

基于模糊聚类的电商企业不平衡财务数据风险预测方法

成 耀¹, 张 铎¹, 周 宇¹, 何金凤^{1,2*}, 程 实¹

¹南通大学人工智能与计算机学院, 江苏 南通

²南通大学人工智能研究院, 江苏 南通

收稿日期: 2024年9月12日; 录用日期: 2024年10月17日; 发布日期: 2025年1月8日

摘 要

针对现行方法在电商企业不平衡财务数据风险预测中存在真负类率和召回率较低的问题, 提出基于模糊聚类的电商企业不平衡财务数据风险预测方法。先采用基于少数类样本的同类样本线性插值的过采样法对其进行处理, 生成平衡财务数据样本集, 然后从电商企业投资与收益、偿债能力、盈利能力、经营能力四个方面选取预测变量, 构建预测变量体系, 最后利用模糊聚类算法对预测变量数据集分类, 预测电商企业财务风险, 实现基于模糊聚类的电商企业不平衡财务数据风险预测。经实验证明, 设计方法真负类率和召回率均在95%以上, 可以实现对电商企业不平衡财务数据风险的精准预测。

关键词

模糊聚类, 电商企业, 不平衡财务数据, 风险预测, 过采样法, 偿债能力

A Method for Predicting the Risk of Imbalanced Financial Data in E-Commerce Enterprises Based on Fuzzy Clustering

Yao Cheng¹, Duo Zhang¹, Yu Zhou¹, Jinfeng He^{1,2*}, Shi Cheng¹

¹School of Artificial Intelligence and Computer Science, Nantong University, Nantong Jiangsu

²Artificial Intelligence Research Institute, Nantong University, Nantong Jiangsu

Received: Sep. 12th, 2024; accepted: Oct. 17th, 2024; published: Jan. 8th, 2025

Abstract

In response to the issues of low true negative rate and recall rate in the risk prediction of imbalanced

*通讯作者。

文章引用: 成耀, 张铎, 周宇, 何金凤, 程实. 基于模糊聚类的电商企业不平衡财务数据风险预测方法[J]. 电子商务评论, 2025, 14(1): 640-647. DOI: 10.12677/ec.2025.141078

financial data of e-commerce enterprises using current methods, this paper proposes a risk prediction method for imbalanced financial data of e-commerce enterprises based on fuzzy clustering. First, an oversampling method based on linear interpolation of similar samples of the minority class is used to process the data, generating a balanced financial data sample set. Then, prediction variables are selected from four aspects of e-commerce enterprises: investment and return, debt repayment ability, profitability, and operational capability, to construct a prediction variable system. Finally, the fuzzy clustering algorithm is used to classify the prediction variable dataset, predict the financial risk of e-commerce enterprises, and achieve risk prediction of imbalanced financial data of e-commerce enterprises based on fuzzy clustering. Experimental results prove that the designed method has a true negative rate and recall rate of over 95%, enabling precise prediction of the risk of imbalanced financial data in e-commerce enterprises.

Keywords

Fuzzy Clustering, E-Commerce Enterprises, Imbalanced Financial Data, Risk Prediction, Oversampling Method, Debt Repayment Ability

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在当今数字经济蓬勃发展的时代，电商企业作为其中的重要组成部分，其财务健康状况直接关系到企业的生存与发展。然而，电商企业在运营过程中，其财务数据往往呈现出高度的不平衡性。一方面，由于市场环境、销售策略、客户群体等多种因素的影响，企业的收入、成本、利润等财务指标波动较大；另一方面，由于财务管理水平、内部控制机制等方面的差异，不同电商企业的财务状况和风险水平也存在显著差异。这种不平衡性使得风险评估难以准确捕捉企业的真实风险状况，进而影响了风险预测的准确性和可靠性。因此，探索一种科学、合理、有效的电商企业不平衡财务数据风险预测方法显得尤为重要。

近几年，相关学者与专家针对企业财务风险预测方面展开了一系列研究，提出了一些方法与思路。文献[1]提出了基于大数据技术的预测方法，主要通过采集、处理和分析企业内外部的海量数据，识别财务报表中的潜在异常和风险点。利用大数据的关联性和预测能力，结合机器学习算法，对财务数据、市场数据、非财务数据进行深度挖掘，预测企业的财务状况、经营成果及未来发展趋势。文献[2]提出了基于 Z-Score 模型的预测方法，通过选取一系列财务指标(如经营利润与总资产比例、净营运资本与总资产比例等)，利用加权方法计算出一个综合得分，反映企业的财务状况和风险水平。虽然在企业财务风险预测方面取得了一定的研究成果，但是当企业面临财务数据分布不均、风险因子错综复杂时，传统方法往往难以有效应对，导致风险预测结果不准确，为此提出基于模糊聚类的电商企业不平衡财务数据风险预测方法。

2. 不平衡财务数据处理

由于记录财务交易时，数据输入错误，如金额记录错误、账户分类错误等，或者在编制财务报表时，如果遗漏了某些重要的交易或信息，会导致财务报表的数据不完整，导致财务数据不平衡，不平衡财务数据的存在会影响到最终风险预测的准确性，因此首先对电商企业不平衡财务数据处理，将原始财务数

据分为不平衡财务数据和平衡财务数据，财务数据的不平衡程度可以用不平衡率反映，其计算公式为：

$$\varepsilon = \frac{n_m}{n_s} \tag{1}$$

式中， ε 表示财务数据不平衡率； n_m 表示多数类财务数据数量； n_s 表示少数类财务数据数量。当不平衡率显著超出 1 时，这明确指示所处理的电商企业财务数据样本集呈现出“不平衡特性”[3]。不平衡率的具体数值越大，实际上反映构成样本集的两类(或多类)样本之间的比例差异愈发极端。在这种情境下，如果不对数据采取任何预处理，直接应用于风险预测任务中，将很可能导致预测结果出现偏颇[4]。因此对不平衡率大于 1 的财务数据集采用基于少数类样本的同类样本线性插值的过采样法对其进行处理，该方法不是简单地复制少数类样本以增加其数量，而是通过构建合成样本的方式来平衡数据集中的类别分布。以一个少数类样本(称之为“根样本”)为起点，随后随机选择该样本的一个或多个近邻样本作为辅助，通过在这两个样本之间进行线性插值，创造出全新的合成样本，具体流程如下：

输入电商财务数据集中少数类样本集，设定采样倍率为 c ，邻近个数为 j [5]。对于集合中的每一个样本计算其在剩余样本中的 k 个最近邻，并将这些近邻放入集合 x_{jk} 中。对于每个样本，根据预设的采样倍率，从 x_{jk} 中随机选择 t 个近邻样本。对于每个选中的近邻样本，生成一个新样本，其用公式表示为：

$$x_{new} = x_j + e * (x_{i,0} - x_j) \tag{2}$$

式中， x_{new} 表示生成的新的财务数据样本集； x_j 表示少数类样本集中选取的样本； e 表示[0, 1]区间内随机数； $x_{i,0}$ 表示邻近样本[6]。将所有生成的 x_{new} 样本加入到新的少数类样本集中。通过上述步骤，扩大原始财务数据集中少数类样本的数量，以此降低财务数据不平衡率，为后续风险预测奠定基础。

3. 预测变量选取

对于电商企业不平衡财务数据风险预测，首先选取具有针对性和代表性的变量，构建预测变量体系，结合电商企业财务指标，从电商企业投资与收益、偿债能力、盈利能力、经营能力四个方面预测风险，构建预测变量体系如下图 1 所示。

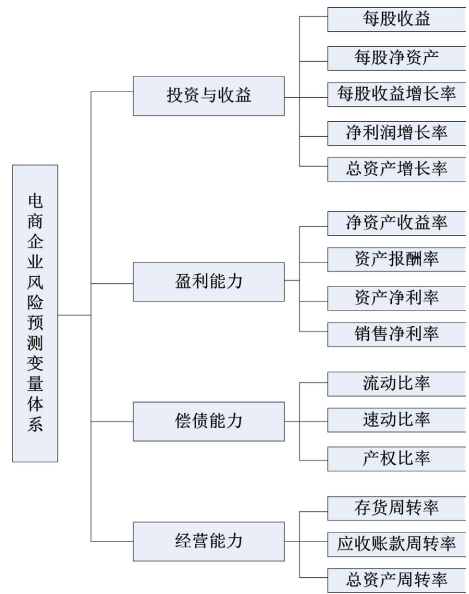


Figure 1. Variable system for risk prediction of e-commerce enterprises
图 1. 电商企业风险预测变量体系

如图 1 所示, 在电商企业投资与收益方面选择每股收益、每股净资产、每股收益增长率、净利润增长率以及总资产增长率五个变量, 结合处理后的电商企业财务数据确定以上五个变量, 其用公式表示为:

$$\begin{cases} EPS = S/N_{we} \\ NPS = X/V \\ BUN = (S_i - S_{i-1})/S_{i-1} \\ FVF = (A_i - A_{i-1})/A_{i-1} \\ NHB = (F_{er} - F_{fe})/F_{fe} \end{cases} \quad (3)$$

式中, EPS 表示电商企业投资股份每股收益; S 表示利润总额; N_{we} 表示股本总数; NPS 表示每股净资产; X 表示股东权益总额; V 表示股本总额; BUN 表示每股收益增长率; S_i 表示本期每股收益; S_{i-1} 表示上期每股收益; FVF 表示净利润增长率; A_i 表示当期净利润; A_{i-1} 表示上期净利润; NHB 表示总资产增长率; F_{er} 表示年末资产总额; F_{fe} 表示年初资产总额[7]。盈利能力体现电商企业通过经营电商活动创造利润的能力, 在盈利方面选取净资产收益率、资产报酬率、资产净利率以及销售净利率四个变量, 其用公式描述为:

$$\begin{cases} ROE = NM/VD \\ FVG = (NM + LK + JN)/VB \\ GBH = (NM - CV)/(F_{er} - F_{fe}) \\ VBF = NM/DR \end{cases} \quad (4)$$

式中, ROE 表示净资产收益率; NM 表示净利润; VD 表示平均净资产; FVG 表示资产报酬率; LK 表示利息费用; JN 表示所得税; VB 表示平均资产总额; GBH 表示资产净利率; CV 表示折旧费用; VBF 表示销售净利率; DR 表示销售收入[8]。偿债能力反映电商企业能否在短期和长期内动用流动资产来偿还流动负债, 从而保障企业的正常运营与资金链安全的能力, 在该方面选择流动比率、速动比率以及产权比率三个变量, 其计算公式为:

$$\begin{cases} ZX = MJ/UYU \\ ATP = (EW + HR + QW)/OI \\ ER = BF/RE \end{cases} \quad (5)$$

式中, ZX 表示流动比率; MJ 表示流动资产; UYU 表示流动负债; ATP 表示速动比率; EW 表示现金及现金等价物; HR 表示短期投资; QW 表示应收账款净额; OI 表示流动负债; ER 表示产权比率; BF 表示负债总额; RE 表示股东权益总额[9]。经营能力反映电商企业资产利用效率与经营效率, 在该方面选择存货周转率、应收账款周转率和总资产周转率三个变量, 其计算公式为:

$$\begin{cases} MK = PU/FSX \\ BG = MH/CS \\ VS = DR/VB \end{cases} \quad (6)$$

式中, MK 表示存货周转率; PU 表示销售成本; FSX 平均存货余额; BG 表示应收账款周转率; MH 表示营业收入; CS 表示平均应收账款余额; VS 表示总资产周转率。通过以上得到电商企业财务风险预测变量数据信息。

4. 基于模糊聚类的电商企业财务风险预测

对其企业风险预测, 主要是通过对预测变量综合分析, 将存在风险电商企业用“ST”, 将无风险的电商企业用“NT”标识, 因此电商企业不平衡财务数据风险预测问题, 实际为一个二分类问题, 此次在

以上基础上利用模糊聚类算法对电商企业财务风险预测变量数据集进行分类。模糊聚类核心在于通过量化的手段评估各样本间的内在联系,进而运用数学方法实现样本的客观、有效聚类[10]。具体而言,当电商企业财务预测变量数据样本集中的某个关键论域集数值显著增大时,这往往预示着财务数据中存在风险现象。相反,若论域集表现为孤立点,则表明该变量数据集中的信息较为深邃,不易直接通过常规手段进行挖掘[11]。为有效应对这一挑战,将论域集元素进行科学分类,并据此构建模糊矩阵,其用公式表示为:

$$\psi = (1 - \delta db_{ga})_{\mu \times \varphi} \tag{7}$$

式中, ψ 表示电商企业财务风险预测变量数据集聚类模糊矩阵; δ 表示模糊论域参数; d 表示马氏距离,衡量样本间差异性; b_{ga} 表示第 g 行 a 列论域; μ 表示模糊论域内模糊子集数量; φ 表示模糊子集序号。利用马氏距离衡量两个样本之间的相似程度,其计算公式为:

$$d = \sqrt{(y_{ga} - b_{ga})^T \kappa (b_{ga} - y_{ga})} \tag{8}$$

式中, y_{ga} 表示与 b_{ga} 不同的样本论域; T 表示聚类周期; κ 表示预测样本分布协方差矩阵[12]。马氏距离越小,意味着两个样本在多维空间中的位置越接近,即相似度越高。以各个样本为初始聚类中心,逐一计算其对应的收敛阈值,其计算公式为:

$$\omega = d_{\min} + \xi (d_{\min} + d_{\max}) \tag{9}$$

式中, ω 表示模糊聚类收敛速度阈值; $d_{\min}$ 、 $d_{\max}$ 分别表示最小和最大马氏距离; ξ 表示风险变量占总数据量的比例。将那些分类量小于收敛阈值的样本进行合并,从而不断迭代更新聚类中心。当模糊聚类达到最大迭代次数时,输出风险与非风险两个聚类中心样本集,即 ST 样本集与 NT 样本集,以此预测到电商企业是否会发展成为财务风险企业,实现基于模糊聚类的电商企业不平衡财务数据风险预测。

5. 实验论证

5.1. 实验准备与设计

采用对比实验的方式验证本文提出的基于模糊聚类的电商企业不平衡财务数据风险预测方法的有效性,将本文方法与文献[1]提出的基于大数据技术的预测方法和文献[2]提出的基于 Z-Score 模型的预测方法对比。以 2020 年 A 股上市电商企业的财务报表数据为实验数据样本,一共 254 家电商企业,为了保证预测难度,将实验数据集中删除当年净利润为负的样本,因此不平衡财务数据具有波动性,如果电商企业净利润为负,则财务报表数据连续两年均为亏损,这类数据不为不平衡财务数据,所以删除当年亏损的样本。剩余数据集包含 2451 个样本。按照以上对剩余不平衡财务数据风险预测,根据实际情况对模糊聚类参数设置:聚类数量设置为 2,模糊因子设置为 0.01,最大迭代次数设置为 100。通过对数据模糊聚类,得到平衡数据集,在此基础上对其进行综合分析,实验随机选取 10 个样本,其预测结果如下表 1 所示。

Table 1. Risk prediction results of imbalanced financial data in e-commerce enterprises
表 1. 电商企业不平衡财务数据风险预测结果

样本序号	预测	实际
1	ST	ST
2	NT	NT

续表

3	ST	ST
4	ST	ST
5	ST	ST
6	NT	NT
7	ST	ST
8	ST	ST
9	ST	ST
10	NT	NT

从上表中数据可以看出，设计方法预测结果与实际一致，可以完成电商企业不平衡财务数据风险预测任务。对于电商企业不平衡财务数据风险预测效果评价，采用真负类率和召回率作为指标，在此次实验中把预测为风险的样本定义为负类，把预测为正常的样本定义为正类，召回率与真负类率可以衡量预测方法能够预测风险与非风险样本的能力，其计算公式为：

$$\left\{ \begin{aligned} TNR &= \frac{TN}{TN + FP} \\ recall &= \frac{TP}{TP + FN} \end{aligned} \right. \tag{10}$$

式中，*TNR* 表示真负类率；*TN* 表示实际为负类被预测为负类的样本数量；*FP* 表示实际为负类但被预测为正类的样本数量；*recall* 表示召回率；*TP* 表示实际为正类被预测为正类的样本数量；*FN* 表示实际为正类但被预测为负类的样本数量。通过对比三种方法真负类率和召回率，评价电商企业不平衡财务数据风险预测精度。

5.2. 实验结果与讨论

三种方法在此次实验中针对电商企业不平衡财务数据风险预测的真负类率和召回率如表 2、表 3 所示。

Table 2. True negative rate of imbalanced financial data risk prediction for e-commerce enterprises (%)

表 2. 电商企业不平衡财务数据风险预测真负类率(%)

负类样本数量/个	本文方法	基于大数据技术的预测方法	基于 Z-Score 模型的预测方法
100	96.52	82.62	76.15
200	97.41	82.14	75.85
300	97.13	81.75	75.48
400	97.05	81.56	75.36
500	96.85	81.45	75.24
600	96.53	81.62	74.59
700	96.14	81.42	74.15
800	96.31	81.06	74.06
900	96.22	81.42	74.11
1000	95.94	80.62	73.62

从表 2 数据可以看出，在所有负类样本数量下，本文方法的真负类率均显著高于其他两种方法，稳定在 95%以上，这充分展示了本文方法在精准区分无风险样本方面的卓越能力。相比之下，基于大数据技术的预测方法虽然表现相对稳定，真负类率维持在 80%至 82%之间，但随着负类样本数量的增加略有波动，整体效果不及本文方法显著。至于基于 Z-Score 模型的预测方法，其表现则明显逊色，真负类率不仅最低，而且随着负类样本数量的增加还呈现出轻微的下降趋势，这表明该方法在识别无风险样本时的准确率较低，且其性能可能较大程度上受到样本量变化的影响。由此证明，本文方法在区分无风险样本方面具有明显优势。

Table 3. Recall rate of imbalanced financial data risk prediction for e-commerce enterprises (%)
表 3. 电商企业不平衡财务数据风险预测召回率(%)

正类样本数量/个	本文方法	基于大数据技术的预测方法	基于 Z-Score 模型的预测方法
100	98.48	76.62	85.61
200	98.36	74.15	85.24
300	98.14	73.25	85.06
400	97.86	73.15	85.42
500	97.68	73.05	85.24
600	97.51	72.89	85.22
700	97.35	72.84	84.59
800	97.15	72.65	84.36
900	97.06	72.47	84.57
1000	97.01	72.15	84.15

从表 3 数据可以看出，在所有正类样本数量下，本文方法的召回率均接近 97%，具有极高的风险识别能力。表明，本文方法能够极为有效地捕捉到绝大多数有风险的财务数据，为电商企业的风险管理提供了强有力的支持。相比之下，基于大数据技术的预测方法在召回率上表现较低，其召回率低于本文方法，且随着正类样本数量的增加略有下降，保持在 72%至 76%之间。由此说明该方法在识别有风险样本时存在较大的遗漏风险，无法全面覆盖所有潜在的风险点。而基于 Z-Score 模型的预测方法则表现中等，其召回率介于本文方法和基于大数据技术的预测方法之间，保持在 84%至 86%之间。综上所述，本文方法在召回率和真负类率上均展现出了显著的优势，能够有效识别出有风险的财务数据，为电商企业的风险管理提供更为可靠和准确的依据。

6. 结束语

此次结合相关文献资料，提出基于模糊聚类的电商企业不平衡财务数据风险预测方法，通过其独特的处理模糊性和不平衡数据的能力，为电商行业的风险管理开辟了新的视角。这一方法不仅提升了风险预测的精准度，还增强了企业对市场波动和内部变化的敏感度和反应速度。期待，在未来的应用中，该方法能够持续优化和完善，为电商企业提供更加全面、深入的风险洞察，助力企业在复杂多变的市场环境中稳健前行。同时，也鼓励更多学者和业界人士参与到相关研究中来，共同推动财务管理和风险预测技术的创新与发展。

参考文献

[1] 张蕊祿. 基于大数据技术的企业财务报表审计风险及防范措施[J]. 中国集体经济, 2024(19): 141-144.

-
- [2] 白皞柏. 基于 Z-Score 模型的上市公司财务风险评价——来自瑞幸咖啡公司 2017-2022 年财务数据的证据[J]. 中国市场, 2024(14): 115-122.
 - [3] 孙海丹. 大数据时代小微企业财务风险预警机制与实践探析[J]. 河北企业, 2024(5): 87-89.
 - [4] 李萍萍, 唐荣红, 袁建华. 财务会计制度视角下高校财务风险预警的实证研究——以 S 大学的财务数据为例[J]. 经济师, 2024(5): 86-87+89.
 - [5] 金晶, 陶杰. 基于深度神经网络与 Logistic 模型的财务风险预警研究——来自制造业上市公司数据[J]. 中国物价, 2024(3): 60-64.
 - [6] 邵永运, 张立莹. 制造业上市公司财务数据异常风险评分卡模型[J]. 沈阳师范大学学报(自然科学版), 2023, 41(6): 556-561.
 - [7] 王冲, 刘亚丽. Benford 与 XGBoost 模型融合的财务风险预警研究——基于 2000-2021 年沪深 A 股上市公司数据分析[J]. 财会研究, 2024(1): 35-46.
 - [8] 张彩凤. 建筑项目风险管理研究: 基于财务数据分析和风险评估的研究[J]. 市场瞭望, 2024(1): 37-39.
 - [9] 张丹瑶, 韩起远. 大数据背景下人工智能在企业财务风险预警中的应用[J]. 中国乡镇企业会计, 2023(11): 169-171.
 - [10] 江乾坤, 王成哲. 互联网企业海外并购财务风险大数据预警研究——基于 Stacking 集成学习[J]. 技术经济, 2023, 42(9): 147-160.
 - [11] 古丽思. 我国医药制造业上市公司财务风险预警研究——基于 Stacking 集成算法[J]. 中小企业管理与科技, 2023(8): 164-166.
 - [12] 阮赛男. 数字化背景下城市商业银行财务风险分析——以北京银行为例[J]. 财务管理研究, 2022(12): 50-58.