# 融合计算机人工智能技术下的数据挖掘系统分析

李志鹏
三峡旅游职业技术学院　湖北宜昌　443001

**摘　要：**数据挖掘是从物联网海量的数据中挖掘出潜在的、不为人知的具有价值的信息知识。这些信息具有一定的潜在价值，并且是可运用的、可理解的，能够为企业带来经济利益或者为科学技术和社会的发展提供参考依据。本研究深入分析了融合人工智能技术的数据挖掘系统的架构、关键技术、数据分析流程及其面临的挑战与机遇。通过探讨人工智能算法在数据挖掘中的应用，如神经网络、遗传算法等，揭示其如何提升数据挖掘的效率与准确性。对数据挖掘中的数据分析环节进行详细阐述，包括数据预处理、特征选择、模型构建与评估等，旨在为相关领域的研究与实践提供理论依据和技术参考。

**关键词：**计算机人工智能技术；数据挖掘系统；数据分析；算法

## 引言

物联网技术的飞速发展，推动着大数据挖掘的变革，信息挖掘逐渐从整体化挖掘对象转变为分布式、碎片化的挖掘模式。为实现物联网中海量信息的有效利用，以及促进数据挖掘的集中化发展，提出了具有前瞻性的人工智能大数据挖掘算法，在智能化时代背景下对于我国工业经济的发展和建设具有重要作用。计算机人工智能技术的出现为数据挖掘系统带来了新的活力和创新能力。人工智能技术中的机器学习、深度学习等算法能够自动学习数据中的模式和规律，大大提高了数据挖掘的效率和准确性。因此，深入研究融合计算机人工智能技术下的数据挖掘系统具有重要的理论和实践意义。

## 一、融合人工智能技术的数据挖掘系统架构

### （一）数据采集与存储层

数据采集与存储层作为数据挖掘系统的基础环节，承担着从多元异构数据源中提取和整合信息的关键任务。这一过程涉及对大规模、多维度数据的高效获取与管理，包括但不限于关系型数据库中的结构化数据、基于XML或JSON格式的半结构化数据，以及以文本、图像、音频等形式存在的非结构化数据。为确保数据的完整性与一致性，需采用高并发数据抓取技术及实时流处理框架，例如Apache Kafka或Spark Streaming，以支持动态数据流的持续捕获与传输。

在存储层面，该阶段通过分层架构设计满足不同数据类型的存储需求。对于结构化数据，可利用高度优化的关系型数据库管理系统（RDBMS），如MySQL和Oracle，实现快速查询与事务处理；而对于海量非结构化数据，则借助分布式NoSQL数据库（如MongoDB、Cassandra）或Hadoop分布式文件系统（HDFS），提供灵活扩展能力和高效的随机访问性能。同时，元数据管理机制被引入以描述数据特征及上下文信息，从而增强数据语义表达能力并简化后续分析流程。通过上述技术手段的综合运用，数据采集与存储层为整个数据挖掘系统的运行奠定了坚实基础。

### （二）数据预处理层

数据预处理作为数据挖掘系统的关键组成部分，旨在解决原始数据中普遍存在的噪声、缺失值及重复值等问题。通过构建严谨的数据清洗流程，可有效识别并剔除异常值与无关数据，同时运用统计学方法对缺失值进行科学估算，确保数据完整性。在多源异构数据环境中，数据集成技术通过实体识别与属性匹配，实现跨数据源的无缝整合，从而形成统一的数据视图。为提升算法适用性，需采用标准化与归一化等转换策略，调整数据尺度以消除量纲差异，并借助离散化手段将连续型变量转化为适合特定模型输入的形式。降维处理可通过主成分分析等方法提取关键特征，进一步优化数据结构，为后续深度分析奠定坚实基础。整个过程严格遵循数据质量控制原则，确保信息的一致性与可靠性，以满足复杂数

**作者简介：**李志鹏，男（1996.7—），汉族，湖北省宜昌市人，硕士，研究方向：图像识别，人工智能。

据分析任务的需求。

**（三）人工智能算法层**

作为数据挖掘系统的核心层，该层深度融合了多种先进的人工智能算法以实现高效的数据分析与模式发现。深度学习技术中的卷积神经网络（CNN）通过多层次的特征映射和非线性变换，能够自动提取图像数据的空间层次结构特征，其在目标检测、图像分类等任务中展现出卓越性能。循环神经网络（RNN）及其改进变体长短期记忆网络（LSTM）凭借对时序依赖关系的有效建模能力，在语音信号处理、自然语言生成以及时间序列预测等领域实现了突破性进展。遗传算法作为一种全局优化方法，基于生物进化理论，利用种群迭代机制探索复杂解空间，尤其适合解决高维、非线性及多模态优化问题。其交叉、变异操作可有效增强搜索多样性，而选择算子则加速收敛至最优解。此外，强化学习通过智能体与环境交互构建奖励机制，动态调整策略函数以最大化长期收益，在推荐系统、游戏AI等场景中表现出强大的决策能力。集成学习方法如随机森林和梯度提升决策树（GBDT）通过组合多个弱学习器形成强学习器，显著提升了模型的泛化能力和稳定性，为大规模数据挖掘任务提供了可靠解决方案。这些算法共同构成了数据挖掘系统的智能化基础，推动了从海量数据中高效提取隐含知识的技术革新。

**（四）数据分析与知识发现层**

在这一层，利用人工智能算法对预处理后的数据进行分析，挖掘出有价值的信息和知识。数据分析过程包括特征选择、模型构建和模型评估等环节。特征选择是从原始数据的众多特征中选择出对目标变量最有影响的特征子集，以降低数据的维度，提高模型的效率和准确性。例如，可以采用信息增益、基尼系数等方法进行特征选择。模型构建则根据数据的特点和挖掘任务的需求选择合适的模型，如分类模型（决策树、支持向量机等）、回归模型（线性回归、岭回归等）等。模型评估则通过各种评估指标（如准确率、召回率、均方误差等）来衡量模型的性能，根据评估结果对模型进行调整和优化。

**（五）结果展示与应用层**

挖掘出的结果需借助直观展示手段呈现，以提升用户对数据内涵的理解与应用效能。专业领域中常用可视化工具如热力图、散点矩阵及三维表面图等，结合交互式仪表板技术，强化信息传达精度。报表生成则依托统计汇总与多维分析方法，确保数据解释的系统性与逻辑性。在商业智能场景下，通过构建客户细分模型，精准定位目标群体特征，优化资源配置效率。医疗健康领域内，基于数据挖掘算法解析复杂病历关联，生成个性化诊疗建议，辅助临床决策制定。金融行业中，风险评估模型整合多元变量，实现动态监控与预警机制建立，为机构提供稳健运营保障。上述实践均体现结果展示与应用层在知识转化中的核心价值。

## 二、人工智能技术在数据挖掘中的数据分析应用

**（一）数据预处理中的数据分析**

在数据预处理阶段，数据分析起到了关键的作用。对于数据清洗，需要分析数据中的噪声和异常值的分布情况，以便选择合适的清洗方法。例如，通过绘制数据的箱线图可以直观地观察到数据中的异常值分布，然后根据异常值的类型（如孤立点、聚类异常值等）选择合适的处理方法。对于数据集成，需要分析不同数据源的数据结构和语义关系，以确保数据能够正确地合并和整合。例如，在将关系型数据库和非关系型数据库中的数据进行集成时，需要分析两者之间的字段对应关系、数据类型转换等问题。对于数据转换，需要分析数据的分布特征，以确定合适的转换方法。例如，对于服从正态分布的数据，可以采用Z-score标准化方法；对于数据分布不均匀的数据，可以采用对数转换等方法。

**（二）特征选择中的数据分析**

特征选择作为数据分析的核心环节，需深入探讨特征与目标变量间的统计关联。通过量化方法评估线性依赖程度，如利用皮尔逊相关系数捕捉连续型变量间的协变趋势，揭示潜在的主导因子。针对非线性关系，互信息等熵基度量可有效衡量特征与目标变量间的信息共享程度，从而揭示复杂依赖结构。特征冗余分析是优化特征集不可或缺的部分，高度相关的特征往往导致模型过拟合或计算效率降低。在实际应用中，相关性矩阵可视化可用于直观识别多重共线性问题，结合方差膨胀因子进一步量化冗余程度。基于主成分分析的降维技术能够从高维空间中提取正交特征，既减少维度灾难风险，又保留关键信息。特征选择过程还需兼顾领域知识约束，以确保所选特征具备解释性和泛化能力，从而提升模型性能与可靠性。

**（三）模型构建与评估中的数据分析**

在模型构建阶段，需要根据数据的特点和挖掘任务的需求选择合适的模型。这需要对数据的分布、规模、特征数量等进行分析。例如，对于数据规模较小、特征

数量较少的数据集，可以选择简单的线性模型，如线性回归模型；对于数据规模较大、特征数量较多且存在非线性关系的数据，可以选择复杂的非线性模型，如神经网络模型。在模型评估阶段，需要根据不同的挖掘任务选择合适的评估指标。例如，在分类任务中，常用的评估指标有准确率、召回率、F1-score等；在回归任务中，常用的评估指标有均方误差、平均绝对误差等。通过对模型评估指标的分析，可以了解模型的性能优劣，从而对模型进行调整和优化。

## 三、融合人工智能技术的数据挖掘系统面临的挑战与机遇

### （一）挑战

1.数据质量问题

数据质量问题在数据挖掘过程中始终是一项严峻挑战，即使经过预处理阶段的优化，其复杂性依然不容忽视。非结构化数据中的语义模糊性可能导致特征提取不准确，从而削弱模型的泛化能力。图像数据中存在的噪声、遮挡及光照变化等现象会进一步降低特征表达的有效性，影响分类或预测性能。多源异构数据融合时的时间戳差异、采样频率不一致以及标注偏差等问题，也可能引入额外的不确定性，进而对数据分析的准确性与可靠性造成干扰。在实际应用中，这些问题需要结合领域知识与统计方法进行精细化处理，以提升数据质量并增强挖掘结果的可信度。

2.算法复杂度与可解释性

先进的人工智能算法，尤其是深度学习方法，在数据挖掘领域展现了卓越的性能，但其高计算复杂度导致模型训练过程需耗费大量算力资源，并对硬件基础设施提出严苛要求。在实际应用场景中，此类算法通常表现为黑箱特性，内部参数与结构的高度非线性关联使得决策机制难以通过直观方式解析。特别是在医疗诊断和金融风控等领域，相关法规与行业标准对模型透明性及可追溯性有严格规范，而深度学习模型因缺乏清晰的因果推断路径，可能无法完全满足这些需求，从而限制了其在高敏感场景中的广泛应用与部署。

3.数据隐私与安全

在数据挖掘的全生命周期中，海量敏感信息的流转与处理对隐私保护及安全性提出了严峻挑战。医疗领域的数据挖掘涉及多源异构的个人健康信息，包括但不限于电子病历、基因组数据和生理监测记录，其高度敏感性要求建立基于差分隐私或联邦学习的保护机制以确保合规性。在商业智能分析中，企业核心数据资产面临网络攻击、非法访问及未授权共享的风险，需依托加密算法、访问控制策略以及区块链技术构建多层次防护体系。数据治理框架下的审计追踪与风险评估亦成为保障数据完整性和可用性的关键环节，从而实现隐私保护与价值挖掘之间的动态平衡。

### （二）机遇

1.大数据与云计算

海量多源异构大数据的持续积累为数据挖掘技术注入了新的活力，其丰富的语义维度与庞大的样本规模显著提升了模式发现的深度与广度。在实际应用中，云计算平台凭借分布式存储与并行计算能力，有效解决了大规模数据处理中的性能瓶颈问题，为复杂人工智能算法的高效训练与部署提供了坚实基础。特别是在深度学习框架下，云原生环境下的弹性资源调度与优化加速了模型迭代周期，进一步推动了数据驱动的知识发现进程。边缘计算与联邦学习等新兴范式的引入，使得数据挖掘能够在保障隐私的前提下实现跨域协同分析，从而拓展了其在高敏感性行业的应用场景。

2.新兴技术的融合

物联网技术通过部署海量感知设备，实现对物理世界多维度数据的实时采集与传输，为数据挖掘提供了高密度、多源异构的数据支撑。区块链技术凭借其分布式账本与加密算法特性，确保数据在存储、共享过程中的不可篡改性与可追溯性，从而构建起数据挖掘所需的信任机制。两者与数据挖掘系统的深度融合，不仅优化了数据质量与完整性，还推动了智能化决策体系的构建与发展。

3.跨领域应用

数据挖掘借助深度学习与知识图谱技术，实现医疗领域中基于多组学数据的疾病预测模型构建及新药靶点发现，提升诊疗精确性；在金融场景下，通过高维特征提取与异常检测算法，强化风险量化评估与实时欺诈识别能力；于智慧交通体系中，运用时空数据分析优化动态路径规划，提高运力调配效率；同时，在教育测评中引入自适应学习框架，依托学生行为数据画像实现个性化教学资源推荐，推动跨学科交叉融合与智能化转型升级。

## 结论

融合计算机人工智能技术的数据挖掘系统为从海量数据中挖掘有价值的信息和知识提供了强大的工具。通

过对数据挖掘系统架构、人工智能技术在数据分析中的应用以及面临的挑战与机遇的分析，我们可以看到，虽然目前还存在一些挑战，但随着技术的不断发展和创新，数据挖掘系统将在更多领域发挥更大的作用。在未来的研究中，需要进一步提高数据质量，降低算法复杂度，提高算法的可解释性，加强数据隐私和安全保护，以推动数据挖掘系统的不断发展和完善。

**参考文献**

[1]张文战.人工智能技术在舰船智能导航系统大数据挖掘中的应用[J].舰船科学技术，2020，42（24）：118-120.

[2]刘国城，张千荷.人工智能技术在国家审计中的应用研究[J].中国审计评论，2024，（02）：108-121.

[3]马成虎.人工智能技术在计算机网络中的应用[J].信息记录材料，2024，25（09）：52-54.

[4]辜凌云.人工智能机器学习内容的著作权合法性基础——以文本和数据挖掘例外规则为视角[J].出版广角，2024，（15）：54-59.

[5]封晶，孙澄宇，董平，等.基于人工智能技术的物联网大数据挖掘算法[J].石河子科技，2022，（05）：14-16.