第 8 卷◆第 12 期◆版本 1.0◆2025 年
文章类型：论文 | 刊号（ISSN）：2630-4686 /（中图刊号）：380GL020

基于机器学习的金融数据挖掘技术探讨

曲思龙
佳木斯大学
DOI:10.12238/er.v8i12.6662

[摘　要] 本文系统探讨了机器学习在金融数据挖掘中的理论基础、核心方法和实际应用。首先分析了金融数据的多维性、非线性和高噪声等特征，阐述了机器学习相比传统统计方法的优势。其次，深入研究了数据预处理、特征工程、核心算法应用以及时间序列预测等关键技术，重点介绍了 SVM、随机森林、深度神经网络和 LSTM 等算法的应用。最后，通过信用风险评估、算法交易、反欺诈检测和智能投顾等案例分析，展示实践效果并进行前瞻性分析。研究表明，机器学习能有效提升金融数据挖掘的准确性和效率，但在模型解释性、实时处理和监管合规等方面仍面临挑战。
[关键词] 机器学习；金融数据挖掘；深度学习；时间序列预测；风险管理
中图分类号：F830 文献标识码：A

Discussion on Financial Data Mining Technology based on Machine Learning

Silong Qu
Jiamusi University

Abstract: This paper systematically discusses the theoretical basis, core methods and practical applications of machine learning in financial data mining. Firstly, the characteristics of multi-dimensional, nonlinear and high noise of financial data are analyzed, and the advantages of machine learning compared with traditional statistical methods are expounded. Secondly, the key technologies such as data preprocessing, feature engineering, core algorithm application and time series prediction are deeply studied, and the application of SVM, random forest, deep neural network and LSTM algorithm is introduced. Finally, through case studies such as credit risk assessment, algorithmic trading, anti-fraud detection and smart investment advisory, the practical effects are demonstrated and forward-looking analysis is carried out. Research shows that machine learning can effectively improve the accuracy and efficiency of financial data mining, but it still faces challenges in model interpretation, real-time processing and regulatory compliance.
Keywords: machine learning; financial data mining; deep learning; time series prediction; risk management

引言

金融行业每天产生海量交易、市场、客户和风险数据，蕴含着丰富的市场规律和行为模式信息。传统金融分析方法主要依赖统计学理论和经验模型，在处理高维、非线性、动态变化的金融数据时存在局限性。随着人工智能技术发展，机器学习作为强大的数据分析工具，能自动发现数据中的隐藏模式，构建复杂的非线性映射关系，在金融预测、风险控制、投资决策等方面表现优异。各种机器学习算法从支持向量机到深度神经网络，从回归分析到强化学习，都在金融数据挖掘中发挥重要作用。然而，金融数据的高噪声、非平稳性、时变性等特殊性质，以及金融行业对模型可解释性和稳健性的严格要求，为机器学习应用带来独特挑战。因此，深入研究机器学习在金融数据挖掘中的应用理论、技术方法和实践经验，对推动金融科技创新、提升服务效率、防范金融风险具有重要意义。

1 金融数据特征与机器学习基础

1.1 金融数据的基本特征与挑战

金融数据具有显著的复杂性，为数据挖掘带来独特挑战。首先，金融数据呈现多维性特征，包括价格、成交量、财务指标、宏观经济数据等，维度间存在复杂相关关系。其次，金融时间序列普遍表现出非线性和非平稳性，市场价格波动受基本面信息、市场情绪、政策变化等多因素影响，数据分布随时间动态变化，传统线性模型难以准确捕捉。此外，金融数据具有高噪声特征，随机波动和异常事件掩盖真实市场信号，增加模式识别难度。同时，异常值和缺失值问题频繁出现，可能源于市场突发事件、交易错误或数据传输问题，

处理不当会严重影响模型训练效果和预测准确性。

1.2 机器学习在金融领域的适用性分析

机器学习技术在金融领域具有天然适用性优势。监督学习方法适用于金融预测和风险评估，能利用历史数据学习特征与目标变量的映射关系，在股价预测、信用评级、欺诈检测等场景发挥重要作用。无监督学习在客户细分、市场结构分析、异常检测等方面具有独特价值，能自动发现隐藏模式和结构。强化学习在算法交易和投资组合优化等序贯决策场景展现巨大潜力。相比传统统计方法，机器学习具有更强的非线性建模能力，能自动发现复杂特征组合和交互关系，具有良好的数据适应性，能处理高维异构数据并通过大数据训练提升性能[1]。然而，金融领域对模型可解释性要求较高，需要在预测精度和可解释性间寻求平衡，选择适合特定场景的算法架构。

2 金融数据挖掘的关键技术与方法

2.1 数据预处理与特征工程技术

数据预处理和特征工程是金融数据挖掘成功的关键环节，直接影响模型的性能和稳定性。金融数据清洗工作包括异常值检测与处理、缺失值填补、数据一致性检查等步骤，其中异常值的处理需要特别谨慎，因为某些看似异常的数据点可能包含重要的市场信息，需要结合业务知识进行判断。数据标准化处理对于金融数据尤为重要，由于不同金融指标的量纲和数值范围差异巨大，需要采用 Z-score 标准化、最大最小值标界化或分位数标准化等方法，确保各特征在模型训练中的贡献度相对均衡。特征工程是提升模型性能的核心技术，包括技术指标构建、统计特征提取、交互特征生成等，常用的技术指标如移动平均线、相对强弱指数(RSI)、布林带等能够有效捕捉价格趋势和波动特征。特征选择策略需要综合考虑特征的相关性、重要性和稳定性，采用过滤式、包装式或嵌入式方法筛选出最具预测价值的特征子集。对于高维金融数据，降维技术如主成分分析(PCA)、线性判别分析(LDA)、t-SNE 等能够有效减少数据维度，降低计算复杂度并缓解维数灾难问题。时间窗口设计和数据分割技术需要考虑金融数据的时序特性，采用滑动窗口、固定窗口或自适应窗口策略，确保训练集和测试集的时间顺序合理性，避免未来信息泄露问题。

2.2 核心机器学习算法应用

支持向量机(SVM)在金融数据挖掘中表现出色，特别适用于小样本、高维数据的分类和回归任务，在股价趋势预测、信用风险评估等场景中广泛应用。SVM 通过核函数技巧能够处理非线性关系，常用的核函数包括径向基函数(RBF)、多项式核等，能够有效捕捉金融数据中的复杂模式。随机森林和梯度提升等集成学习方法在金融风险评估中发挥重要作用，这类方法通过组合多个弱学习器形成强学习器，具有良好的泛化能力和鲁棒性，能够处理混合类型的特征变量，并提供特征重要性评分，有助于理解模型的决策逻辑。神经网络和深度学习技术在量化交易中展现出强大的潜力，多层感知机(MLP)能够学习复杂的非线性映射关系，而卷积神经网络(CNN)在处理图像化的金融数据(如 K 线图、热力图)方面具有天然优势。深度神经网络的强大表达能力使其能够自动学习层次化的特征表示，无需人工进行复杂的特征工程。聚类算法在客户分群和市场细分中发挥重要作用，K-means、层次聚类、DBSCAN 等算法能够根据客户的交易行为、风险偏好、资产配置等特征进行精准分群，为个性化金融服务提供基础支撑，同时也可用于识别市场中的不同交易模式和异常行为模式[2]。

2.3 时间序列分析与预测模型

长短期记忆网络(LSTM)和门控循环单元(GRU)在金融时序预测中具有显著优势，这类递归神经网络能够有效处理序列数据中的长期依赖关系，克服传统 RNN 的梯度消失问题。LSTM 通过门控机制选择性地保留和遗忘历史信息，特别适合捕捉金融时间序列中的周期性模式和长期趋势，在股价预测、汇率预测、商品价格预测等任务中表现优异。GRU 作为 LSTM 的简化版本，具有更少的参数和更快的训练速度，在某些金融预测任务中能够达到与 LSTM 相当的性能。注意力机制在多变量金融预测中的应用为模型带来了更强的解释性和预测能力，通过动态分配不同输入变量的权重，模型能够自动识别对预测目标最重要的特征和时间步，这对于理解复杂金融系统中的因果关系具有重要价值。Transformer 架构和自注意力机制的引入进一步提升了时序建模的效果，能够并行处理序列数据并捕捉长距离依赖关系。时间卷积网络(TCN)作为一种新兴的序列建模方法，通过膨胀卷积和残差连接实现了高效的时序特征提取，在金融数据建模中展现出良好的性能和可解释性。混合模型的集成策略通过组合不同类型的预测模型,如将传统时间序列模型(ARIMA、GARCH)与机器学习模型相结合，或将多个深度学习模型进行集成，能够充分利用各模型的优势，提高预测的准确性和稳健性，在实际的金融预测应用中取得了良好的效果。

3 应用实践与发展趋势

3.1 典型应用场景与案例分析

机器学习在金融领域已形成四大核心应用场景并取得显著成效。信用风险评估与智能授信系统通过整合客户多维度数据，运用 XGBoost、随机森林等算法构建精准风险模型，某大型银行的应用案例显示，相比传统评分卡模型，风险识别准确率提升 15%，审批时间从数天缩短至分钟级。算法交易与量化投资策略优化利用深度强化学习从海量市场数据

中识别交易机会，某量化对冲基金采用 LSTM 结合强化学习的策略，实现年化收益率超 20%，最大回撤控制在 5%以内[3]。反欺诈检测系统运用支持向量机和孤立森林算法实时分析交易模式，某支付平台的系统可在毫秒级完成风险评估，欺诈检测准确率超 99%。智能投顾服务则通过机器学习分析客户风险偏好和市场环境，提供个性化资产配置建议，实现风险收益的动态平衡。

3.2 技术挑战与未来发展方向

机器学习在金融数据挖掘中面临四大技术挑战与发展机遇。模型解释性与监管合规是首要挑战，深度学习模型的“黑盒”特性难以满足监管透明度要求，SHAP 值、LIME 等可解释 AI 技术为解决此问题提供新思路，未来需在保持性能的同时提升解释性。实时处理与高频数据挖掘面临计算压力和延迟限制，边缘计算、模型压缩、知识蒸馏等技术将提升实时处理能力[4]。联邦学习在数据隐私保护中前景广阔，允许多机构在不共享原始数据下协同训练，预计在跨机构风险建模、反洗钱等领域广泛应用。量子机器学习虽处于早期阶段，但在优化问题和复杂金融系统模拟方面具有指数级计算优势，随着量子计算发展，将在投资组合优化、风险建模、衍生品定价等领域发挥重要作用。

4 结论

本文系统探讨了机器学习在金融数据挖掘领域的理论基础、技术方法和实践应用。研究表明，机器学习技术能有效应对金融数据的高维性、非线性和动态变化特征，从支持向量机到深度神经网络的各种算法在金融领域都找到了合适应用场景。在信用风险评估、算法交易、反欺诈检测和智能投顾等典型应用中，机器学习不仅提升了业务效率和决策准确性，也为金融服务创新提供了技术支撑。然而，应用中仍面临模型可解释性、实时处理能力、数据隐私保护等挑战。随着可解释 AI、联邦学习、量子计算等前沿技术发展，机器学习在金融领域的应用将更加成熟，未来将朝着智能化、个性化、安全化方向演进。总体而言，机器学习与金融数据挖掘的深度融合代表了金融科技发展的重要方向，对推动金融行业数字化转型、提升服务质量、防范系统性风险具有重要意义，必将为构建高效、智能、安全的现代金融体系提供有力支撑。

[参考文献]

[1]刘光仿.金融业数据挖掘与机器学习应用实践[J].中国金融电脑,2024(1):12-15.

[2]谢荣,温蜜.基于差分隐私的敏感数据挖掘技术研究[J].上海电力学院学报,2020,036(004):401-407.

[3]夏鑫,牟玮,李艳芬,等.基于机器学习技术挖掘中医名家医案数据的方法探讨[J].医学新知,2024,34(4):448-457.

[4]许盛伟,牟健.基于机器学习的政务大数据定级技术研究[J].保密科学技术,2021(5):6.

作者简介：

曲思龙（1977.10-），男，汉族，黑龙江佳木斯人，硕士，实验师，研究方向为数据挖掘、数据安全方向。

