文章编号:1671-7872(2025)03-0344-08

基于成对样本比较的相对贫困识别特征正交筛选方法

常志朋[1],陈闻鹤[2,3]
(1.安徽工业大学 商学院,安徽 马鞍山 243032;2.安徽师范大学 经济管理学院,安徽 芜湖 241000;3.复杂系统多学科管理与控制安徽普通高校重点实验室(安徽工业大学),安徽 马鞍山 243032)

摘要:为解决相对贫困识别特征筛选问题,提出一种基于成对样本比较的正交筛选方法。采用成对比较方式收集“存在相对贫困”和“不存在相对贫困”两类成对样本集;基于同类样本拉近、非同类样本推远的思想设计特征子集评估函数,且采用正交试验筛选特征。为验证方法有效性,以大别山区356户建档立卡农户和212户非建档立卡农户为样本,随机构建四组成对样本集筛选四组关键特征,采用逻辑回归、决策树、支持向量机、深度神经网络、随机森林、Boosting和朴素贝叶斯7种分类器进行性能测试。结果表明:除决策树分类器外,其余6种分类器在四组关键特征上的识别准确率、灵敏度、特异度和AUC值均超过90%;不同样本集筛选的特征识别性能差异较小,四组关键特征均能达到全特征集的识别效果。本文方法原理简单、操作便捷,适用于缺乏相对贫困划分标准或难以制定相对贫困划分标准的情形,能有效筛选识别特征。
关键词:相对贫困;关键特征;特征筛选;成对比较;正交试验;马氏距离;大别山区;识别
中图分类号:O 212.4 **文献标志码:**A **doi:**10.12415/j.issn.1671-7872.24029

Orthogonal Feature Screening Method for Relative Poverty Identification Based on Pairwise Sample Comparison

CHANG Zhipeng[1], CHEN Wenhe[2,3]
(1. School of Business, Anhui University of Technology, Maanshan 243032, China; 2. School of Economics and Management, Anhui Normal University, Wuhu 241000, China; 3. Key Laboratory of Multidisciplinary Management and Control of Complex Systems of Anhui Higher Education Institutes, Anhui University of Technology, Maanshan 243032, China)

Abstract: To address the issue of feature selection for relative poverty identification, an orthogonal selection method based on pairwise-sample comparison was proposed. Paired sample sets of “relative poverty” and “non-relative poverty” were collected by means of pairwise by means of pairwise comparison. Then, a new feature subset evaluation function was designed based on the idea of pulling similar samples closer and pushing dissimilar samples further apart. Finally, orthogonal experimental design was employed to select features. To validate the effectiveness of the method, 356 registered poor households and 212 non-registered poor households from the Dabie Mountain area were considered as research subjects. Four sets of paired sample sets were randomly constructed to screen four groups of key features,and seven classifiers including logistic regression, decision tree, support vector machine, deep

收稿日期:2024-03-03
基金项目:国家自然科学基金项目(71673001);安徽省高校人文社会科学基金重大项目(SK2021ZD0034);安徽普通高校重点实验室开放基金项目(CS2024—12)
通信作者:常志朋(1978—),男,吉林榆树人,博士,教授,主要研究方向为马田系统理论、模式识别等。
引文格式:常志朋,陈闻鹤.基于成对样本比较的相对贫困识别特征正交筛选方法[J].安徽工业大学学报(自然科学版),2025,42(3):344-351.

neural network, random forest, Boosting, and naive Bayes were tested for performance evaluation. The results indicate that,with the exception of the decision tree, accuracy, sensitivity, specificity, and AUC values exceeding 90% are achieved by the other six classifiers across all four sets of key features. Minimal variation is observed in the identification performance of features selected from different sample sets, and comparable performance to that of the full feature set is attained by all four sets of key features.The proposed method is characterized by its simple principle and operational convenience, making it suitable for scenarios where relative poverty classification standards are lacking or difficult to establish, thereby enabling effective screening of identification features.

Keywords: relative poverty; key features; feature selection; paired comparison; orthogonal experiment; Mahalanobis distance; Dabie mountain area; identification

2020 年我国脱贫攻坚战取得全面胜利, 历史性地解决了绝对贫困问题。但消除绝对贫困并非减贫工作的终点, 要解决发展不平衡不充分问题、缩小城乡区域发展差距、实现共同富裕, 仍需缓解相对贫困, 而缓解相对贫困的首要前提是准确识别相对贫困。当前国内学者对相对贫困识别开展了系列研究, 文献 [1−6] 主张采用单一收入维度方法进行识别, 但王小林等[7] 认为这种单一收入维度方法存在明显的局限性, 与我国 2035 年发展战略目标不相适应, 建议构建包含经济“贫”、社会“困”和生态环境三维度的综合识别体系。理论上识别维度越多、覆盖面越广, 就越能真实反映相对贫困的深度和广度, 从而更精准锁定相对贫困人群。然而现实中受数据采集成本、数据可得性等因素的制约, 构建全覆盖的相对贫困特征识别体系面临诸多困难。因此, 亟需筛选出一组具有代表性、可操作性强、认可度高且判别力突出的关键特征。这种基于关键特征的识别方法不仅可避免单一收入维度对贫困深度和广度把握不足的缺陷, 还可在国家或区域层面建立统一识别特征体系, 确保评估尺度的一致性。同时, 有助于基层开展相对贫困监测识别工作, 为政府决策、科学研究和公众认知等提供持续可靠的基础数据支撑。

然而, 如何科学筛选识别相对贫困的关键特征仍是一个亟待解决的重要问题。当前主流的特征筛选方法如 Decision Tree[8−9]、Random Forest[10−11]、XGBoost[12]、Lasso[13] 和 Logistic Regression[14−17] 等均需要事先收集先验信息较强的类标签样本数据进行筛选。但现实困境在于, 无论是基于单一收入维度识别相对贫困还是多维度识别相对贫困, 学术界对相对贫困的划分标准还未达成共识, 导致难以收集标签样本数据。此外, 这些方法对使用者的统计学和机器学习专业知识要求较高, 在基层部门的实际应用中面临较大障碍。值得注意的是, 相较于直接判定单个农户的贫困状态, 人们往往更容易判断两个农户之间是否存在相对贫困关系。鉴于此, 提出采用“成对比较”样本采集策略: 若两农户之间存在相对贫困则归为一类, 若两农户之间不存在相对贫困则归为另一类, 即将存在明显相对贫困的农户归为一类, 无明显差异的归为另一类。在此基础上, 通过构建基于样本对距离度量的特征评估函数, 实现同类样本聚集、异类样本分离的特征选择目标; 同时引入操作简便的正交试验设计[18−20] 筛选关键特征, 便于在基层使用和推广。这种方法不仅规避了传统方法对标签数据的依赖, 还大幅降低了对使用者专业背景的要求, 为相对贫困识别提供了切实可行的技术路径。

1　基于成对样本比较的特征子集评估函数构建

设 $X=\{X_1,X_2,\cdots,X_n\}$ 为 n 个相对贫困识别特征, $E=\{x_1,x_2,\cdots,x_m\}$ 为 m 户农户样本数据。由于缺乏相对贫困划分标准, 无法将 E 中的农户划分为“相对贫困”和“相对不贫困”两类样本。但是可在 E 中任意选取 p 对农户进行成对比较, 若任意一对农户 x_i 和 x_j 之间不存在相对贫困, 则 x_i 和 x_j 归为相似样本集; 若农户 x_i 和 x_j 之间存在相对贫困, 则 x_i 和 x_j 归为非相似样本集。因此构建如下相似和非相似两类成对样本集:

$$\mathcal{S}=\{(i,j)|x_i\text{和}x_j\text{是相似的，相互之间不存在相对贫困}\}$$

$$\mathcal{D}=\{(i,j)|x_i\text{和}x_j\text{是非相似的，相互之间存在相对贫困}\}$$

从分类判别角度而言, 对于任意特征子集 $K(K\subset X$ 且不为空), 若其能拉近 (Pull) 所有不存在相对贫困样本对之间的距离, 同时能推远 (Push) 所有存在相对贫困样本对之间的距离, 则表明特征子集 K 能有效提高“相对贫困”和“相对不贫困”两类样本的可分离性, 如图 1。

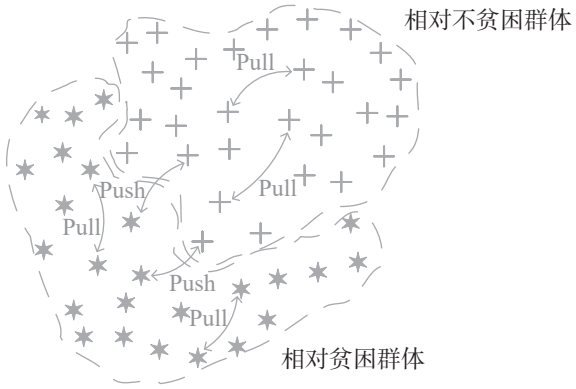

图1 特征子集评估函数构建原理

Fig. 1 Construction principle of evaluation function of feature subset

换言之, 如果任意特征子集K能使所有不相似样本对之间的距离之和越大, 同时使所有相似样本对之间的距离之和越小, 说明特征子集K的区分能力越强, 重要性越高。马氏距离 (Mahalanobis distance, MD) 作为一个协方差距离, 能够充分考虑特征之间的相关性, 更适合用于构建特征子集的重要性评价指标。因此, 本文采用马氏距离构建关于特征子集K的重要程度计算公式, 具体如下:

$$\omega_K = \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^K, x_j^K) - \frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^K, x_j^K) \tag{1}$$

其中: $K \subset X$且K不为空; $d_{\mathrm{MD}}^2(x_i^K, x_j^K) = (x_i^K - x_j^K)^{\mathrm{T}} S_K^{-1}(x_i^K - x_j^K)$; x_i^K和x_j^K为样本x_i和x_j在特征子集K上的取值向量; S_K为p对样本在特征子集K上的协方差矩阵。

2 基于正交试验设计的特征筛选

设对n个备选的相对贫困识别特征进行筛选, 则需计算2^n个特征子集的重要程度, 当n较大时计算难度较高甚至难以实现。正交试验是一种基于数理统计与正交性原理的科学试验设计方法, 其利用规格化的表格——正交表来安排试验, 具有“均衡分散, 整齐可比”的特点, 能够在考察范围内选出代表性强的少数试验条件, 实现均衡抽样。由于正交试验的均衡性, 可通过少量试验找出最优特征组合[21]。因此, 文中采用正交试验计算每个特征的重要程度, 以 6 个相对贫困识别特征为例说明特征筛选过程。

2.1 成对样本集的构建

收集m个相对贫困样本$\mathcal{X} = \{x_1, x_2, \cdots, x_m\}$和$n$个相对不贫困样本$\hat{\mathcal{X}} = \{\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n\}$, 构建不存在相对贫困样本集$\mathcal{S}$和存在相对贫困样本集$\mathcal{D}$, 构建方法如下:

$$\mathcal{S} = \{(x_i, x_j) | (x_i, x_j) \in \mathcal{X}, (x_i, x_j) \in \hat{\mathcal{X}}\},$$

$$\mathcal{D} = \left\{(x_i, x_j) \mid x_i \in \mathcal{X}, x_j \in \hat{\mathcal{X}}\right\}$$

集合$\mathcal{S}$和$\mathcal{D}$中的成对样本数分别为:

$$|\mathcal{S}| = \frac{1}{2}\left(|\mathcal{X}|^2 + |\hat{\mathcal{X}}|^2 - |\mathcal{X}| - |\hat{\mathcal{X}}|\right), \quad |\mathcal{D}| = |\mathcal{X}| \times |\hat{\mathcal{X}}|$$

2.2 正交试验的设计

常用的 2 水平正交表有$L_4(2^3)$, $L_8(2^7)$, $L_{16}(2^{15})$, $L_{20}(2^{19})$, $L_{32}(2^{31})$, $L_{64}(2^{63})$和$L_{256}(2^{127})$等, 针对本例 6 个特征, 选择$L_8(2^7)$正交表。6 个特征可以安排在正交表的任意 6 列中, 文中将 6 个特征安排在正交表的前 6 列中, 具体见表 1。

表 1 2 水平正交试验设计

Tab. 1 Two-level orthogonal test design

试验	相对贫困识别特征							$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i, x_j)$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i, x_j)$	ω_K
	X_1	X_2	X_3	X_4	X_5	X_6				
	1	2	3	4	5	6	7			
1#	1	1	1	1	1	1	1	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_1}, x_j^{K_1})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_1}, x_j^{K_1})$	ω_{K_1}
2#	1	1	1	2	2	2	2	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_2}, x_j^{K_2})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_2}, x_j^{K_2})$	ω_{K_2}
3#	1	2	2	1	1	2	2	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_3}, x_j^{K_3})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_3}, x_j^{K_3})$	ω_{K_3}
4#	1	2	2	2	2	1	1	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_4}, x_j^{K_4})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_4}, x_j^{K_4})$	ω_{K_4}
5#	2	1	2	1	2	1	2	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_5}, x_j^{K_5})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_5}, x_j^{K_5})$	ω_{K_5}
6#	2	1	2	2	1	2	1	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_6}, x_j^{K_6})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_6}, x_j^{K_6})$	ω_{K_6}
7#	2	2	1	1	2	2	1	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_7}, x_j^{K_7})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_7}, x_j^{K_7})$	ω_{K_7}
8#	2	2	1	2	1	1	2	$\sum_{(i,j)\in\mathcal{D}} d_{\mathrm{MD}}^2(x_i^{K_8}, x_j^{K_8})$	$\sum_{(i,j)\in\mathcal{S}} d_{\mathrm{MD}}^2(x_i^{K_8}, x_j^{K_8})$	ω_{K_8}

2.3　特征重要程度的计算

表 1 中，“1”表示该特征参与计算ω_K；“2”表示该特征不参与计算ω_K。因此，表 1 中的每行都能生成特征子集K用于计算ω_K。q次试验的ω_K可表示为：

$$\boldsymbol{\omega} = (\omega_{K_1}, \omega_{K_2}, \cdots, \omega_{K_q})^{\mathrm{T}} \tag{2}$$

若采用$u_{ij}=1$表示特征X_j参与第i次试验，用$u_{ij}=0$表示特征X_j未参与第i次试验，则 6 个特征的 8 次正交试验可表示为：

$$\boldsymbol{U} = (u_{ij})_{8\times 6} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

因此，根据式 (2) 和矩阵$\boldsymbol{U}$可得到各特征参与所有试验的平均ω_K值，计算公式如下：

$$\bar{\omega}^{+} = \frac{2}{q}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{U} \tag{3}$$

进而，可计算各特征未参与所有试验的平均ω_K值，计算公式如 (4)：

$$\bar{\omega}^{-} = \frac{2}{q}\omega^{\mathrm{T}}(\mathbf{1}_{q\times n} - \boldsymbol{U}) \tag{4}$$

最后，根据式 (3), (4) 计算出特征的重要程度向量，如式 (5)：

$$\boldsymbol{\eta} = \bar{\omega}^{+} - \bar{\omega}^{-} \tag{5}$$

$\boldsymbol{\eta}$的每个分量对应一个初始特征，因此第j个分量的η_j即为特征X_j的重要程度。

2.4　关键特征的筛选

根据各特征的重要度 (η_j)，可采用 3 种方法选取关键特征：若$\eta_j>0$则保留特征X_j，若$\eta_j\leqslant 0$则删除特征X_j；指定阈值λ，若$\eta_j\geqslant\lambda$则保留特征X_j，反之删除特征X_j；确定需选取的特征数k，然后根据重要程度分值从$\boldsymbol{\eta}$中选取前k个最大分量，这k个分量对应的特征即为选取的关键特征。

3　实例应用

以大别山区农村家庭为研究对象，参考文献 [22] 的研究框架，从人力资本、社会资本、自然资本、物质资本、金融资本和生计环境 6 个维度初选出 18 个特征，构成大别山区相对贫困识别的初始特征集，具体特征见表 2。

表 2　相对贫困识别特征集

Tab. 2　Identification feature set of relative poverty

维度	特征	特征解释
人力资本	家庭成员健康状况X_1	非常健康=5，健康=4，一般=3，比较健康=2，非常不健康=1
	家庭人均受教育程度X_2	未上过学=0，小学/私塾=6，初中=9，普通高中/职业高中/技校/中专=12，大专=15，大学本科=16，硕士=19，博士=23
	教育支出X_3	家庭教育支出占总消费的比例
	医疗支出X_4	家庭医疗支出占总消费的比例
	职业技能X_5	无劳动能力=0，普通农业劳动力=1，技术型农业劳动力=2，企业普通工人=3，企业技术人员=4，在政府等非企业组织供职或从事个体经营的劳动者=5
社会资本	联系资本X_6	户主月均通信费用
	就业资本X_7	寻找外出务工机会可求助的亲友数量
自然资本	耕地数量X_8	家庭成员人均耕地面积
	粮食产量X_9	家庭所用耕地的亩均粮食产量
物资资本	住房情况X_{10}	家庭成员人均住房面积
	生产资料数量X_{11}	家庭拥有的汽车、摩托车、拖拉机、农业机械等生产型资产数量
	耐用消费品X_{12}	家庭拥有电视、空调、冰箱洗衣机、电脑等耐用消费品数量
	食品支出X_{13}	家庭食品支出占总消费的比例
金融资本	信贷资本X_{14}	从信用社、商业银行等获取贷款的机会
	人均年收入X_{15}	农村住户上年从各个来源得到的总收入相应的扣除所发生的费用后的收入总和除以家庭人口数。
生计环境	自然灾害X_{16}	年均受到滑坡、泥石流、塌方、洪涝等自然灾害或威胁的数量
	基础设施X_{17}	是否通自来水、通电、通公路
	公共服务X_{18}	公共服务可达性，距离最近的卫生站/医院、小学、公交站的平均时间

为验证本文方法的有效性, 根据表 2 中的特征共收集整理 568 户有效农户数据, 其中 356 户农户为建档立卡农户, 将其定义为相对贫困农户; 剩余的 212 户农户为非建档立卡农户, 将其定义为相对不贫困农户。需要说明的是, 本文方法在实际应用中无需收集相对贫困和相对不贫困两类样本, 只需通过成对比较存在相对贫困和不存在相对贫困两类成对样本。运用本文方法筛选相对贫困的关键特征, 具体实施步骤如下。

3.1 成对样本集的构建

为检验采用不同样本对筛选出的特征是否具有明显的识别性能差异, 构建四组“存在相对贫困”和“不存在相对贫困”的成对样本集 (见表 3), 分别利用这四组成对样本集进行特征筛选, 并对筛选出的特征子集进行性能评估。

表 3 成对样本数据集

Tab. 3 Paired sample data set

组号	相对贫困/对	非相对贫困/对
第一组	20	20
第二组	30	30
第三组	30	20
第四组	20	30

3.2 正交试验的设计

针对初选的 18 个特征, 选取$L_{20}(2^{19})$正交表设计正交试验, 表 4 为采用第一组成对样本数据集计算得到的正交试验结果。其余各组样本数据的正交试验结果计算过程与此相同。

表 4 正交试验方案与结果

Tab. 4 Orthogonal experimental design and results

试验	相对贫困识别特征								相对贫困样本对的马氏距离之和	非相对贫困样本对的马氏距离之和	ω_K
	X_1	X_2	X_3	…	X_{16}	X_{17}	X_{18}				
	1	2	3	…	16	17	18	19			
1#	2	2	2	…	2	1	2	1	16.16	11.21	4.95
2#	2	2	2	…	1	2	1	2	11.67	11.06	0.61
3#	2	2	1	…	2	2	1	1	10.25	9.79	0.45
4#	2	2	1	…	2	1	2	2	13.22	6.86	6.37
5#	2	2	1	…	1	2	2	1	12.95	9.95	3.00
6#	2	1	2	…	1	1	2	1	16.78	12.86	3.92
7#	2	1	2	…	2	1	1	2	17.58	12.03	5.55
8#	2	1	2	…	1	1	1	1	12.71	9.88	2.83
9#	2	1	1	…	1	2	2	2	12.63	11.68	0.96
10#	2	1	1	…	2	2	1	2	15.41	10.24	5.17
11#	1	2	2	…	1	2	1	2	18.30	14.00	4.30
12#	1	2	2	…	2	2	2	1	13.13	11.85	1.28
13#	1	2	2	…	1	1	2	2	16.29	12.69	3.60
14#	1	2	1	…	1	1	1	2	17.13	11.96	5.17
15#	1	2	1	…	2	1	1	1	14.05	10.45	3.60
16#	1	1	2	…	2	2	2	2	14.24	12.02	2.22
17#	1	1	2	…	2	2	1	1	12.80	7.92	4.88
18#	1	1	1	…	1	2	2	1	12.22	8.72	3.50
19#	1	1	1	…	2	1	2	2	14.30	10.75	3.55
20#	1	1	1	…	1	1	1	1	27.03	22.71	4.32

3.3 特征重要程度的构建

根据每次试验的ω_K值, 利用式 (5) 计算 18 个特征的相对重要程度并进行降序排序。图 2 为由四组不同样本对计算得到的特征重要性排序结果。

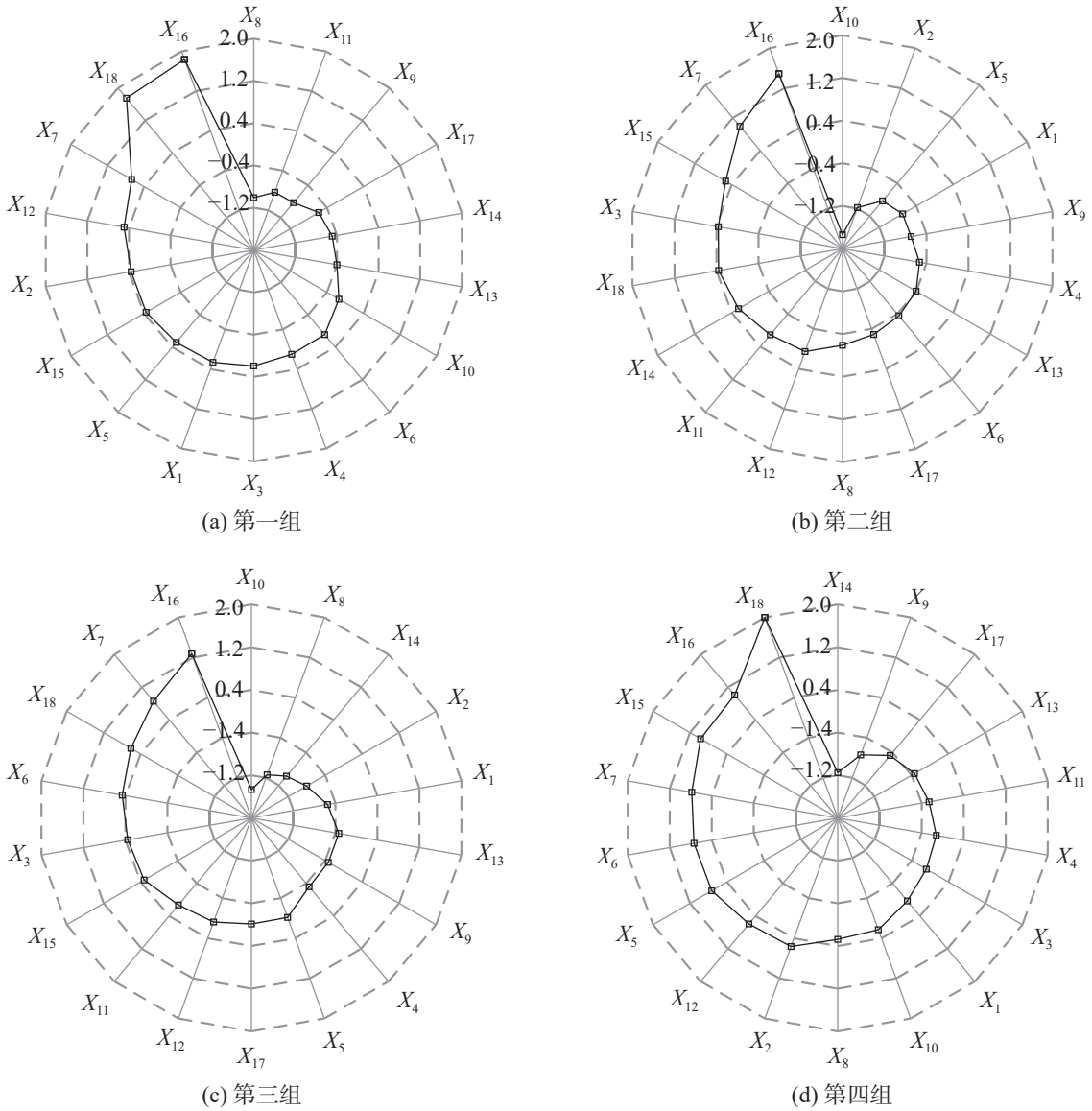

图2　由四组成对样本数据集计算得到的特征重要性排序结果

Fig. 2　Feature importance ranking results calculated from four paired sample datasets

3.4　关键特征的筛选

根据特征重要性排序结果, 从每组特征中选取重要性得分最高的 6 个特征作为相对贫困识别的关键指标, 具体筛选结果如下:

$$K_1=\{X_{16},X_{18},X_7,X_{12},X_2,X_{15}\}$$
$$K_2=\{X_{16},X_7,X_{15},X_3,X_{18},X_{14}\}$$
$$K_3=\{X_{16},X_7,X_{18},X_6,X_3,X_{15}\}$$
$$K_4=\{X_{18},X_{16},X_{15},X_7,X_6,X_5\}$$

选取逻辑回归 (logistic regression, LR)、决策树 (decision tree, DT)、支持向量机 (support vector machine, SVM)、深度神经网络 (deep neural network, DNN)、随机森林 (random forest, RF)、朴素贝叶斯 (naive Bayesian, NB) 和 Boosting 7 种分类器验证 K_1, K_2, K_3 和 K_4 能否有效识别相对贫困, 每种分类器采用 Matlab 自带的训练函数, 具体参数设置见表 5。实验环境为 Intel(R) Core(TM) i9–9900 处理器, Win10 家庭中文版 64 位操作系统, 32 GB 内存。编程环境为 Matlab 2021a 版本。

表 5　分类器训练函数及参数设置

Tab. 5　Training function of classifier and its parameter setting

分类器	训练函数	参数设置
LR	fitglm	系统默认
DT	fitctree	系统默认
SVM	fitcsvm	采用径向基核函数，核参数自动优化
DNN	fitcnet	系统默认
RF	treebagger	集成20棵决策树
NB	fitcnb	系统默认
Boosting	fitensemble	采用RobustBoost算法，集成30棵决策树，误差阈值0.01

选取准确率 (A)、灵敏度 (S_e)、特异度 (S_p) 和 S_{AUC}(area under curve)4 个指标定量评估 K_1, K_2, K_3 和 K_4 的分类性能。其中: A 表示被分对的样本数除以所有的样本数; S_e 表示所有正例中被分对的比例; S_p 表示所有负例中被分对的比例; S_{AUC} 反映分类器在正负两类数据上的识别能力。上述 4 个指标的计算公式如下:

$$A = \frac{T_P + T_N}{T_P + F_N + F_P + T_N} \tag{6}$$

$$S_e = \frac{T_P}{T_P + F_N} \tag{7}$$

$$S_p = \frac{T_N}{T_N + F_P} \tag{8}$$

$$S_{AUC} = \frac{\sum_{i \in \text{positiveclass}} r_i - \frac{M(M+1)}{2}}{MR} \tag{9}$$

式中: F_P 表示假正, 被分类器预测为正的负样本数; T_N 表示真负, 被分类器预测为负的负样本数; T_P 表示真正, 被分类器预测为正的正样本数; F_N 表示假负, 被分类器预测为负的正样本数; r_i 表示第 i 个正样本的序号; M 表示实际类为正类的个数; R 表示实际类为负类的个数[23]。

试验随机抽取 30% 样本为训练集, 剩余 70% 样本为测试集, 每组试验重复运行 50 次后计算 A, S_e, S_p 和 S_{AUC} 的平均值和标准差, 结果如表 6。由表 6 可知: 在每组关键特征中, 7 种分类器中除决策树 (DT) 外, 其余分类器的准确率 (A)、灵敏度 (S_e)、特异度 (S_p) 和 S_{AUC} 均超过 90%, 多数特征超过 95%; 对比四组关键特征发现, 各分类器的分类性能波动幅度很小, 说明不同样本对组合筛选的关键特征具有稳定的分类效果; 与全特征集相比, 四组关键特征的分类性能与之基本相当, 证明筛选出的四组关键特征均能有效代替原始 18 个特征, 适用于相对贫困识别、日常监测及数据收集等工作。

表 6 试验结果

Tab. 6 Experimental results

特征	分类器	A	S_e	S_p	S_{AUC}
	LR	0.96±0.01	0.96±0.01	0.96±0.02	0.99±0.00
	SVM	0.95±0.01	0.95±0.02	0.95±0.02	0.98±0.03
	DT	0.89±0.02	0.92±0.04	0.85±0.05	0.83±0.04
K_1	DNN	0.94±0.01	0.95±0.02	0.94±0.03	0.96±0.03
	RF	0.93±0.01	0.94±0.02	0.92±0.03	0.98±0.01
	NB	0.93±0.02	0.92±0.03	0.95±0.02	0.98±0.01
	Boosting	0.94±0.01	0.94±0.03	0.94±0.02	0.98±0.00
	LR	0.95±0.01	0.95±0.02	0.95±0.02	0.99±0.00
	SVM	0.94±0.01	0.95±0.02	0.93±0.03	0.98±0.01
	DT	0.91±0.02	0.94±0.03	0.87±0.03	0.85±0.03
K_2	DNN	0.93±0.02	0.94±0.03	0.91±0.03	0.93±0.04
	RF	0.94±0.01	0.96±0.02	0.92±0.02	0.97±0.01
	NB	0.92±0.01	0.93±0.03	0.91±0.01	0.98±0.01
	Boosting	0.94±0.01	0.94±0.02	0.93±0.02	0.98±0.01
	LR	0.97±0.01	0.98±0.01	0.96±0.02	0.99±0.00
	SVM	0.97±0.01	0.97±0.02	0.96±0.02	0.98±0.03
	DT	0.92±0.01	0.93±0.02	0.90±0.03	0.88±0.03
K_3	DNN	0.96±0.01	0.98±0.01	0.95±0.02	0.97±0.02
	RF	0.96±0.01	0.96±0.02	0.95±0.02	0.98±0.01
	NB	0.96±0.01	0.96±0.01	0.97±0.01	0.99±0.00
	Boosting	0.95±0.01	0.96±0.01	0.95±0.02	0.99±0.00
	LR	0.96±0.01	0.97±0.02	0.95±0.02	0.99±0.00
	SVM	0.96±0.01	0.97±0.02	0.95±0.02	0.98±0.02
	DT	0.92±0.01	0.93±0.03	0.90±0.04	0.87±0.03
K_4	DNN	0.95±0.01	0.96±0.02	0.93±0.02	0.93±0.03
	RF	0.95±0.01	0.95±0.02	0.95±0.02	0.98±0.02
	NB	0.96±0.01	0.97±0.01	0.95±0.01	0.99±0.00
	Boosting	0.94±0.01	0.95±0.02	0.94±0.02	0.98±0.01
	LR	0.97±0.01	0.97±0.01	0.97±0.01	0.99±0.00
	SVM	0.96±0.01	0.97±0.02	0.96±0.02	0.97±0.03
	DT	0.92±0.01	0.94±0.03	0.90±0.04	0.86±0.03
X	DNN	0.97±0.01	0.97±0.01	0.96±0.02	0.96±0.01
	RF	0.95±0.01	0.95±0.02	0.94±0.02	0.98±0.01
	NB	0.94±0.01	0.94±0.02	0.96±0.02	0.99±0.00
	Boosting	0.96±0.01	0.97±0.02	0.95±0.02	0.99±0.01

4 结论

提出一种基于成对样本比较的筛选相对贫困识别特征方法, 该方法突破了传统方法需要通过明确定义相对贫困划分标准来收集相对贫困与相对不贫困样本的限制, 而是通过构建“存在相对贫困”和“不存在相对贫困”的成对样本集进行特征筛选。这种方法特别适用于缺乏明确相对贫困划分标准或难以制定相对贫困划分标准的场景, 大大提升了特征筛选的适用性。实例验证结果表明该方法是有效可行的, 筛选的多组关键特征均表现出优异的识别性能, 其识别准确率与使用全部特征时基本相当。但需要指出的是, 本文研究在最优样本对的选取策略方面仍存在改进空间, 未来需要进一步优化样本对的构建方法, 以确保筛选出的特征组合具有更强的识别能力。

参考文献:

[1] 陈宗胜, 沈扬扬, 周云波. 中国农村贫困状况的绝对与相对变动: 兼论相对贫困线的设定 [J]. 管理世界, 2013, 29(1): 67−75, 77, 76, 187−188.
CHEN Z S, SHEN Y Y, ZHOU Y B. On the absolute and relative changes in the poverty in China's villages and on the setting of the relative poverty line[J]. Management World, 2013, 29(1): 67−75, 77, 76, 187−188.

[2] 叶兴庆, 殷浩栋. 从消除绝对贫困到缓解相对贫困: 中国减贫历程与 2020 年后的减贫战略 [J]. 改革, 2019(12): 5−15.
YE X Q, YIN H D. From eliminating absolute poverty to alleviating relative poverty: China's history of poverty reduction and poverty reduction strategies after 2020[J]. Reform, 2019(12):5−15.

[3] 孙久文, 夏添. 中国扶贫战略与 2020 年后相对贫困线划定: 基于理论、政策和数据的分析 [J]. 中国农村经济, 2019(10):98−113.
SUN J W, XIA T. China's poverty alleviation strategy and the delineation of the relative poverty line after 2020: an analysis based on theory, policy and empirical data[J]. Chinese Rural Economy, 2019(10):98−113.

[4] 沈扬扬, 李实. 如何确定相对贫困标准?: 兼论“城乡统筹”相对贫困的可行方案 [J]. 华南师范大学学报 (社会科学版), 2020(2):91−101,191.
SHEN Y Y, LI S. How to determine the standards of relative poverty after 2020?: with discussion on the feasibility of “urban-rural coordination” in relative poverty[J]. Journal of South China Normal University (Social Science Edition), 2020(2):91−101,191.

[5] 周力. 相对贫困标准划定的国际经验与启示 [J]. 人民论坛·学术前沿, 2020(14):70−79.
ZHOU L. International experience and insight of relative poverty standard[J]. Frontiers, 2020(14):70−79.

[6] 祝振华, 张红丽, 李洁艳. 城乡相对贫困的动态识别与量化分解: 兼论相对贫困线的设定 [J]. 农业经济与管理, 2023(5):83−94.
ZHU Z H, ZHANG H L, LI J Y. Dynamic identification and quantitative decomposition of relative poverty in urban and rural areas: also on the setting of relative poverty line[J]. Agricultural Economics and Management, 2023(5):83−94.

[7] 王小林, 冯贺霞. 2020 年后中国多维相对贫困标准: 国际经验与政策取向 [J]. 中国农村经济, 2020(3):2−21.
WANG X L, FENG H X. China's multidimensional relative poverty standards in the post-2020 era: international experience and policy orientation[J]. Chinese Rural Economy, 2020(3):2−21.

[8] ZHOU H F, ZHANG J W, ZHOU Y Q, et al. A feature selection algorithm of decision tree based on feature weight[J]. Expert Systems with Applications, 2021, 164:113842.

[9] FAN W, LIU K P, LIU H, et al. Interactive reinforcement learning for feature selection with decision tree in the loop[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2):1624−1636.

[10] SUN X R, CHAI J. Random forest feature selection for partial label learning[J]. Neurocomputing, 2023, 561: 126870.

[11] MAO X J, PENG L H, WANG Z L. Nonparametric feature selection by random forests and deep neural networks[J]. Computational Statistics & Data Analysis, 2022, 170: 107436.

[12] ZHANG B, ZHANG Y, JIANG X C. Feature selection for global tropospheric ozone prediction based on the BOXGBoost−RFE algorithm[J]. Scientific Reports, 2022, 12(1):9244.

[13] COELHO F, COSTA M, VERLEYSEN M, et al. LASSO multi-objective learning algorithm for feature selection[J]. Soft Computing, 2020, 24(17):13209−13217.

[14] WANG J Y, WANG H M, NIE F P, et al. Feature selection with multi-class logistic regression[J]. Neurocomputing, 2023, 543:126268.

[15] WICHITAKSORN N, KANG Y Y, ZHANG F Q. Random feature selection using random subspace logistic regression[J]. Expert Systems with Applications, 2023, 217:119535.

[16] SATO T, TAKANO Y, MIYASHIRO R, et al. Feature subset selection for logistic regression via mixed integer optimization[J]. Computational Optimization and Applications, 2016, 64(3):865−880.

[17] LI L Y, LIU Z P. A connected network-regularized logistic regression model for feature selection[J]. Applied Intelligence, 2022, 52(10):11672−11702.

[18] WU C F J, HAMADA M S, BOOKS X I. Experiments: Planning, Analysis, and Optimization, Second Edition[M]. Hoboken, N J: Wiley, 2013.

[19] TAGUCHI G, JUGULUM R. New trends in multivariate diagnosis[J]. The Indian Journal of Statistics, 2000, 62B(2):233−248.

[20] TAGUCHI G, JUGULUM R. The Mahalanobis−Taguchi Strategy[M]. New York: John Wiley & Sons, 2002.

[21] 何为, 薛卫东, 唐斌. 优化试验设计方法及数据分析 [M]. 北京: 化学工业出版社, 2012.
HE W, XUE W D, TANG B. Optimized Experimental Design Methods and Data Analysis[M]. Beijing: Chemical Industry Press, 2012.

[22] 何仁伟, 李光勤, 刘运伟, 等. 基于可持续生计的精准扶贫分析方法及应用研究: 以四川凉山彝族自治州为例 [J]. 地理科学进展, 2017, 36(2):182−192.
HE R W, LI G Q, LIU Y W, et al. Theoretical analysis and case study on targeted poverty alleviation based on sustainable livelihoods framework: a case study of Liangshan Yi Autonomous Prefecture, Sichuan Province[J]. Progress in Geography, 2017, 36(2):182−192.

[23] 崔少泽, 赵森尧, 王延章. 基于 ADASYN−IFA−Stacking 的再入院患者风险预测方法 [J]. 系统工程理论与实践, 2021, 41(3):744−758.
CUI S Z, ZHAO S Y, WANG Y Z. Risk prediction method for readmission patients based on ADASYN−IFA−Stacking[J]. Systems Engineering-Theory & Practice, 2021, 41(3):744−758.

责任编辑: 丁吉海