Operations Research and Fuzziology 运筹与模糊学, 2023, 13(2), 447-456
Published Online April 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.132044

Hans汉斯

基于mRMR-XGB-LS算法的信用评估分类模型

刘文杰

上海工程技术大学数理与统计学院，上海

收稿日期：2023年2月13日；录用日期：2023年3月31日；发布日期：2023年4月7日

摘　要

信用评估领域时刻都在产生大量数据，一方面，这些数据中隐藏的数据结构具有很大价值；另一方面，信用数据通常存在着高维冗余、缺乏标签的特点，无法直接进行分析和研究。因此，为了去除信用数据冗余性和挖掘数据结构，本文结合过滤法高效简单、嵌入式分类性能优越和无监督特征选择方法不需要样本类别标签的优点，提出一种基于mRMR、XGBoost和拉普拉斯得分算法的信用评估分类模型。首先，对半监督数据中的有标记训练集分别执行mRMR算法和XGBoost算法，在有标记和无标记训练集上执行拉普拉斯得分算法，分别得到特征的排序。其次，根据特征的排序分别赋予特征相应的权重，并进行简单求和得到每个特征的最终权重。接着，按照权重大小选出最优的特征子集，去除无关冗余的特征。最后，基于XGBoost、LightGBM和CatBoost构建信用评估分类模型，以*G*值和*F*值来度量不同特征选择方法下模型的分类性能。实验结果表明，本文模型在不同的标记样本率和数据集下均具有较高的*G*值和*F*值，能够有效筛选特征、减少数据冗余性，提高数据的分类性能和对少数类的识别能力。

关键词

信用评估，半监督特征选择，分类模型，XGBoost，mRMR，拉普拉斯得分

Credit Evaluation Classification Model Based on mRMR-XGB-LS Algorithm

Wenjie Liu

School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai

Received: Feb. 13th, 2023; accepted: Mar. 31st, 2023; published: Apr. 7th, 2023

Abstract

A large amount of data is generated in the field of credit evaluation all the time. On the one hand, the data structure hidden in these data is of great value. On the other hand, credit data is usually

文章引用: 刘文杰. 基于 mRMR-XGB-LS 算法的信用评估分类模型[J]. 运筹与模糊学, 2023, 13(2): 447-456.
DOI: 10.12677/orf.2023.132044

characterized by high dimensional redundancy and lack of labels, so it cannot be directly analyzed and studied. Therefore, in order to remove the redundancy of credit data and mine the data structure, a credit evaluation classification model based on mRMR, XGBoost and Laplacian score algorithm is proposed in this paper, combining the advantages of efficient and simple filtering method, superior embedded classification performance and unsupervised feature selection method without sample category label. Firstly, the mRMR algorithm and XGBoost algorithm were respectively performed on the labelled training sets in the semi-supervised data, and the Laplacian score algorithm was performed on the labelled and unlabelled training sets to get the feature ordering respectively. Secondly, according to the ranking of features, the corresponding weights are assigned to each feature, and the final weights of each feature are obtained by simple sum. Then, the optimal feature subset is selected according to the weight size, and the irrelevant redundant features are removed. Finally, a classification model was constructed based on XGBoost, LightGBM and CatBoost, and *G* and *F* values were used to measure the classification performance of the model under different feature selection methods. Experimental results show that the model presented in this paper has high *G* and *F* values under different labeling sample rates and data sets, which can effectively screen features, reduce data redundancy, improve data classification performance and recognition ability of minority classes.

Keywords

Credit Assessment, Semi-Supervised Feature Selection, Classification Model, XGBoost, mRMR, Laplacian Score

1. 引言

企业信用评估可以被看作通过模型分析企业信用数据来评估信用风险的过程，其本质是一个二分类问题。现实中标注样本类别的成本十分昂贵，信用数据中往往含有少量的标记数据和大量的未标记数据。在实际应用中如何有效利用大量的无标记样本和少量的有标记样本去除信用数据冗余性、挖掘数据结构无疑具有重要的实践意义。

特征选择是在不损失数据信息的前提下，去除冗余的可能会对模型产生负面影响的特征，使得最终所选的特征子集是最优的。基于特征子集评价策略的不同，特征选择可以分为：过滤式、包裹式以及嵌入式[1]。基于不同监督信息，特征选择可以分为有监督特征选择、无监督特征选择和半监督特征选择[2]。传统的半监督特征选择方法包括：基于伪标签的方法、基于图的方法[3]、基于 SVM 的方法[4]以及其他半监督特征选择方法[5]。

在信用评估领域中的特征选择方法以有监督特征选择方法为主，包括：方差、信息增益[6]、信息价值和 Pearson 相关系数[7]等。然而，有监督特征选择方法需要获得大量的具有标记信息的样本，运算成本高，无法对未知数据进行处理。数据添加标签成本巨大，耗费时间长，信用数据中的大量无标记数据被浪费，其隐藏的有效信息无法被合理利用。区别于有监督的特征选择算法，无监督的特征选择算法运算成本大幅降低，不需要事先对数据做标签，且能够挖掘数据的潜在特征，但训练样本的歧义性高。其中，拉普拉斯得分(Laplacian Score, LS)算法[8]是一种典型的无监督特征选择算法。

单一的特征选择方法从单一维度以某个特定的评价指标来对特征进行筛选，获得局部最优的特征子

集，而集成多种特征选择方法的特征子集可以获得近似全局最优的特征子集。此外，集成特征子集还可以提高算法的稳定性，降低特征子集的不稳定性[9]。传统的特征集成方法包括：简单平均、加权平均、Borda 计数法、投票法和 SVM-Rank 等[10]。Wang 等[11]对特征选择算法进行集成时发现，对少数算法进行集成时模型的预测性能要优于集成所有特征选择算法的模型。

为了更好地对信用评估数据进行预测，针对存在标记样本和未标记样本的半监督数据，本文提出一种基于 mRMR、XGBoost 和拉普拉斯得分(mRMR-XGBoost-LS)算法的信用评估分类模型。在特征选择阶段，结合了过滤式泛化性能好、计算开销小、嵌入式算法效率高、性能好和无监督特征选择算法不需要数据标签的优点进行特征选择。首先，在半监督数据的有标记训练集上执行有监督的过滤式算法 mRMR 和嵌入式算法 XGBoost，在有标记和无标记训练集上执行无监督的过滤型算法拉普拉斯得分法，分别得到特征的排序。其次，根据特征的排序分别赋予特征相应的权重，并简单求和得到特征最终的权重。最后，按照权重大小选出最优的特征子集。在分类阶段，分别基于 XGBoost、LightGBM 和 CatBoost 算法构建信用评估分类模型，以 G 值和 F 值来度量不同特征选择方法、不同分类算法下信用评估模型的分类性能。

本文将采用未筛选的特征子集和三种特征选择算法所构建的信用评估分类模型以及本文提出的基于 mRMR-XGB-LS 算法的信用评估分类模型进行研究对比，结果表明本文所提信用评估分类模型不仅大大减少了无关冗余特征，有效避免单一特征选择方法的不稳定性，而且还能对半监督数据进行有效处理，提高信用评估分类模型的泛化能力。

2. 相关理论

2.1. mRMR 算法

本模板过滤法不依赖特征分类器，通用性强，特征子集冗余度较低，但所选特征子集在分类精确度方面通常低于包裹法和嵌入法。mRMR [12]是一种典型的过滤式特征选择算法，该算法同时考虑特征之间的相关性以及特征与目标变量之间的相关性。mRMR 以不同的方式对特征之间的相关性和冗余性进行衡量。用 f_i 表示数据集 S 的第 i 个特征，度量 S 与类别 c 之间的最大相关最小冗余的两种方式如下：

$$\max_S \left\{ \frac{1}{|S|}\sum_{f_i \in S} R(f_i, c) - \frac{1}{|S|^2}\sum_{f_i, f_j \in S} D(f_i, f_j) \right\} \tag{1}$$

$$\max_S \left\{ \frac{1}{|S|}\sum_{f_i \in S} R(f_i, c) \Big/ \frac{1}{|S|^2}\sum_{f_i, f_j \in S} D(f_i, f_j) \right\} \tag{2}$$

研究表明[12]选择特征(2)式比(1)式更加有效，已有研究多选择式(2)进行研究[13] [14]。针对 mRMR 算法中的冗余性度量函数，罗康洋和王国强[14]构造两个新的选入第 k 个特征的评价函数：

FACQ (F-test AC quotient)：

$$\max_{f_k \in F - S_{k-1}} \left\{ F(f_k, y) \Big/ \frac{1}{|S_{k-1}|}\sum_{f_j \in S_{k-1}} AC(f_k, f_j) \right\} \tag{3}$$

FDAQ (F-test DAC quotient)：

$$\max_{f_k \in F - S_{k-1}} \left\{ F(f_k, y) \Big/ \frac{1}{|S_{k-1}|}\sum_{f_j \in S_{k-1}} 1/\left(1 - AC(f_k, f_j)\right) \right\} \tag{4}$$

其中，

$$AC\left(f_k,f_j\right)=\frac{\left|\sum_{l=1}^{n}f_{kl}f_{jl}\right|}{\sqrt{\sum_{l=1}^{n}f_{kl}^2\sum_{l=1}^{n}f_{jl}^2}} \tag{5}$$

为绝对值余弦度量。该评价函数能够更加容易的识别出变量之间的冗余信息，得到更为简洁的特征子集。因此，本文在构建集成型特征选择的方法时选择文献[14]中改进的 mRMR 特征选择算法。

2.2. XGBoost 特征选择算法

本模板 XGBoost [15]的基础是梯度提升算法。相比于传统的梯度提升算法，XGBoost 求解损失函数极值时使用泰勒二阶展开，另外在损失函数中加入了正则化项，使得算法收敛速度更快、求解效率更高。XGBoost 算法在构建树的过程中，贪婪的选择能使分割后整棵树增益值最大的特征作为叶子节点，即对树分割时选择当前使得信息增益最大的特征。信息增益的计算如下所示：

$$Gain=\frac{1}{2}\left[\frac{g_L^2}{h_L^2+\lambda}+\frac{g_R^2}{h_R^2+\lambda}-\frac{\left(g_L+g_R\right)^2}{\left(h_L+h_R\right)^2+\lambda}\right]-\gamma \tag{6}$$

因此，特征被分割的次数越多，其对树模型的增益越大，该特征重要度越大。XGBoost 特征选择算法中对特征的筛选取决于各个特征对模型贡献的重要度，即特征用于树分割次数的总和。XGBoost 通过统计特征在树模型构建过程中被用于分割的次数总和来确定特征的重要度，通过对特征的重要度从高到低排序进行特征选择。XGBoost 算法作为典型的嵌入式算法，运算效率高，在以往研究中被广泛应用。

2.3. 拉普拉斯得分算法

没有标签信息指导的情况下，无监督特征选择方法通过引入相关性、数据流形和聚类等技术来筛选有效特征。无监督特征选择算法也可以分为过滤式、封装式和嵌入式三大类。

拉普拉斯得分算法[8]是由 He 等人于 2005 年提出的一种基于拉普拉斯特征映射及局部保留投影的无监督过滤式特征选择算法，实现方便、计算成本较低且效率较高。关键性假设是同一类别的数据样本相互靠近，不同类别的数据样本相距较远。目的是在特征选择过程中保持数据的局部几何结构。其步骤如下所述。

输入：样本资料矩阵 X

$$X=\left(x_{ij}\right),i=1,\cdots,N;j=1,\cdots,m,$$

其中第 i 行表示第 i 个样本，记为 x_i；第 j 列表示第 j 个特征，记为 f_j。

输出：特征的拉普拉斯得分

步骤 1：构建样本的相似度矩阵

若样本 x_i 是样本 x_j 的 p 最近邻，$s_{ij}=\mathrm{e}^{-\frac{\left\|x_i-x_j\right\|^2}{t}},i,j=1,\cdots,N$；否则 $s_{ij}=0$，则样本的相似度矩阵为 $S=\left(s_{ij}\right)_{N*m}$。

步骤 2：构建拉普拉斯矩阵

令对角矩阵 $D=\mathrm{diag}\left(d_{11},d_{22},\cdots,d_{NN}\right)$，则拉普拉斯矩阵 $L=D-S$，其中

$$d_{ii}=\sum_{j=1}^{N}s_{ij},i,j=1,\cdots,N。$$

步骤 3：计算特征 f_j 的拉普拉斯得分

令 $\widetilde{f}_j = f_j - \frac{f_j^{\mathrm{T}} D f_j}{1^{\mathrm{T}} D 1} 1$，$1 = [1,1,\cdots,1]^{\mathrm{T}}$，则特征 f_k 的拉普拉斯得分 $L_j = \frac{\widetilde{f}_j^{\mathrm{T}} L \widetilde{f}_j}{\widetilde{f}_j^{\mathrm{T}} D \widetilde{f}_j}$。

对于特征 f_j，L_j 越小，该特征越重要。

3. 基于 mRMR-XGB-LS 算法的信用评估分类模型

由于在拉普拉斯得分算法中，特征的拉普拉斯得分 L 越小，代表特征越重要。为了方便统一的计算和衡量，采用高斯函数将 L 进行转换，得到新的衡量特征的重要性 sig_L，则有：

$$sig_L = \frac{1}{\sqrt{2\pi}} \exp(-L/2) \tag{7}$$

当 L 越小，sig_L 越大，特征的重要性越大。

本文提出一种基于 mRMR-XGB-LS 算法的信用评估分类模型，在特征选择阶段结合了过滤式泛化性能好、计算开销小和嵌入式算法效率高、性能好，无监督特征选择算法不需要数据标签的优点。详细步骤如下。

输入：无标记训练特征 X^U、有标记训练特征集 X^L 及其标签 Y^L，特征个数 m，选择特征个数 n。

输出：特征子集 X^{sub} 和信用评估分类模型。

步骤 1：计算特征的重要度并进行特征排序

对于数据集的每个特征 f_k，根据 LS 算法计算特征在所有训练集(X^U 和 X^L)上的重要度得到其特征排序为 $rank_L$；根据 mRMR 特征选择算法计算在有标记特征 X^L 上的重要度并得到其特征排序 $rank_{mRMR}$；

利用 XGBoost 特征选择算法得到各个特征重要度并得到其特征排序 $rank_{XGBoost}$。

步骤 2：计算特征的最终权重

根据特征排序 $rank_L$、$rank_{mRMR}$、$rank_{XGBoost}$ 得到特征权重 $weight_L$、$weight_{mRMR}$、$weight_{XGBoost}$，对其进行加和，得到特征总权重 $weight_{all}$，其中

$$weight_j = m + 1 - rank_j, j = 1,\cdots,m \tag{8}$$

步骤 3：选择最优特征子集

对 $weight_{all}$ 进行排序，选择前 $n(n < m)$ 个权重最大的特征子集。

步骤 4：构建信用评估分类模型

基于最优特征子集 X^{sub}，利用 XGBoost、LightGBM 和 CatBoost 构建信用评估分类模型。

信用评估分类模型的流程图如图 1 所示。

4. 实验结果与分析

4.1. 数据集和数据预处理

为了验证本文提出的基于 mRMR-XGB-LS 算法的信用评估分类模型的可行性，选取 UCI 中的 Taiwanese Bankruptcy Prediction Data Set 数据集和 Kaggle 上 Financial Distress Prediction 数据集进行试验研究。前者是预测公司是否破产的数据集，当类标签为 0 时，表示该公司正常；当类标签为 1 时，表示该公司破产。后者是预测样本公司是否陷入财务困境的数据集，当目标变量大于−0.5 时，该公司应被视为健康(0)；否则被视为财务困境(1)。这两个数据集中的其他数据为样本公司的财务特征和非财务特征。将两个数据集中财务困境或者破产的样本记为正样本，否则记为负样本。

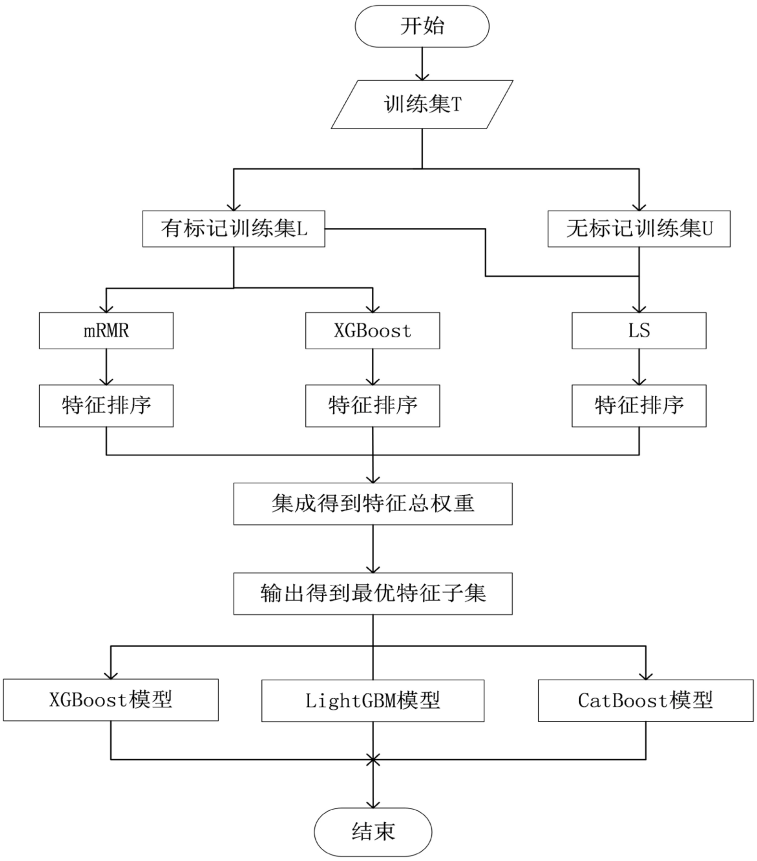

Figure 1. Credit evaluation classification model based on mRMR-XGB-LS algorithm
图 1. 基于 mRMR-XGB-LS 算法的信用评估分类模型

对两个数据集进行数据预处理。首先，删去数据集中重复的样本和特征。其次，进行缺失值处理：删去数据中缺失程度大于 1/3 的特征；针对代表破产的样本，删去超过 10 个缺失值的样本。针对代表未破产的样本，删去所有含有缺失值的样本。接着，对于缺失值，使用每个特征的均值对其填充。最后，使用离差标准化来去除数据的量纲。数据处理后，结果如表 1 所示。

Table 1. Data set distribution
表 1. 数据集分布

数据集	样本量	特征数	类别	正样本
Taiwanese	4618	91	2	4.8%
Financial	3670	83	2	3.7%

从两个数据集的分布可以看出，数据集特征较多，需要进行特征选择以去除冗余、无关的特征。

4.2. 实验设计

首先，将公开数据集按类别分层划分，抽取 20%的样本作为测试集 T_{est}，将 80%的样本作为训练集 T。其次，为了探究不同情况下本文所提出的基于 mRMR-XGB-LS 算法的信用评估分类模型的有效性，在训练集 T 中依照比例分层划分为有标记训练集 L (包括样本特征 X^L 和样本类别 Y^L)和无标记训练集 U (包括

样本特征 X^U 和样本类别 Y^U)，并将 U 中标签删去当作无标记样本。不失一般性，取 L 和 U 的比例分别为 3:2、1:1、1:2、1:3 和 1:4。接着，在 X^L 上训练 mRMR 和 XBoost 特征选择算法，在 X^L 和 X^U 上训练 LS。最后，分类器选择 XGBoost、CatBoost 和 LightGBM。

为了选择不同特征算法的最优参数使其在测试集上达到最高的性能以及充分验证模型的有效性，对不同模型进行五折交叉验证，所有比例下同一模型的参数保持一致，将其评估指标的均值进行对比。对于分类算法的参数选择默认值，不对其进行设定。在本文算法中，LS 中 p 近邻数是较为重要的参数。根据实验得到 LS 中的 p 值和最佳特征个数如下表 2 所示。

Table 2. Parameters of feature selection
表 2. 特征选择参数

数据集	LS 中 p 近邻数	最佳特征个数
Taiwanese	19	40
Financial	13	29

4.3. 分类评估指标

由于本文所选择数据集是高维不平衡的数据集，通常可以构建混肴矩阵，利用正类(少数类)样本召回率 rr_p 、正类(少数类)样本查准率 pr_p 、G-means (G)值和 F-value (F)值来衡量模型性能。具体定义见表 3：

Table 3. Confusion matrix
表 3. 混肴矩阵

	真实正类	真实负类
预测正类	TP	FP
预测负类	FN	TN

$$rr_p = TP/(TP+FN) \tag{9}$$

$$pr_p = TP/(TP+FP) \tag{10}$$

$$rr_n = TN/(FP+TN) \tag{11}$$

$$G = \sqrt{rr_p \times rr_n} \tag{12}$$

$$F = \frac{2rr_P \times pr_p}{rr_P + pr_p} \tag{13}$$

其中，rr_p 衡量了违约样本被正确预测的概率，rr_n 衡量了非违约样本被正确预测的概率。G 值综合评估了模型对两种类别的样本预测正确的性能，G 值越大，代表模型整体的性能越强。F 值综合考虑了违约样本的召回率和精准率，F 值越大，表明模型对违约样本的识别能力越强。此外，ROC 曲线下的面积 AUC 也常用来衡量分类模型的性能。

4.4. 实验结果分析

利用 XGBoost、LightGBM 和 CatBoost 三种算法，分别通过使用原始特征集、改进的 mRMR 特征选择算法、XGBoost 特征选择算法、LS 算法以及 mRMR-XGB-LS 特征选择算法(简记为 m-X-L)构建信用评估分类模型，对比不同模型的性能以验证本文所提出的基于 mRMR-XGB-LS 算法的信用评估分类模型的

有效性，其中 LS 使用的是包含有标记和无标记数据的所有训练集。结果如表 4~6 所示，粗体表示精度最高。

Table 4. Comparison of *G* values of XGBoost, LightGBM and CatBoost models
表 4. XGBoost、LightGBM 和 CatBoost 三种模型的 *G* 值对比

G	Taiwanese Bankruptcy Prediction Data Set						Financial Distress Prediction					
L:*U*	3:2	1:1	1:2	1:3	1:4	均值	3:2	1:1	1:2	1:3	1:4	均值
mRMR + XGBoost	0.618	0.618	0.566	0.617	0.584	0.601	0.459	0.505	**0.498**	0.522	0.485	0.494
LS + XGBoost	0.625	0.635	0.619	0.564	0.565	0.602	0.429	0.388	0.414	0.483	0.423	0.427
XGBoost + XGBoost	0.625	0.644	0.632	0.612	**0.593**	0.621	**0.493**	0.470	0.429	0.522	0.481	0.479
XGBoost	0.616	**0.665**	0.606	**0.636**	0.573	0.619	0.463	0.435	0.421	0.488	0.441	0.450
m-X-L + XGBoost	**0.632**	0.653	**0.633**	0.611	0.588	**0.623**	0.483	**0.533**	0.481	**0.540**	**0.489**	**0.505**
mRMR + CatBoost	0.593	0.576	0.537	0.543	0.492	0.548	0.438	0.418	0.432	0.480	0.402	0.434
LS + CatBoost	0.617	0.598	0.552	0.521	0.487	0.555	0.419	0.398	0.378	0.404	0.360	0.392
XGBoost + CatBoost	0.632	**0.603**	**0.574**	0.547	0.505	0.572	0.435	0.444	0.424	0.419	0.381	0.421
CatBoost	0.628	0.594	**0.574**	0.540	0.479	0.563	0.463	0.435	0.421	0.488	0.441	0.450
m-X-L + CatBoost	**0.644**	0.596	0.551	**0.549**	**0.525**	**0.573**	**0.483**	**0.533**	**0.481**	**0.540**	**0.489**	**0.505**
mRMR + LightGBM	0.591	0.595	0.591	0.605	0.572	0.591	0.444	0.449	0.460	**0.544**	**0.505**	0.481
LS + LightGBM	**0.651**	0.624	0.606	0.588	0.582	0.610	0.413	0.404	0.428	0.395	0.415	0.411
XGBoost + LightGBM	0.645	**0.645**	**0.615**	**0.612**	0.601	**0.623**	0.467	0.456	0.406	0.521	0.497	0.470
LightGBM	0.613	0.620	0.611	0.604	0.586	0.607	0.437	0.444	**0.467**	0.498	0.450	0.459
m-X-L + LightGBM	0.644	0.625	**0.615**	0.611	**0.622**	**0.623**	**0.508**	**0.457**	0.436	0.533	0.475	**0.482**

Table 5. Comparison of *F* values of XGBoost, LightGBM and CatBoost models
表 5. XGBoost、LightGBM 和 CatBoost 三种模型的 *F* 值对比

F	Taiwanese Bankruptcy Prediction Data Set						Financial Distress Prediction					
L:*U*	3:2	1:1	1:2	1:3	1:4	均值	3:2	1:1	1:2	1:3	1:4	均值
mRMR + XGBoost	0.485	0.493	0.432	0.476	0.450	0.467	0.300	0.348	**0.331**	0.357	0.306	0.329
LS + XGBoost	**0.517**	0.525	0.500	0.427	0.447	0.483	0.252	0.212	0.268	0.311	0.248	0.258
XGBoost + XGBoost	0.508	0.534	0.511	0.475	0.468	0.500	**0.320**	0.304	0.256	0.348	0.296	0.305
XGBoost	0.495	**0.560**	0.483	**0.514**	0.452	0.501	0.301	0.272	0.260	0.328	0.270	0.286
m-X-L + XGBoost	0.512	0.541	**0.520**	0.492	**0.475**	**0.508**	0.316	**0.373**	0.301	**0.365**	**0.317**	**0.334**
mRMR + CatBoost	0.455	0.452	0.418	0.412	0.360	0.420	0.278	0.265	0.278	0.333	0.245	0.280
LS + CatBoost	0.514	**0.499**	0.434	0.388	0.357	0.438	0.252	0.227	0.215	0.248	0.201	0.229
XGBoost + CatBoost	0.529	0.494	0.460	0.417	0.384	**0.457**	0.273	0.289	0.266	0.264	0.226	0.263
CatBoost	0.524	0.482	0.455	0.416	0.355	0.447	0.301	0.272	0.259	0.328	0.270	0.286
m-X-L + CatBoost	**0.535**	0.483	**0.545**	**0.425**	**0.406**	**0.457**	**0.316**	**0.373**	**0.301**	**0.365**	**0.317**	**0.334**
mRMR + LightGBM	0.466	0.473	0.467	0.483	0.431	0.464	0.292	**0.300**	0.307	**0.378**	0.325	0.320

Continued

LS + LightGBM	**0.549**	0.522	0.490	0.484	0.469	0.503	0.243	0.240	0.260	0.255	0.248	0.249
XGBoost + LighGBM	0.539	**0.534**	0.500	0.493	0.477	0.509	0.301	0.298	0.247	0.354	**0.331**	0.306
LightGBM	0.513	0.506	0.499	0.495	0.471	0.497	0.280	0.290	**0.318**	0.325	0.280	0.299
m-X-L + LightGBM	0.540	0.519	**0.508**	**0.498**	**0.521**	**0.517**	**0.342**	**0.300**	0.275	0.364	0.305	**0.317**

Table 6. Comparison of AUC values of XGBoost, LightGBM and CatBoost models
表 6. XGBoost、LightGBM 和 CatBoost 三种模型的 AUC 值对比

AUC	Taiwanese Bankruptcy Prediction Data Set						Financial Distress Prediction					
$L:U$	3:2	1:1	1:2	1:3	1:4	均值	3:2	1:1	1:2	1:3	1:4	均值
mRMR + XGBoost	0.931	0.935	0.940	0.935	0.932	0.935	**0.912**	0.905	**0.906**	0.897	0.892	0.902
LS + XGBoost	**0.951**	0.949	0.941	0.933	0.931	0.941	0.881	0.888	0.872	0.877	0.887	0.881
XGBoost + XGBoost	**0.951**	**0.952**	0.946	**0.941**	0.933	0.945	0.903	**0.915**	0.897	0.912	0.901	0.906
XGBoost	0.949	0.947	**0.952**	**0.941**	**0.941**	**0.946**	**0.912**	0.910	**0.906**	**0.914**	**0.913**	**0.911**
m-X-L + XGBoost	0.943	0.943	0.943	0.940	0.933	0.940	0.906	0.901	0.892	0.899	0.901	0.900
mRMR + CatBoost	0.947	0.945	0.948	0.943	0.941	0.945	**0.927**	**0.928**	0.927	**0.931**	**0.927**	**0.928**
LS + CatBoost	0.952	0.950	0.948	0.942	0.939	0.946	0.913	0.910	0.913	0.911	0.912	0.912
XGBoost + CatBoost	**0.956**	**0.958**	**0.953**	0.945	0.941	**0.950**	0.921	0.924	**0.928**	0.928	0.923	0.925
CatBoost	0.955	0.953	0.951	0.946	**0.944**	**0.950**	0.911	0.910	0.906	0.914	0.913	0.911
m-X-L + CatBoost	0.952	0.953	0.951	**0.947**	0.943	0.949	0.906	0.901	0.892	0.899	0.901	0.900
mRMR + LightGBM	0.939	0.939	0.942	0.940	0.938	0.940	0.914	0.915	0.912	0.910	0.900	0.910
LS + LightGBM	**0.953**	0.949	0.943	0.943	0.935	0.944	0.899	0.898	0.889	0.896	0.895	0.895
XGBoost + LightGBM	0.951	**0.957**	0.950	**0.948**	0.940	0.949	0.912	0.915	0.911	0.914	0.908	0.912
LightGBM	**0.953**	0.952	**0.952**	0.946	**0.946**	**0.950**	**0.918**	**0.923**	**0.915**	**0.913**	**0.918**	**0.917**
m-X-L + LightGBM	0.946	0.949	0.949	0.945	0.945	0.947	0.911	0.921	0.909	0.912	0.905	0.912

由表 4 和表 5 可以看出：在大多数情况下，本文所提出的基于 mRMR-XGB-LS 算法的信用评估分类模型的 G 值和 F 值不仅高于 LS、mRMR 和 XGBoost 三种特征选择下的信用评估分类模型，而且高于原始特征数据集。此外，在各个数据集和信用评估分类模型下，本文模型的 G 值和 F 值的均值都是最高，这说明本文所提出的基于 mRMR-XGB-LS 算法的信用评估分类模型稳定性强，可以有效剔除冗余、无关的特征，且能够有效利用无标记样本识别出少数类样本，提升模型总体性能。由表 6 可得，虽然本文算法的 AUC 值排名并不靠前，但总的来说与其他算法相比，差距不大，分别维持在 0.94 和 0.92 左右，这也从侧面表明了本文模型的优越性和稳定性。

5. 结论

信用评估的实际应用中，数据集中存在着大量标签不完整的数据，单一的有监督特征选择方法和无监督特征选择方法都各有其局限性。为此，本文提出一种基于 mRMR-XGB-LS 算法的信用评估分类模型，能够利用半监督数据有效筛选特征构建模型。筛选后的特征子集更加准确，避免了单一特征选择方法无法有效筛选出冗余、无关特征及单一特征选择方法性能的不稳定性。此外，在多个数据集和信用评估分

类模型中，基于 mRMR-XGB-LS 算法的信用评估分类模型的 G 值和 F 值均为最高，显示了本文所提方法的优越性。

今后有待进一步研究的工作包括：

1) 现有的研究大部分是针对分类问题提出的，针对回归问题的半监督特征选择方法是一个有意义的课题。

2) 本文是基于单一模型对数据进行分类的，如何构建有效的集成信用评估分类模型以克服单一模型的不足，提高模型的稳定性和鲁棒性是一个值得研究的课题。

参考文献

[1] Venkatesh, B. and Anuradha, J. (2019) A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, **19**, 3-26. https://doi.org/10.2478/cait-2019-0001

[2] Peralta, D. and Saeys, Y. (2020) Robust Unsupervised Dimensionality Reduction Based on Feature Clustering for Single-Cell Imaging Data. *Applied Soft Computing*, **93**, Article ID: 106421. https://doi.org/10.1016/j.asoc.2020.106421

[3] Benabdeslem, K. and Hindawi, M. (2011) Constrained Laplacian Score for Semi-Supervised Feature Selection. *Machine Learning and Knowledge Discovery in Databases*: *European Conference*, *ECML PKDD*, Athens, 5-9 September 2011, 204-218. https://doi.org/10.1007/978-3-642-23780-5_23

[4] Ang, J.C., Haron, H. and Hamed, H.N.A. (2015) Semi-Supervised SVM-Based Feature Selection for Cancer Classification Using Microarray Gene Expression Data. *Current Approaches in Applied Artificial Intelligence*: 28*th International Conference on Industrial*, *Engineering and Other Applications of Applied Intelligent Systems*, *IEA*/*AIE*, Seoul, 10-12 June 2015, 468-477. https://doi.org/10.1007/978-3-319-19066-2_45

[5] 张东方, 陈海燕, 王建东. 半监督特征选择综述[J]. 计算机应用研究, 2021, 38(2): 321-329.

[6] 马学俊. GSIS 超高维变量选择[J]. 统计与信息论坛, 2015, 30(8): 16-19.

[7] Dash, M. and Liu, H. (1997) Feature Selection for Classification. *Intelligent Data Analysis*, **1**, 131-156. https://doi.org/10.1016/S1088-467X(97)00008-5

[8] He, X., Deng, C. and Niyogi, P. (2005) Laplacian Score for Feature Selection. *Neural Information Processing Systems*, Vancouver, 5-8 December 2005, 507-514.

[9] 胡宽. 融合集成特征选择与 LightGBM 的多联机制冷剂充注量故障诊断策略[D]: [硕士学位论文]. 武汉: 华中科技大学, 2020.

[10] Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V., *et al*. (2017) Ensemble Feature Selection: Homogeneous and Heterogeneous Approaches. *Knowledge-Based Systems*, **118**, 124-139. https://doi.org/10.1016/j.knosys.2016.11.017

[11] 姜丽, 姜淑娟, 于巧. 软件缺陷预测中基于排序集成的特征选择方法[J]. 小型微型计算机系统, 2018, 39(7): 1410-1414.

[12] Peng, H., Long, F. and Ding, C. (2005) Feature Selection Based on Mutual Information Criteria of Max Dependency, Max-Relevance, and Min Redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

[13] 张田华, 罗康洋. 基于集成学习的上市公司高送转预测实证研究[J]. 计算机工程与应用, 2022, 58(10): 255-262.

[14] 罗康洋, 王国强. 基于改进的 MRMR 算法和代价敏感分类的财务预警研究[J]. 统计与信息论坛, 2020, 35(3): 77-85.

[15] Chen, T. and Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785