镍透光率的影响，并对实验结果进行了逐步回归分析和最优化计算，最终得到固相合成纳米氧化镍的最优工艺参数，在该条件下得到的氧化镍的收率为95. 5%，平均粒径在7 nm。该方法对科研工作中的小型的数据库的工艺条件的优化处理具有较强的实用性。

周永生等针对生化企业的复杂性和所产生的数据的无序性、交互性等特点提出了带有确定性决策项的关联规则挖掘算法。相较于传统的关联规则挖掘算法，该算法从环境因子和环境因子数据项的关系入手，减少了无关频繁项目集的产生，缩短了算法的运行时间，提高了算法的运行效率。该算法对生化企业生产环境的优化具有很好的适应性。同年，该算法已成功的运用于发酵法生产柠檬酸的实验中，并能够较快的得到高转化率和高总酸量的优化生产环境。该算法的提出有利于提高生化企业优化工艺参数的效率和准确性。

He 等通过对基因组进行数据挖掘，找出了在甘露醇作为底物的条件下，NADH 还原酶( PgCR)对4-氯乙酰乙酸乙酯( COBE) 还原成( S) -4-氯-3-羟基丁酸乙酯( ( S) -CHBE) 具有较高的催化活性和优良的立体选择性。并对其反应条件进行了优化，在反应温度为30 ℃，反应pH 为7，底物甘露醇为2. 5 mmol /mmol COBE，NAD + 为0. 1 μmol /( mmol COBE) ，细胞用量为1 g /mL 时，CHBE 的产量可达到99%，高对映体过量值 99%。郭羽婷等以吉林大学“无机制备与合成化学国家重点实验室”所建立的开放骨架磷酸铝合成反应数据库为研究对象，采用MWMR 算法，探究了溶剂、模板剂等合成参数对于开放骨架磷酸铝合成的影响。该研究成果对含有( 8，6) 元环结构开放骨架磷酸铝定向合成参数的优化具有一定的指导意义。

2. 3 数据库平台的应用

数据库是指支持管理决策过程的、集成的、交互的一个数据集合。它是对大量无序、非线性的数据通过聚类、关联、模糊集、最大权重和最小冗余等特征算法等整合的信息系统。通过一个直观的、简洁的人机交互界面，人们可以方便的查找与分析各种数据之间的关联，并且在查询的条件中引入模糊偏好，使得搜索的结果更具有灵活性。因而通过数据挖掘技术对收集的数据进行关联、聚类，建立化工数据库或开发化工应用软件，也是数据挖掘技术在化学化工领域的一个应用热点。

杨善升等运用数据挖掘技术，开发了一种适用于合成氨过程优化和监测的应用软件。该软件集成了模式识别、人工智能、统计学习理论等多种优化方法，具有操作简单、功能强大和适应性强等特点。而且该软件已成功运用于云南云维集团有限公司合成氨过程的生产优化中，并取得了良好的效果。

Peng 等基于J2EE 通过对中药中有效成分的信息进行归纳、总结、关联、聚类等研究建立了中药有效成分化合物的数据库。通过该数据库可以方便研究者对药物的化学性质和药理性质有一个系统的认识，并且通过各类化合物的化学性质与药理性质的关联与聚类，有助于研究者进行新药的开发。Ghadbeigi 等对200 个出版刊物中锂离子电池材料进行了数据挖掘，整合并创建了包括各类电池材料的超过16 000 个数据的锂离子电池的大型数据库。该数据库不仅包含了电极材料的关键性能参数，如能量密度、功率密度、放电容量等，还对相关电极材料的资源分布和储量等信息也进行了分析。利用该数据库不仅可以分析出阴极材料和阳极材料所具有的特性，而且根据对不同材料性能的挖掘，有助于研究人员探索新型的锂离子电极材料。肖建茂等利用数据挖掘技术探究了各种锂离子电池材料的物理化学性质与材料组成和组织结构等的关系，并根据材料基因工程的基本思想通过高通量第一性原理构建了锂电池材料设计的技术平台。该平台通过数据挖掘可以总结出性能优良的锂离子电池材料的结构特性，有助于新型锂离子电池材料的探索和现有材料的改性研究。

2. 4 安全生产的应用

化工厂是安全事故频发的一类企业，控制安全事故的传统办法小心防范和密切监控只能被动的对安全事故进行排查，对安全事故的控制效果不佳。随着数据挖掘技术的不断发展，使得化工厂在安全事故的管控方面占据主动成为了可能。数据挖掘技术提供了模糊集理论、主观Bayes 算法等不确定性推理方法，来对事故原因和事故征兆之间具有的模糊性和随机性等不确定关系进行研究。根据化工厂存储的大量的运行数据，通过数据挖掘技术的特定算法，挖掘出各个事故征兆对结论的支持度，得到相应地征兆权值，进而计算出相应的阀限值。通过采集到的事故征兆的多少，根据合理的规则可信度公式，改变规则可信度，从而对生产进行监控并对潜在的危险进行预测。

曹顺安针对火电厂的水汽化学过程的故障的诊断方法的缺乏，利用FP\_growth 算法设计了一种基于可信度理论的诊断模型。该模型不仅适用性较广，而且对故障具有较强的识别能力，能够在故障初期征兆尚未完全出现、某些次要征兆值未能获取等情况下，对可能会发生的故障进行识别，并作出警告。这对于火电厂的安全生产具有十分重要的意义。

Cheng 等通过对202\_ ～ 202\_ 年之间在台湾发生的349 例石化行业的重大职业事故进行了数据挖掘技术的分类和回归树分析，并对事故的分布和发生的原因以及事故的频发点进行了统计学分析。认为对于石化行业，天气的变化、材料结构的缺陷、内部材料温度和压力的变化、人们的疏忽等都会导致安全事故的发生。故因对事故的频发点定期进行风险评估，制造商也应该提高机械和设备入口点的安全和保护装置，高危作业的施工人员也应当定期进行安全教育培训。

李鹏等运用数据挖掘技术对中国石油化工股份有限公司的催化裂化装置所积累的海量数据进行深入探索研究与分析，解决催化裂化装置报警问题、结焦问题和收率问题，进一步提高了催化裂化装置运行能力，并取得了良好的经济收益和社会效益。石荣雪等通过对文献数据库进行大数据分析，挖掘出了国内因锅炉化学水工况选取不当而造成事故情况，并进行了数据分析，确定了一些常见锅炉机组的最适化学水工况，并给出了相应的建议。对国内锅炉安全生产具有一定的参考价值。

>3 结束语

自20 世纪90 年代以来，随着科研工作者的不断努力，数据挖掘技术已在化学化工领域得到了广泛的应用。它所包含的丰富的算法不仅对化学化工所产生的大量、无序、多维的数据具有很好的适应性，而且也为解决化工建模、工艺参数的优化、化学数据库平台和安全生产等多方面的难题提供了新的思路。随着化工过程自动化、智能化的进一步发展，计算机控制系统将会采集并存储越来越多的数据，届时数据挖掘技术在化学化工领域的新应用将不断涌现。已经形成雏形的数据挖掘集成技术已经显示出了对比单一技术的优越性，届时也将得到更加广泛的应用。

本DOCX文档由 www.zciku.com/中词库网 生成，海量范文文档任你选，，为你的工作锦上添花,祝你一臂之力！