非均等误判代价下信用特征选择的 HHG-CB-CSR 协同模型

王静赛[1]，熊志斌[2]
1. 河南财政金融学院金融学院，河南 郑州 451464
2. 华南师范大学数学科学学院，广东 广州 510631
DOI:10.61369/ASDS.2025090010

摘　　要： 特征选择是信用评估的关键环节。针对信用数据的类别不平衡、非均等误判代价及类别型特征多等问题，本文首先基于代价敏感学习提出代价敏感查全率（CSR）指标；进而融合 Heller–Heller–Gorfine（HHG）检验与 CatBoost，构建 HHG–CB–CSR 信用特征选择方法——以 HHG 检验指导序列后向搜索，CatBoost 为学习器，CSR 为特征子集评价与停止准则。该方法可解决高基数类别型特征数值化难题，精准度量特征相关性并赋予选择过程代价敏感性。4 个信用数据集的实证表明，HHG–CB–CSR 在传统指标与 CSR 指标上均表现优异，稳健性与实际应用性突出。

关 键 词： 信用评估；特征选择；代价敏感查全率；HHG 检验；CatBoost

A Collaborative HHG-CB-CSR Framework for Feature Selection in Credit Scoring with Asymmetric Misclassification Costs

Wang Jingsai[1], Xiong Zhibin[2]
1.School of Finance, Henan Finance University, Zhengzhou, Henan 451464
2.School of Mathematical Sciences, South China Normal University, Guangzhou, Guangdong 510631

Abstract: Feature selection constitutes a pivotal stage in building effective credit scoring models. Confronting the typical challenges of credit datasets—namely class imbalance, asymmetric misclassification costs, and an abundance of categorical variables—this study initially introduces a novel metric termed Cost–Sensitive Recall (CSR), which is grounded in the principles of cost–sensitive learning. Building upon this, we formulate a collaborative feature selection framework, HHG–CB–CSR, by synergizing the Heller–Heller–Gorfine (HHG) test with the CatBoost algorithm. The framework employs the HHG test to direct a sequential backward search procedure, leverages CatBoost as the learning model, and adopts CSR as the definitive criterion for feature subset evaluation and termination. This methodology effectively tackles the numerical encoding of high–cardinality categorical features, enables precise quantification of feature dependencies, and endows the selection process with inherent cost–sensitivity. Empirical validation on four public credit datasets confirms that the HHG–CB–CSR approach yields outstanding results across both conventional and CSR–based evaluation metrics, underscoring its significant robustness and value for real–world applications.

Keywords: credit scoring; feature selection; cost–sensitive recall; HHG test; CatBoost

引言

信用风险防范是我国金融管理核心任务之一，提高信用风险管理水平不仅是风险防范的迫切需求，更对社会信用体系建设具有重大现实与战略意义。信用评估是信用风险管理的基础，其质量直接决定风险管理水平，因此改善并提高信用评估效果是学界与业界的关注重点。

当前信用评估方法按技术类型可分为两类：一是基于统计计量技术的模型，如判别分析法、Logistic 回归、Probit 模型等（Altman

基金项目：教育部人文社科研究规划基金（16YJA790053）；广东省普通高校特色创新类项目（人文社科）（2017WTSCX019）。
作者简介：
王静赛（1997.01–），男，汉族，河南南阳人，助教职称，河南财政金融学院专任教师，硕士学位，主要研究方向：信用风险管理与机器学习交叉领域，电子邮箱：18738952967@163.com；
熊志斌（1972.03–），男，汉族，江西宜黄人，华南师范大学数学科学学院副教授，博士学位，主要研究方向：智能算法与信用风险管理，电子邮箱：xiongzhibin@m.scnu.edu.cn。

等1994[1]；方匡南等2016[2]；王正位等2020[3]）；二是机器学习方法，其因无需严格假设、擅长处理非线性数据的优势被广泛应用，包括支持向量机、决策树、神经网络等。其中，决策树类模型（如随机森林、GBDT）及改进算法 XGBoost（Chen 等，2016）[4]、LightGBM（Ke 等，2017）[5]，因解释性强、超参数少等优点表现突出（Zhou 等，2019）[6]。王重仁和韩冬梅（2019）[7] 通过贝叶斯优化改进 XGBoost 超参数，提升了互联网信贷风险评估区分能力；朱磊等（2023）[8] 基于 LightGBM 构建企业信用预警模型，验证了树模型实用性。Prokhorenkova 等人（2018）[9] 提出的 CatBoost 算法（以对称决策树为基学习器的 GBDT 改进算法），因鲁棒性强、不易过拟合、擅长处理类别型特征，更受研究者青睐。但信用特征变量间存在强相关性与冗余性，直接使用会影响评估效果且增加数据采集成本，故特征选择成为信用评估建模的关键环节，其目的是剔除冗余变量，简化模型并提升预测准确性，同时降低数据成本。

信用评估变量选择以特征变量与目标变量（信用状况）的相关性、冗余性为基础。吴星泽（2011）[10] 指出，有预测能力的特征变量必与目标变量相关。传统变量筛选方法包括：一是基于相关性度量的搜索策略，如 Fritz 等（2000）[11] 的前向搜索、胡心瀚等（2012）[12] 的后向搜索及逐步回归；二是基于惩罚函数的模型驱动策略，如王小燕（2014）[13] 的 LASSO 方法、方匡南等（2014）[14] 的岭回归约束。近年来，王小燕等（2024）[15] 提出 CMCP-LMCL 深度神经网络，通过组变量选择同步处理高维信用数据的非线性相关性与冗余特征，提升筛选效率。上述方法均依赖变量间相关程度的精确识别，而有效度量复杂信用变量关联关系是核心问题。

现有相关性度量指标存在缺陷：互信息估计计算概率密度难且对异常值敏感（Sakar 等，2012）[16]；Pearson 相关系数仅适用于线性关系；Spearman 秩相关系数仅能度量线性与简单单调非线性关系，无法处理复杂非线性或非函数关系（曾津等，2017）[17]。Reshef 等（2011）[18] 基于互信息提出最大信息系数（MIC），可度量线性与各类非线性关系、挖掘非函数依赖关系，且解决大样本概率分布求解问题、抗异常值能力强（樊嵘等，2014）[19]，袁哲明等（2020）[20] 将其与冗余分摊策略结合用于特征选择，效果良好。但 MIC 仅适用于一维随机变量相关性度量且依赖大样本（Heller 等，2013）[21]；此外，信用数据中类别型特征（如学历、职业）经独热编码易引起维数灾难，传统变量选择还可能破坏类别变量完整性，降低模型解释性。Heller 等（2013）提出的 HHG 检验法，可检测任意维随机向量关联关系且适用于小样本，Santos 等（2013）[22] 实验表明其能识别函数 / 非函数相关及局部相关性，故 HHG 检验法适合信用变量选择。

模型预测性能评价方面，传统以准确度（Acc）为核心指标，但信用评估数据存在样本不平衡（好样本远多于坏样本）与非均等误判代价（坏样本误判为好的代价更高）两大问题。李妍峰和李文豪（2025）[23] 提出相对混合支持向量前沿方法，缓解样本不平衡影响，但 Acc 易因过度拟合多数类忽略风险信号，不适用于信用评估（Dushimimana 等，2020）[24]。平衡精度（Bacc）虽考虑样本不平衡，却无法反映非均等代价；代价敏感错误率（CSE）虽考虑非均等代价，但在极度不平衡样本中效果有限（Junior 等2020[25]；Ning 等2016[26]），故构建合适的评价指标亟待解决。

综上，针对信用数据样本不平衡、非均等代价、高维性及多类别型特征的特点，本文提出改进评价指标——代价敏感查全率（CSR），可在两类问题并存时客观反映模型预测情况；同时融合 HHG 检验与 CatBoost（CB）学习器，提出新信用变量选择方法——HHG-CB-CSR 模型。该模型可有效进行变量选择，且能处理样本不平衡与非均等代价问题，研究成果可为监管层、投资者决策及后续研究提供参考与借鉴。

一、改进的综合评价指标——代价敏感查全率指标的构建

在预测任务中，给定样本集

$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，$x_i \in X \subseteq R^n$，$y_i \in y \subseteq R^n$，i=1,2,…,m。

其中 y_i 是样本 x_i 对应的真实类别标签，用0代表信用好的样本（正例），1代表信用差的样本（负例）。预测与真实情况的对应关系如表1所示。

表1：二分类混淆矩阵

	预测结果	
真实情况	第0类	第1类
第0类	TP	FN
第1类	FP	TN

常用的模型评价指标主要有总体准确率（Accuracy，Acc）、查全率 (Recall，又称召回率)、查准率 (Precision) 及 F_a 指标等。其中 Recall 是以真实情况为基准的一种评价指标。记好样本和坏样本的查全率分别为 Recall_0（R_0）、Recall_1（R_1），具体定义为：

$$R_0=P\left(\hat{y}=0|y=0\right)=\frac{TP}{TP+FN},\ (R_0 \in [0,1]) \tag{1}$$

$$R_1=P\left(\hat{y}=1|y=1\right)=\frac{TN}{FP+TN},\ (R_1 \in [0,1]) \tag{2}$$

它们表示的是真实的好样本和坏样本中分别被预测正确的比例。

信用评估面临两大难题：首先，误判代价不等，将坏客户识别为好客户的损失远大于反向错误；其次，类别不平衡导致模型天然倾向于多数类（好客户），削弱了对少数坏客户的预测力。核心挑战在于，模型需在类别不平衡（好客户多，坏客户少）和非均等误判代价（将坏客户误判为好的代价极高）的双重约束

下，有效识别出关键的少数坏客户。为应对此问题，我们提出代价敏感查全率（Cost-Sensitive Recall, CSR），一个能同时反映这两大特性、并灵敏衡量模型对坏客户预测能力的评价指标。

分类任务的误判代价可以用代价矩阵来表示。以二分类为例，其代价矩阵如表2所示，其中 $cost_{ij}$（i=0,1; j=0,1）表示将第 i 类样本预测为第 j 类的代价。

令 $\beta=\frac{cost_{10}}{cost_{01}}$，CSR 的具体定义如下：

$$CSR=(1+\beta)\times\frac{R_0\times R_1}{\beta\times R_0+R_1} \tag{3}$$

表2：二分类代价矩阵

	预测类别	
真实类别	第0类	第1类
第0类	0	$cost_{01}$
第1类	$cost_{10}$	0

β 是一个先验参数，称为代价因子。对于信用评估问题，将坏客户误判为好客户的代价远高于将好客户误判为坏客户的代价，因此代价因子 β 通常设定为大于1的值，以加大对坏客户召回率的关注。代价因子使得 CSR 具有代价敏感性。相较于 F_β 或 Acc、AUC 等常用指标，CSR 指标具备以下优势：

1. 抗类别不平衡：传统指标（如 Acc、F_β）会被多数类主导，忽略对少数类的预测效果。CSR 通过对两类查全率进行调和平均，从根本上解决了这一偏见。

2. 代价敏感性：AUC 等指标不考虑误判代价的差异，不适用于风险评估。CSR 通过引入具有明确经济含义的代价因子 β，确保了模型选择的代价敏感性，使其更符合实际业务需求。

二、基于 CSR 指标的 HHG-CB-CSR 协同模型构建

（一）相关理论介绍

1. HHG 检验

HHG (Heller, Heller, and Gorfine) 检验是一种强大的非参数统计方法，它用于检测两个随机变量 X 和 Y 的独立性。其零假设为：

$$H_0:F(X,Y)=F(X)F(Y) \tag{4}$$

即 X 与 Y 相互独立。该检验基于样本点之间的成对距离来捕捉变量间的依赖结构。

假设 X 与 Y 不相互独立，并且有连续的联合密度函数。记 $d(\cdot,\cdot)$ 是 X 或 Y 的样本点之间的范数距离，那么 X 的任意两个分量之间的距离 $d(x_i,x_j)$ 也必然与 X 的分布一致，对于变量 Y 也是同样。令 $R_x=d(x_i,x_j)$ 和 $R_y=d(y_i,y_j)$，考虑下面两个随机变量：$I\{d(x_0,X)\le R_x\}$ 和 $I\{d(y_0,Y)\le R_y\}$，对于 N 次独立观测，记：

$$A_{11}=\sum_{i,j}^{\cdot\cdot}I\{d(x_0,X)\le R_x\}I\{d(y_0,Y)\le R_y\} \tag{5}$$

$$A_{12}=\sum_{k=1}^{N}I\{d(x_0,X)\le R_x\}I\{d(y_0,Y)> R_y\} \tag{6}$$

$$A_{21}=\sum_{k=1}^{N}I\{d(x_0,X)> R_x\}I\{d(y_0,Y)\le R_y\} \tag{7}$$

$$A_{22}=\sum_{k=1}^{N}I\{d(x_0,X)> R_x\}I\{d(y_0,Y)> R_y\} \tag{8}$$

假设有样本集 $\{(x_i,y_i)\}$，容量为 N。HHG 检验的核心思想是，对于任意一对样本点 i 和 j，以它们之间的距离 $d(x_i,x_j)$ 和 $d(y_i,y_j)$ 作为半径，构建一个2×2列联表来统计其余 N-2 个点 k 的分布情况。具体地，对于固定的 i 和 j（$i\neq j$），我们定义两个二元指示变量：$I\{d(x_i,X)\le d(x_i,x_j)\}$ 和 $I\{d(y_i,Y)\le d(y_i,y_j)\}$。通过遍历所有其他样本点 k（$k\neq i,j$），我们可以构建如表3所示的列联表。

表3：变量 $I\{d(x_i,X)\le d(x_i,x_j)\}$ 与 $I\{d(y_i,Y)\le d(y_i,y_j)\}$ 列联表

	$d(y_0,\cdot)\le d(y_i,y_j)$	$d(y_0,\cdot)>d(y_i,y_j)$	和
$d(x_0,\cdot)\le d(x_i,x_j)$	$A_{11}(i,j)$	$A_{12}(i,j)$	$A_{1\cdot}(i,j)$
$d(x_0,\cdot)>d(x_i,x_j)$	$A_{21}(i,j)$	$A_{22}(i,j)$	$A_{2\cdot}(i,j)$
和	$A_{\cdot1}(i,j)$	$A_{\cdot2}(i,j)$	N-2

基于此列联表，计算其皮尔逊卡方统计量 $S(i,j)$，公式如下：

$$S(i,j)=\frac{(N-2)[A_{12}(i,j)A_{21}(i,j)-A_{11}(i,j)A_{22}(i,j)]^2}{A_{1\cdot}(i,j)A_{2\cdot}(i,j)A_{\cdot1}(i,j)A_{\cdot2}(i,j)} \tag{9}$$

当分母中任意一项为0时，定义 $S(i,j)=0$。$S(i,j)$ 量化了以点 i 和 j 为锚点时所观察到的局部依赖性。

为了得到一个总体的检验统计量，HHG 将所有 N(N-1) 个可能的点对 (i,j) 所计算出的 $S(i,j)$ 值进行求和：

$$G=\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}S(i,j) \tag{10}$$

统计量 G 综合了所有局部依赖性的证据，其值越大，表明反对独立性原假设的证据越强。

为确定 G 值的统计显著性，HHG 检验采用置换检验（Permutation Test）来计算 P 值。具体方法是：保持变量 X 的样本序列不变，对变量 Y 的样本序列进行 L 次随机重排（打乱顺序），每次重排后都计算一个新的统计量 G。这样便得到了一个在零假设（独立性）下 G 的经验分布。原始样本计算出的 G 值在该经验分布中的分位数即为 P 值。

对于给定的显著性水平 α，当 $P<\alpha$ 时，我们拒绝原假设 H_0，即有充分的统计证据认为变量 X 与 Y 不独立。

2. CatBoost 算法

CatBoost（Categorical Boosting）是一种基于梯度提升决策树（Gradient Boosting Decision Tree, GBDT）框架的高性能机器学习算法，由俄罗斯科技公司 Yandex 于2017年提出。该算法在标准 GBDT 的基础上进行了多项关键创新，旨在原生、高效地处理类别型特征，并有效抑制过拟合。其核心创新之一在于有序目标统计（Ordered Target Statistics）方法，用于高效处理类别型特征。与独热编码（One-Hot Encoding）或哈希编码等传统方法不同，OTS 方法通过目标变量的统计信息将类别标签转换为数值特征。具体而言，对于某个类别特征的第 i 个样本取值，其转换

值计算如下所示：

$$X'_{j,i}=\frac{\sum_{j=1}^{q} I\left(x'_{j,i}=x'_{k,i}\right)\cdot y'_j+\lambda\cdot P}{\sum_{j=1}^{q} I\left(x'_{j,i}=x'_{k,i}\right)+\lambda} \tag{11}$$

其中 $x'_{j,i}$ 代表第 j 行第 i 列的特征取值 $x'_{j,i}$ 的数值化结果 (OTS 值)，P 是目标变量的先验值（通常取数据集均值），λ> 为平滑参数，这种数值化处理方法可以较好地消除噪声数据干扰以及防止过拟合。

另一关键特性是 CatBoost 采用了有序提升（Ordered Boosting）机制，通过构建与训练数据独立排列的多个模型来克服梯度偏差问题，显著减少了预测偏移，提升了模型泛化能力。此外，CatBoost 采用对称树（Oblivious Trees）作为基学习器，这种树在同一层使用相同的分裂规则，不仅加快了预测速度，也降低了模型复杂度。算法还内置了对类别不平衡问题的处理能力，可通过调整类别权重或使用代价敏感学习来优化少数类的分类性能。

由于其卓越的鲁棒性和对类别型特征的天然支持，CatBoost 尤其适用于信用风险评估、推荐系统等包含大量类别属性的实际任务，并在多个公开基准测试中表现出色。

（二）HHG-CB-CSR 协同模型

在构建特征选择模型时，一个核心挑战是如何平衡计算效率与搜索的完备性。穷尽所有特征组合的纯包装法（Wrapper）在计算上成本过高。而纯过滤法（Filter）虽然高效，却忽略了特征间的交互与冗余。

为此，本研究提出一种过滤 - 包装混合（Filter-Wrapper Hybrid）的协同策略。该策略的核心思想是：利用 HHG 检验的 P 值作为一种高效的启发式信息，来指导和简化后续基于 CatBoost 模型的包装法搜索过程。具体而言，我们假设与目标变量独立性越强的特征（即 P 值越大），其对最终模型的贡献越小，因此应被优先考虑剔除。这种基于 P 值的排序并非最终的裁决标准，而是为后续的迭代剔除提供一个合理的、有序的搜索路径。最终特征子集的选择，依然由包含多变量交互信息的 CatBoost 模型性能（CSR 得分）来决定。该方法在保证较高计算效率的同时，通过包装法评估来捕获特征间的复杂关系，从而寻求在效率和效果之间的最佳平衡。

由于 HHG 检验具有强大的变量相关性检测能力，而 CatBoost 作为一种树基集成模型算法，具备很强的稳健性和类别型特征数据处理能力；本研究在基于 CSR 评价指标基础上，将 HHG 检验法与 CatBoost 相结合，构建了 HHG-CB-CSR 协同模型，具体方法如下：

设原始数据集 D 含有 v 个特征，m 个样本，特征矩阵为 X，X 是 m 行 v 列的矩阵，目标变量为 y，是 m 行 1 列的矩阵。用 X_i (i=1,2…v) 表示第 i 个特征，则 $X=(X_1,X_2,X_3,?\ X_v)_{m?v}$，$D=(X,y)_{m\times(v+1)}$，设定显著性水平为 α，后向搜索的步长为 h ∈ (0,1)。

HHG-CB-CSR 模型变量选择方法具体步骤为：

步骤 1：遍历所有特征变量 X_i 与目标变量 y 进行 HHG 检验，得到 v 个特征的检验 P 值；

步骤 2：将 v 个特征的 P 值与 α 进行比较，若 P 值大于等于 α 则将对应变量划分到剔除变量目标池，否则划分到保留变量目标池；

步骤 3：使用序列后向搜索策略，对剔除变量目标池的特征变量根据 P 值从大到小按照搜索步长 h 进行逐级剔除，每次剔除后都将剩余特征数据输入 CatBoost 分类器，获得交叉验证的 CSR 平均得分；

步骤 4：重复步骤 3，直至遍历剔除变量目标池的所有特征变量。将获得的 $\frac{1-\alpha}{h}$ 个 CSR 得分与全特征 CSR 得分进行比较，选取最高得分对应的保留特征集合作为最优特征子集。

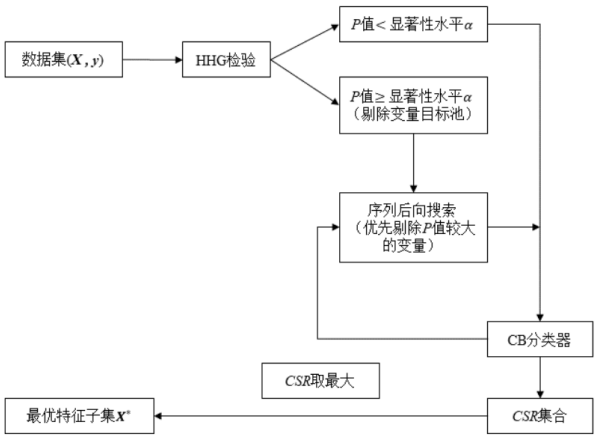

图 1：HHG-CB-CSR 方法流程图

三、HHG-CB-CSR 模型在信用评估中的实证研究

（一）样本数据来源与描述

本文选取 4 个公开的信用评估数据集进行实证研究，分别是：阿里云天池平台的“贷款违约预测”数据集（D_1）、UCI 机器学习库中的“German Credit Data”数据集（D_2）和“Default of Credit Card Clients Dataset”数据集（D_3）以及 Kaggle 平台的“Credit Card Fraud Detection”数据集（D_4）。如表 4 所示，这四个数据集的共同特点是存在显著的样本类别不平衡问题和大量类别型特征，坏样本比例从 30.00% 到仅 0.17% 不等，这为检验模型处理不平衡数据的能力提供了合适的实验环境。

表 4：四种信用数据集样本状况一览表

数据集	样本总量	信用好样本数量	信用差样本数量	信用差样本比例
D_1	80 万	64.04 万	15.96 万	19.95%
D_2	1000	700	300	30.00%
D_3	30000	23364	6636	22.12%
D_4	28.48 万	28.43 万	492	0.17%

（二）基于 HHG-CB-CSR 方法的信用评估建模

为验证所提 HHG-CB-CSR 方法的有效性，本文选取方差阈

值法 (VTFS)、最大信息系数法 (MICFS)、递归特征消除法 (RFE) 作为对比，并引入使用全部变量的模型 (ALL-CB) 作为基准。所有方法筛选出的特征子集均输入到经过超参数优化的 CatBoost 分类器中进行性能评估。

本文提出的 HHG-CB-CSR 方法结合了 HHG 检验的非线性相关性评估能力与后向搜索策略，以最大化交叉验证的 CSR 为目标，自适应地筛选特征。不同方法最终选择的变量数量汇总于表5。

表5：各特征选择方法在不同数据集上选择的变量数

数据集	原始变量数	VTFS	MICFS	RFE	HHG-CB-CSR
D_1	45	41	42	42	42
D_2	20	17	11	13	15
D_3	23	15	16	16	21
D_4	30	13	30	15	15

（三）实证研究结果

本文通过几种经典评价指标和本研究侧重的 CSR 指标，对5种模型在4个数据集上的预测性能进行了综合评估。总体而言，本文提出的 HHG-CB-CSR 模型在多数情况下表现出最优或接近最优的性能，尤其在处理类别不平衡问题和控制误分类成本方面具有显著优势。

各模型在 D_1 数据集上的性能对比如表6所示。

表6：各模型在数据集 D_1 上的性能对比 (%)

模型	Acc	BAcc	AUC	R_1	平均 CSR
VTFS-CB	70.50	70.02	74.18	69.23	69.44
MICFS-CB	67.50	68.16	73.94	69.23	68.94
RFE-CB	72.00	69.98	74.10	66.67	67.48
ALL-CB	68.50	68.78	72.78	69.23	69.11
HHG-CB-CSR	73.00	72.54	73.99	71.80	71.99

从表6可以看出，本文提出的 HHG-CB-CSR 模型在 D_1 上的综合表现最佳。不仅在 Acc、BAcc 和 R_1 等多个传统指标上取得领先，更重要的是，在为不平衡问题设计的平均 CSR（β=3,5,10,50）指标上达到了71.99%，显著高于次优模型（69.44%），证明了其在控制误分类成本、精准识别少数类（坏）样本方面的卓越能力。

为验证模型的稳健性与普适性，我们在 D_2、D_3、D_4 上进行了测试。表7汇总了各模型在关键指标上的表现。

表7：各模型在另外3个数据测试样本集上的预测结果对比

（经典评价指标）(%)

数据集	模型	Acc	R_0	R_1	BAcc	AUC
D_2	VTFS-CB	66.50	63.12	74.58	68.85	75.23
	MICFS-CB	64.00	62.41	67.80	65.10	75.29
	RFE-CB	64.00	60.99	71.86	66.09	74.82
	ALL-CB	65.00	63.12	69.49	66.31	74.76
	HHG-CB-CSR	68.00	65.25	74.58	69.91	76.28
D_3	VTFS-CB	74.96	77.94	64.64	71.29	77.73
	MICFS-CB	75.51	78.81	64.04	71.43	78.25
	RFE-CB	76.36	79.77	64.55	72.16	78.49
	ALL-CB	76.08	79.38	64.64	72.01	78.19
	HHG-CB-CSR	76.08	79.38	64.64	72.01	78.19
D_4	VTFS-CB	98.56	98.58	86.21	92.39	96.69
	MICFS-CB	98.68	98.70	86.21	92.45	98.75
	RFE-CB	98.59	98.61	86.21	92.41	98.47
	ALL-CB	98.68	98.70	86.21	92.45	98.75
	HHG-CB-CSR	98.33	98.34	89.66	94.00	98.38

表8：各模型在另外3个数据集上的 CSR 指标对比 (%)

数据集	模型	CSR				平均 CSR
		β=3	β=5	β=10	β=50	
D_2	VTFS-CB	71.34	72.39	73.37	74.32	72.86
	MICFS-CB	66.37	66.84	67.27	67.69	67.04
	RFE-CB	68.79	69.79	70.71	71.61	70.23
	ALL-CB	67.78	68.34	68.86	69.35	68.58
	HHG-CB-CSR	72.01	72.84	73.62	74.37	73.21
D_3	VTFS-CB	67.52	66.53	65.66	64.86	66.14
	MICFS-CB	67.19	66.10	65.15	64.28	65.68
	RFE-CB	67.78	66.67	65.69	64.79	66.23
	ALL-CB	67.79	66.70	65.75	64.88	66.28
	HHG-CB-CSR	67.79	66.70	65.75	64.88	66.28
D_4	VTFS-CB	89.00	88.05	87.20	86.42	87.67
	MICFS-CB	89.03	88.07	87.21	86.42	87.68
	RFE-CB	89.01	88.06	87.21	86.42	87.68
	ALL-CB	89.03	88.07	87.21	86.42	87.68
	HHG-CB-CSR	91.68	91.00	90.39	89.82	90.72

结果进一步证实了本文方法的优势。如表8所示，HHG-CB-CSR 模型在所有3个数据集的平均 CSR 指标上均排名第一（或并列第一）。尤其在欺诈样本比例极低的 D_4 数据集上，其 CSR 值（90.72%）相比其他模型（约87.68%）实现了近3个百分点的提升，充分表明了该方法在极端不平衡场景下的高效性与稳健性。

四、结束语

鉴于信用数据通常具有类别不平衡性和误判代价不均等性，以及包含大量类别型特征，本文提出了一种综合评价指标——代价敏感查全率（CSR），作为特征子集的评价函数，使特征选择具备代价敏感性。同时，本文将 HHG 检验与 CatBoost 和 CSR 指标结合，提出了 HHG-CB-CSR 协同模型。HHG 检验算法在特征选择过程中相较于其他相关性度量方法具有显著优势。此外，利用 CatBoost 内置的 OTS 方法处理类别型特征，有效解决了传统方法在处理大量类别型特征时易出现的维数灾难问题。通过在4个不同信用数据集上的实证研究，结果表明，HHG-CB-CSR 协同模型在信用评估任务中具有良好的稳健性和泛化能力。在 CSR 指

标对比中，本文提出的方法给出的最优特征子集更偏重于误判代价大的类别，更关注“坏样本”的预测准确性。与大多数研究仅依据分类精度、AUC 等不具备代价敏感性的指标评价模型效果不同，本文的方法更契合信用评估的实际应用需求。

当然，本文方法也存在局限性。例如，CSR 评价指标仅探讨了二分类任务，对于多分类任务，还需给出 CSR 的广义化定义并拓展到多分类信用评估研究中。此外，HHG-CB-CSR 方法具有较高的计算复杂度，有待进一步优化以降低计算成本。

参考文献

[1] Altman E I, Marco G, Varetto F. Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks [J]. Journal of Banking and Finance, 1994, 18(3): 505-529.

[2] 方匡南 , 范新妍 , 马双鸽 . 基于网络结构 Logistic 模型的企业信用风险预警 [J]. 统计研究 , 2016, 33(4): 50-55.

[3] 王正位 , 周从意 , 廖理 , 等 . 消费行为在个人信用风险识别中的信息含量研究 [J]. 经济研究 , 2020, (1): 149-163.

[4] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System [C]. Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[5] Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [C]. In Advances in Neural Information Processing Systems, 2017: 3149-3157.

[6] Zhou Z H, Feng J. Deep Forest: Towards An Alternative to Deep Neural Networks [J]. National Science Review, 2019, 6(1): 74-86.

[7] 王重仁 , 韩冬梅 . 基于超参数优化和集成学习的互联网信贷个人信用评估 [J]. 统计与决策 , 2019(1): 87-91.

[8] 朱磊 , 应瑛 , 陈怡桐 , 等 . 基于 LightGBM 和 SHAP 值的企业信用预警模型和实证分析 [J]. 征信 , 2023, 41(11): 49-56.

[9] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: Unbiased boosting with categorical features [C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2018: 6639-6649.

[10] 吴星泽 . 财务危机预警研究 : 存在问题与框架重构 [J]. 会计研究 , 2011(2): 59-65, 97.

[11] Fritz S, Hosemann D. Restructuring the Credit Process: Behaviour Scoring for German Corporates [J]. Intelligent Systems in Accounting Finance & Management, 2000, 9(1): 9-21.

[12] 胡心瀚 , 叶五一 , 缪柏其 . 上市公司信用风险分析模型中的变量选择 [J]. 数理统计与管理 , 2012, 31(6): 1117-1124.

[13] 王小燕 , 方匡南 , 谢邦昌 . Logistic 回归的双层变量选择研究 [J]. 统计研究 , 2014, 31(9): 107-112.

[14] 方匡南 , 章贵军 , 张惠颖 . 基于 Lasso-logistic 模型的个人信用风险预警方法 [J]. 数量经济技术经济研究 , 2014, 31(2): 125-136.

[15] 王小燕 , 江建伟 , 姚欣悦 . 基于 CMCP-LMCL 的多分类深度神经网络及其应用 [J]. 统计研究 , 2024, 41(7): 148-160.

[16] Sakar C O, Kursun O. A Method for Combining Mutual Information and Canonical Correlation Analysis: Predictive Mutual Information and Its Use in Feature Selection [J]. Expert Systems with Applications, 2012, 39(3): 3333-3344.

[17] 曾津 , 周建军 . 高维数据变量选择方法综述 [J]. 数理统计与管理 , 2017, 36(4): 678-692.

[18] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting Novel Associations in Large Data Sets [J]. Science, 2011, 334(6062): 1518-1524.

[19] 樊嵘 , 孟大志 , 徐大舜 . 统计相关性分析方法研究进展 [J]. 数学建模及其应用 , 2014, 3(1): 1-12.

[20] 袁哲明 , 杨晶晶 , 陈渊 . 基于最大信息系数与冗余分摊的特征选择方法 [J]. 计算机工程 , 2020, 46(8): 101-105.

[21] Heller R, Heller Y, Gorfine M. A Consistent Multivariate Test of Association Based on Ranks of Distances [J]. Biometrika, 2013, 100(2): 503-510.

[22] Santos S S, Takahashi D Y, Nakata A, et al. A Comparative Study of Statistical Methods Used to Identify Dependencies between Gene Expression Signals [J]. Briefings in Bioinformatics, 2014, 15(6): 906-918.

[23] 李妍峰 , 李文豪 . 面向信用风险评估的相对混合支持向量前沿方法研究 [J/OL]. 系统科学与数学 , 1-21 [2025-10-11].

[24] Dushimimana B, Wambui Y, Lubega T, et al. Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans [J]. Journal of Risk and Financial Management, 2020, 13(8): 180.

[25] Junior L M, Nardini F M, Renso C, et al. A Novel Approach to Define the Local Region of Dynamic Selection Techniques in Imbalanced Credit Scoring Problems [J]. Expert Systems with Applications, 2020, 152: 113351.

[26] Ning C, Ribeiro B M, An C. Financial Credit Risk Assessment: A Recent Review [J]. Artificial Intelligence Review, 2016, 45(1): 1-23.

