基于随机森林算法的贫困精准识别模型研究

罗　丽

(西北工业大学 管理学院,陕西 西安 710129)

摘　要　扶贫对象的精准识别是实现精准扶贫的重要条件。实现贫困数据的精准分类与识别以及贫困识别由定性到定量、由单维瞄准向多维瞄准的转变是精准扶贫的重要基础。精准识别可以采用大数据分析中的分类算法实现。本文基于可持续生计分析框架,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六个方面建立了多维贫困指标体系,运用随机森林算法构建了精准识别模型,并采用中国家庭追踪调查数据(CFPS),对扶贫对象精准识别模型的分类及识别效果进行了评价,结果表明模型效果良好。

关键词　可持续生计;多维贫困指标;分类;随机森林算法;精准识别

中图分类号:D 422.6;F 323.8　　**文献标识码**:A　　**文章编号**:1008-3456(2019)06-0021-09

DOI 编码:10.13300/j.cnki.hnwkxb.2019.06.004

随着我国扶贫脱贫工作进入关键决胜阶段,扶贫工作的精细化、精准化要求越来越高。精准识别作为我国实现精准扶贫的首要条件,主要基于人口的生存情况,通过调查分析将真正贫困的人口定位、识别出来。现阶段,我国主要通过对扶贫地区外部环境的考察以及统计收集扶贫地区各类贫困信息,而后由扶贫主体基于贫困数据对贫困人口进行识别[1]。然而,收集到的人口贫困信息是海量的、复杂的、多维度的,根据这些信息判断贫困户并分析贫困程度的难度很大,且费时费力。并且贫困人口的识别是基于扶贫主体的主观意识决定的,缺乏客观性与准确性。

现阶段,全国性的建档立卡数据库仅实现了数据录入,扶贫工作人员只能查询统计贫困相关数据,缺乏对贫困数据的深入分析,很多隐藏的高价值贫困信息未能得到发掘与利用。由于贫困人口建档立卡工作只是实现了贫困数据的采集与初步分析利用,并未为我国扶贫工作的开展提供坚实的数据支撑[2],所以经常会出现一些贫困人口识别脱靶现象。2015年,我国审计署在对广西马山县进行审计时,发现该县违规鉴定的贫困人口多达3 000多名[3]。在2015年8月到2016年6月近一年时间内的建档立卡"回头看"工作中,发现全国扶贫系统中有900多万人是不符合贫困人口标准的,而多达800万人符合贫困人口标准却没有被真正识别出来。2017年,在全国性的精准识别自查自纠工作中,国务院扶贫办在全国范围内发现将近250万贫困人口未被正确识别[4]。这些现象都反映了我国当前的精准识别工作存在着识别精确率不高的问题。

如果能采用数据分析技术对贫困人口数据进行系统深入的分析,挖掘出隐藏在数据背后的贫困特征和规律,将有助于准确识别贫困人口并发现扶贫对象的致贫原因,从而为精准扶贫提供数据依据。贫困的精准识别即从大量的贫困数据中,将人口分成贫困和非贫困两类。通过对数据进行分析训练,进而实现数据的分类预测是大数据领域的一个重要研究方向。通过运用分类算法对样本数据进行训练,就能分析出数据中隐含的特征与规律,并据此实现对未知数据的分类与识别[5]。因此,贫困的精准识别可以采用大数据领域的分类算法进行研究,完成贫困数据的精准分类与识别,实现贫困

收稿日期:2019-05-27

基金项目:国家社会科学基金项目"新时代我国西部地区军民融合创新生态系统演化机理及发展战略研究"(18BGL033);陕西省自然科学基础研究计划项目"商业模式创新视域下陕西省军民融合发展模式研究"(2018JM7006)。

作者简介:罗　丽(1984-),女,博士研究生;研究方向:大数据管理。

识别由定性到定量的、由单维瞄准向多维瞄准的转变,使贫困深度看得见、摸得着,进而减少了人为因素的影响,确保了扶贫对象的精准定位。

一、研究综述

1.**国外研究现状**

扶贫是一个世界性的难题,不同国家的众多学者对精准识别进行了研究。近年来,随着大数据技术的不断发展,许多学者开始尝试将大数据技术运用到实际的扶贫工作中,并取得了一系列的研究成果。

在国外,大数据技术在贫困精准识别方面的运用已经取得了一定的研究成果。Karlan 等运用大数据技术,构建了"贫困排序"和"家庭情况验证调查"两步法的贫困精确瞄准方法,并通过对秘鲁和洪都拉斯的实证研究证明了该方法的有效性[6]。Sano 等针对印度尼西亚的贫困情况,运用 K 均值聚类分析算法,对各省的绝对贫困、各省贫困人口的相对数、贫困率、各省贫困深度指数等数据进行了统计分析,并以集群成员可视化的形式呈现了聚类分析的结果[7]。Permana 等通过对 351 个村庄进行观测与生计资产分析,将影响村庄贫困的因素分为教育、卫生和经济 3 大类 11 个指标,运用 C4.5 算法对数据进行分类和处理后,发现最敏感的贫困指标为辍学率(教育方面)、营养不良率(健康方面)和农民工率(经济方面)[8]。Vijaya 等针对贫困测度主要从家庭维度出发而缺少对个人贫困程度现状的分析,认为个体的性别差异也是影响家庭贫困情况的一个重要因素,并构建了个体多维贫困测度模型,对印度卡纳塔克邦的个体贫困情况进行了测度,分析得出个体多维贫困测度能更好地体现出因性别导致的贫困差异[9]。

2.**国内研究现状**

在我国,大数据技术在贫困精准识别方面的运用还主要停留在理论分析的层面上。邓维杰获取了多个贫困村的生计资本数据,据此分析了贫困村各级指标间的关系及影响机理,并提出了二元检索贫困村分类法[10]。王瑜使用 K 均值聚类方法对我国农村贫困地区的贫困人口进行聚类分析,将贫困人口分为特色地区贫困人口和连片贫困人口,并对这两种贫困人口的结构进行了深入分析,由此得到其各自的特点和区域分布[11]。田宇利用 Kriging 算法对数据进行了空间插值,构建了基于"单维度、多维度识别及贫困加总/分解"的多维贫困测度算法,对武陵山贫困地区的贫困程度进行了测算与分析[12]。张传华探讨了大数据下的扶贫管理机制并分析大数据精准扶贫的实施对策以及未来所面对的挑战,指出利用大数据技术对贫困数据进行分析,能够进行科学的预测,实现帮扶对象的精准定位[13]。邓华丽采用 K 均值算法,从家庭人口、平均年龄、儿童和老人的比例、病人和残疾人的比例、平均工作能力、平均教育水平六个方面分析了符合最低生活保障条件的中低收入群体的关键特征,并根据聚类结果将生活保障体系中的家庭分为患者家庭、孤独长者家庭、学生家庭和贫困家庭四类[14]。

综上所述,国外许多研究人员在贫困人口大数据分析方面已经取得了不少研究成果,主要集中在贫困数据的聚类分析和贫困影响因子挖掘方面,而基于分类算法的贫困人口识别研究较少。另外,由于其他国家的贫困人口识别指标体系和机制与我国的有所不同,因此,对我国的贫困人口精准识别参考意义有限。而我国有关大数据在贫困精准识别方面的研究大多集中在可行性方面,主要以定性的角度为主,缺乏定量的研究。现有的贫困人口识别定量研究也主要是基于某个地区的贫困情况而进行的,对其他贫困地区的贫困人口识别缺乏适用性。

因此,从我国当前的贫困人口现状和贫困人口识别机制出发,构建基于可持续生计的多维贫困指标体系,运用随机森林算法构建了贫困人口精准识别模型,以期实现贫困人口识别由定性到定量、由单维瞄准向多维瞄准的转变。

二、贫困测度指标的选择

我国当前的贫困线,即贫困测度指标主要是农村纯收入水平。这种以家庭收入为衡量标准的贫

困测度方法虽然操作方便,但不能从多个角度真实地反映家庭的生活状况。贫困不仅仅表现为生活所需的物质资源匮乏,还表现为个体、家庭或者群体缺乏获取人类发展所需的医疗卫生条件、健康、自由、尊严、社会地位和福利等的能力。因此,对贫困的理解应从多维层面展开。

目前,我国关于多维贫困测度方面的研究还处于起步阶段,还没有建立符合我国实际的多维贫困测度方法。虽然各个省份基于实际情况也建立了精准扶贫多维识别标准,但整体上仍需完善,并不能正确地反映居民的贫困情况。国际上较为成熟的多维贫困测度方法包括"牛津贫困与人类发展中心"的 Alkire 和 Foster 两位学者提出的双临界值法和联合国开发计划署所采纳的多维贫困指数 MPI。我国有关多维贫困的测度研究也主要运用了 MPI 指数和 AF 方法。但是,由于我国存在地域广阔、自然环境复杂、区域特性明显等特征,所以国外的 MPI 指数和 AF 方法在我国的应用效果一般。此外,现阶段我国采用的扶贫方法主要是社会救助,基于社会救助的反贫困战略只能暂时性地帮助贫困户走出困境,并不能从根本上解决其生计问题。因此,为从根本上解决居民的贫困问题,需要从根本上解决居民的可持续生计问题。

运用可持续生计分析框架的基本思路,可以把农户贫困放在开放的系统中做动态考察。在脆弱性的生存环境下,贫困人口拥有的生计资本非常有限,进而导致了其生计的逐步下降并陷入贫困状态;如果政府能够及时针对贫困人口的生计问题展开救助,采取政策支持、资金援助和技能培训等多样化的生计策略,通过经济发展构建群体性的可持续生计,逐步改善贫困人口的生计,进而使得贫困人口生计进入良性循环,逐步从贫困的生活状态中走出来。

本文基于可持续生计分析框架,在李小云等[15]和何仁伟[16]构建的生计资本指标体系的基础上,建立了包括 5 大生计资本和生计环境的多维贫困识别指标体系,如表 1。

表 1　基于可持续生计的多维贫困识别指标体系

项目	指标
人力资本	劳动能力
	教育文化
	职业技能
社会资本	政治资本
	联系成本
	就业资本
自然资本	人均耕地面积
	粮食单产
物资资本	住房情况
	卫生设备
	拥有财产
金融资本	现金资本
	信贷资本
生计环境	自然灾害情况
	户外道路情况
	公共服务状况

人类的生计主要受内外两方面因素的影响,包括人力资本、社会资本、自然资本、物质资本和金融资本在内的生计资本是贫困人口生计状况的内因,生计环境是贫困人口生计状况的外因。生计资本是影响贫困人口贫困的主要因素,也是影响贫困人口脱贫的关键因素。其中,人力资本包括个人的教育水平、职业技能、劳动力和健康状况等;社会资本指个体在社会环境中的亲朋好友、上下级同事、宗教组织等各种社会资源;自然资本是个体所生存的自然环境中的地质资源、植被和生物等自然资源;物质资本是指个人在日常生活中所拥有的基础设施以及生产用具;金融资本是指个人拥有的储蓄资金和获得信贷机会等。生计环境是影响人口贫困程度和脱贫进程的外部因素,包括自然灾害情况、基础设施状况和公共服务状况三个指标。其中,基础设施和公共服务状况良好能够加快脱贫的进程,而严重的自然灾害能够延缓脱贫进程。

三、数据来源与处理

中国家庭追踪调查数据(Chinese family panel studies,CFPS)是由北京大学中国社会科学调查中心主持的一项调查项目,主要是采用调查问卷的方式获取我国个体、家庭、社区三个层次的数据。本文主要采用 CFPS 2016 年的调查数据,并按照构建的可持续生计多维贫困指标体系筛选出无缺失值的农村面板数据;其中家庭数 7 860 户,个体样本数 25 382 个。

本文主要采用从中国家庭追踪调查信息系统提取的方式获取贫困数据,然后根据定义的可持续

生计多维贫困指标体系完成各项数据由符号数据转化为计量数据,并剔除一些不规范的贫困数据,最后将贫困数据按照统一的格式进行存放,完成贫困数据的转化、清洗与存储。

1.**符号数据向数值数据的转换**

为了便于数据统计分析,需要将调查数据转化为计量数据。本文采用对照表的形式,将贫困指标的符号数据根据指标定义一一对应转化为可以直接分析利用的贫困数据。根据贫困分析指标和具体贫困数据,将贫困指标的具体数值定义如表2所示。

表2 贫困指标的具体数值定义

指标	含义	0	1	2	3	4	5
劳动能力	按身体健康状况划分	丧失劳动能力	很差	较差	一般	良	优
教育文化	按学历划分	文盲(不会讲汉语)	文盲(能讲汉语)	小学	初中	高中或中专	大专以上
职业技能	根据从事职业划分	无劳动能力	普通农业劳动力	技术型农业劳动力	企业普通工人	企业技术员	具有职业技能鉴定
政治资本	家庭成员是否有乡村干部	有	没有	—	—	—	—
联系成本	家庭月人均通信消费/元	0	(0,30]	(30,50]	(50,100]	(100,+∞)	
就业资本	查找外出打工机会能求助的亲友人数	0	1～3人	4～6人	7～10人	10人以上	—
人均耕地面积	正在经营的耕地面积总和与家庭人数比/亩	[0,0.2]	(0.2,0.5]	(0.5,0.8]	(0.8,1.3]	(1.3,+∞)	—
粮食单产	亩均耕地粮食产量/千克	[0,100]	(100,200]	(200,300]	(300,500]	(500,800]	(800,+∞)
住房情况	人均住房面积/平方米	0	(0,20]	(20,30]	(30,40]	(40,50]	(50,+∞)
卫生设备	根据厕所结构衡量	无	旱厕	水冲式	—	—	—
耐用消费品	拥有手机和电视的情况	都没有	只有一种	两者都有	—	—	—
现金资本	家庭年人均现金收入/元	(0,500]	(500,1000]	(1 000,2 000]	(2 000,3 000]	(3 000,4 000]	(4 000,+∞)
信贷资本	获得信贷的机会	没有	有	—	—	—	—
自然灾害	所处环境有滑坡、泥石流、崩塌等灾害或威胁	没有	有	—	—	—	—
基础设施	是否安装自来水、通电、户外通公路,三项之和	0	1	2	3	—	—
公共服务	是否有卫生站、小学、公交站,三项之和	0	1	2	3	—	—

根据贫困指标的具体数值定义,完成贫困数据由符号数据向数值数据的转换,转化过程如表3。在贫困符号数据向数值数据的转换过程中,可以编程实现从原始数据集中读取贫困指标的数值,并根据贫困指标的具体数值定义依次完成贫困数据的转化。

2.**比例变换**

在采用数据挖掘算法分析贫困数据的过程中,算法要求将贫困数据归一到[0,1]的数值数据,这就需要将相关字段数据进行比例变换。本文所定义的贫困指标均为标量数据,可以通过线性转化的方式,直接映射到一定的数据范围内。这里采用最小-最大规范化数据标准化方法对贫困数据进行处理,如式(1)。

$$X^{*}=\frac{x-min}{max-min} \tag{1}$$

式(1)中,min 代表变量 X 的最小值,max 代表变量 X 的最大值,x 为变量 X 的原始数值,X^{*} 代表转化后的数值。由于各个贫困指标的最小值均为0,因此,各个贫困指标转化后的数据均为其原始数据与其最大值的比值。

表 3　贫困数据向数值数据的转化过程

指标	样本 1	数值数据	样本 2	数值数据	样本 3	数值数据
劳动能力	很差	1	较差	2	一般	3
教育文化	文盲(能讲汉语)	1	初中	3	小学	2
职业技能	企业普通工人	3	技术型农业劳动力	2	普通农业劳动力	1
政治资本	有	1	没有	0	有	1
联系成本	100 元以上	4	50～100 元	3	30～50 元	2
就业资本	7～10 人	3	1～3 人	1	4～6 人	2
人均耕地面积	0.5～0.8 亩	2	0.5～0.8 亩	2	0.8～1.3 亩	3
粮食单产	100～200 千克	1	200～300 千克	2	300～500 千克	3
住房情况	30 m^2 以下	1	40～50 m^2	4	30～40 m^2	3
卫生设备	水冲式	2	水冲式	2	旱厕	1
耐用消费品	只有手机	1	手机电视都有	2	只有电视	1
现金资本	1 000～2 000 元	3	4 000 元以上	5	2 000～3 000 元	3
信贷资本	没有信贷机会	0	有信贷机会	1	—	
自然灾害	没有自然灾害	0	有自然灾害	1	有自然灾害	1
基础设施	已装自来水、已通电、未通公路	2	已装自来水、已通电、已通公路	3	未装自来水、已通电、未通公路	1
公共服务	有卫生站、小学、没有公交站	2	有卫生站、小学、公交站	3	没有卫生站和小学、有公交站	1

对各个指标贫困数据进行归一化处理后,可以将每个人口简化为一个贫困向量,贫困向量的元素坐标即为该样本的基本生活情况,贫困向量定义如式(2)。

$$poverty=(p_1,p_2,p_3,p_4,p_5,p_6,p_7,p_8,p_9,p_{10},p_{11},p_{12},p_{13},p_{14},p_{15},p_{16}) \tag{2}$$

基于贫困向量,所有样本的基本情况就可以通过多个单一贫困向量汇成一个贫困矩阵,进而构建出贫困数据集,即贫困信息模型 P,具体定义如式(3)。

$$P=\begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,16} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,16} \\ \cdots & \cdots & \cdots & \cdots \\ P_{n,1} & P_{n,2} & \cdots & P_{n,16} \end{bmatrix} \tag{3}$$

四、贫困精准识别模型的构建

贫困的精准识别即从大量的贫困数据中,将个体分成贫困和非贫困两类。分类作为数据挖掘的一个重要研究方向,主要是根据一些给定的已知类别标号的样本,通过训练得到分类函数,进而利用分类函数对未知类别的样本进行分类。现阶段,解决分类问题的方法很多,主要包括单一的分类方法和集成学习算法。

随机森林算法作为一种集成学习算法,相比 SVM、神经网络和 KNN 等分类算法,分类效果较优。本文分析的人口数据具有非平衡性、数据缺失异常和数据变量多等特征,而随机森林算法能够很好地应对数据集数据缺失、非平衡及多元共线性问题,在对多元数据进行分类预测时能取得良好的分类效果,是当前分类效果较好的算法之一。因此,本文采用随机森林算法构建贫困人口识别模型实现对贫困样本的识别。

1.基于随机森林算法的贫困识别模型构建

基于随机森林算法的贫困户识别过程的主要步骤就是决策树的生成。随机森林算法在构建决策树的时候需要经过采样与完全分裂两个步骤。在构建决策树之前,随机森林算法会对贫困数据集进行行、列两次随机采样。行采样主要是采用有放回的方式,从原有数据集中随机采取与样本集中样本数量相同的样本,组成新的样本集;这样就保障了随机样本集在不改变原数据集数据分布的同时,解决了贫困数据集过拟合的问题。列采样主要是从贫困数据集的特征集中随机选取多个特征作为决策

树构建的依据。随机采样完成后,随机森林算法采用完全分裂的形式,根据样本数据的特征值构建贫困决策树。

随机森林算法主要是通过随机生成训练集和生成多棵决策树来完成随机森林模型的构建,并通过投票选择分类结果,完成测试样本集的分类,其过程如图 1 所示。

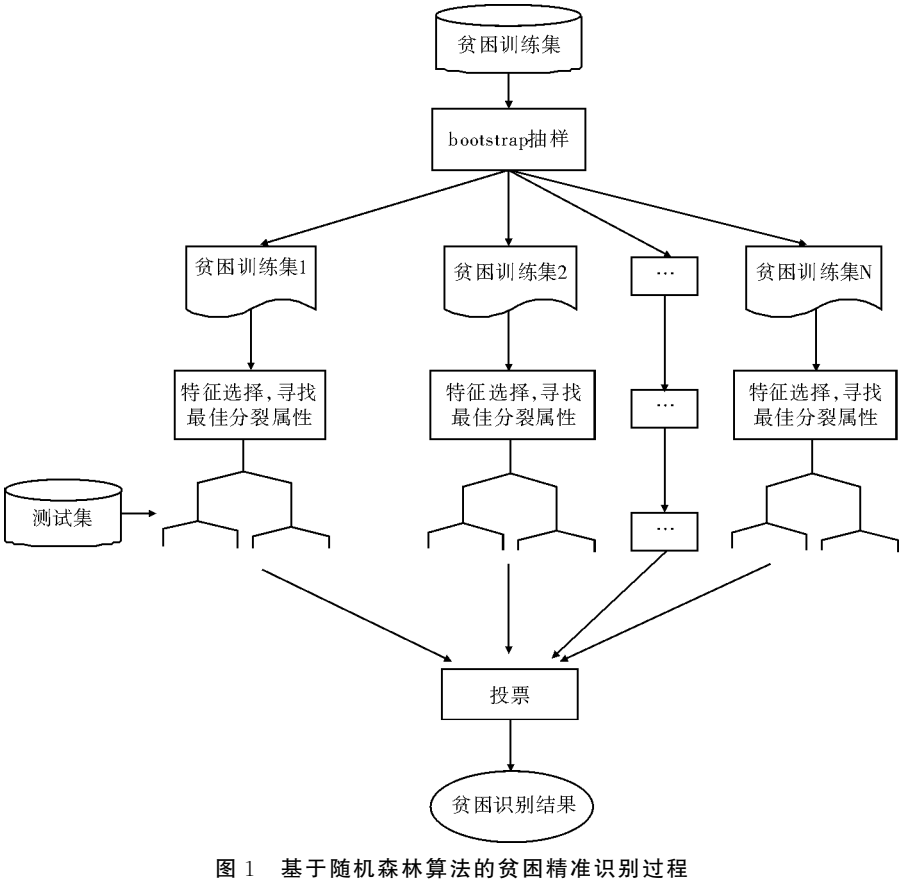

图 1 基于随机森林算法的贫困精准识别过程

具体步骤描述如下:

(1)采用 bootstrap 抽样法从贫困数据集中有放回地随机抽样,构建与原数据集样本数量相同的贫困训练样本集;对于未被抽中的样本,组成贫困测试样本集,用来测量贫困训练集构成的决策树的识别误差。

(2)随机选取多个特征(即贫困指标)作为决策树构建的依据,基于构建的多个训练样本集,构建多个决策树;在选择特征时,一般选择的特征个数远小于样本总共拥有的特征数;假设样本有 M 个特征,一般选择 $m=\log_2 M$ 个属性作为构建决策树的依据。根据 CART 树算法的原理,采用 Gini 系数,从 m 个特征中选取 Gini 系数最小的分类特征作为支节点,并基于分类特征分裂后的 Gini 系数确定最优切分点,完成 CART 树的构建。

贫困精准识别是一个二分类问题,贫困数据集的 Gini 系数计算如式(4)。

$$Gini(P)=\sum_{i=0}^{1}p_i^2=2p(1-p) \tag{4}$$

式(4)中,P 为贫困数据集,由于贫困为二分类问题,当 $i=0$ 时,表示该家庭为非贫困家庭,当 $i=1$ 时,表示该家庭处于贫困状态;p 表示家庭处于贫困状态的概率。

对于每个特征属性 p_i,按照大于其特征值的样本归为数据集 $D1$,不大于其特征值的样本归为数据集 $D2$ 的规则,根据各个特征值,依次将贫困数据集进行分裂,并计算出各个特征值的分裂 $Gini_{pi}(D)$,计算方法如式(5)。

$$Gini_{pi}(D)=\frac{|D1|}{|D|}Gini(D1)+\frac{|D2|}{|D|}Gini(D2) \tag{5}$$

每个特征属性 p_i 根据特征值划分,具有多个 Gini 系数;依次计算出所有特征属性的 Gini 系数后进行对比,选择最小 Gini 系数对应特征属性作为划分节点,并将对应的特征属性的特征值作为最优切分点。

(3)从根节点开始,按照步骤 2,采用贪心策略从上到下依次选择分类特征,直到节点不能分裂为止,完成决策树的构建。

(4)将上述过程重复多次,构建多棵决策树,形成随机森林。

(5)当需要对贫困测试集中的样本进行分类时,通过随机森林得出该样本的多个识别结果,计算样本 x 识别结果为 c 的概率 $P(c|x)$,并运用多数投票法,将概率最大的贫困识别结果确定为该样本的贫困识别结果;贫困识别结果确定方法如式(6)。

$$C=arg\ max\ P(c|x) \tag{6}$$

基于以上分析,本文运用 R 语言的 randomForest 包构建了基于随机森林算法贫困的精准识别算法。在 randomForest 包中,randomForest 函数用来构建随机森林模型,Predict 函数使用训练后的随机森林对新数据进行预测。randomForest 函数有预测属性、生成决策树的数目和分裂属性的个数等三个参数,并能根据三个参数通过分析训练集生成随机森林模型;设置样本是否贫困为需要预测的属性。然后,使用 Predict 函数依据生成的随机森林模型对测试集进行预测。

2.模型参数确定

在基于随机森林的贫困精准识别模型中,ntree 和 mtry 是两个重要的属性,对贫困精准识别的性能有重要影响。因此,确定合适的 ntree 和 mtry 值对模型的构建非常重要。基于上述算法,对不同 ntree 和 mtry 值下算法的性能进行了对比分析,并最终确定了最合适的值。

首先,采用十折交叉验证法,将 7 860 个样本数据构成的贫困数据集划分为 D1-D10 十个贫困数据子集,其中每个数据集中有 786 个样本数据;其次,采用其中 9 组贫困数据子集作为训练集构建贫困精准识别模型,剩余 1 组作为测试数据集,运用混淆矩阵方法测试所构建的贫困精准识别模型的精准度和误差率;最后,重复 10 次实验,求得平均精准度和误差率作为贫困精准识别模型的精准度和误差率。

混淆矩阵具体定义如表 4。

根据混淆矩阵可以计算出算法的正确率、误差率、精确率、召回率和 f 值。

$$正确率=(TP+TN)/(TP+FP+FN+TN)$$

$$误差率=(FN+FP)/(TP+FP+FN+TN)$$

$$精确率(P)=TP/(TP+FP)$$

$$召回率(R)=TP/(TP+FN)$$

$$f\ 值=2PR/(P+R)$$

表 4　混淆矩阵

实际结果	预测结果	
	正类	负类
实际正类	TP	FN
实际负类	FP	TN

表 5　误差对比分析

mtry	误差率
1	0.147 1
2	0.146 3
3	0.146 6
4	0.136 5
5	0.142 4
6	0.144 5
7	0.148 7
8	0.154 3
9	0.154 6
10	0.157 7

对于决策树分裂属性的个数 mtry,一般取值为变量数的平方根。为了提高算法的效率,采用不同 mtry 值建立了多个基于随机森林的贫困精准识别模型,并对这些模型进行了误差对比分析,具体如表 5。

通过对比可以发现,当决策树分裂属性的个数 mtry 为 4 时,贫困精准识别模型的误差率最低。因此,将决策树分裂属性个数 mtry 设置为 4。

确定了决策树分裂属性个数 mtry 后,可以通过对比不同决策树个数 ntree 组成的贫困精准识别模型的性能,确定组成贫困精准识别模型的最佳决策树个数。将样本的一半作为训练集,用另一半样本作为测试集,通过对训练集进行训练构建随机森林模型;构建完成后,用训练集和测试集对贫困人口进行识别,基于识别结果对构建的随机森林模型的误差率进行计算。不同决策树个数 ntree 组成的随机森林模型的误差率如图 2。

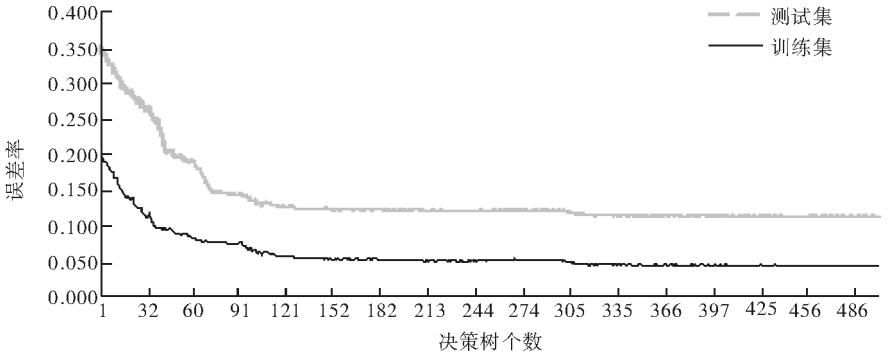

图 2 随机森林模型的误差率

根据图 2 可知,本文构建的随机森林贫困识别模型的识别误差并没有随着决策树数量的增加而出现剧烈的波动。当决策树个数 ntree 小于 100 时,随机森林模型识别贫困人口的误差率随着决策树个数的增加急剧下降;当决策树个数 ntree 大于 100 小于 300 时,随机森林模型识别贫困人口的误差率的下降趋势随着决策树个数的增加趋向缓和;当决策树个数 ntree 大于 300 时,随机森林模型识别贫困人口的误差率基本稳定,不再下降。因此,当决策树个数 ntree 大于 300 时,个数的增加并不能带来误差率的下降,反而降低了模型的运行效率。基于以上情况,将决策树个数 ntree 设置为 300。同时,当 ntree 为 300 时,训练集的误差率为 0.05,测试集的误差率为 0.12;相比测试集,训练集的误差率没有出现过度的波动。因此,本文构建的随机森林贫困识别模型整体比较稳定,没有出现过拟合的情况。

基于以上分析和 R 语言函数,通过对数据样本进行训练构建了随机森林贫困精准识别模型,其中模型中的决策树个数设置为 300,决策树分裂属性的个数 mtry 设置为 4。

贫困精准识别模型的参数确定后,再次采用十折交叉验证法和混淆矩阵法对贫困精准识别模型进行了评价。基于随机森林贫困精准识别模型的贫困家庭识别能力评价主要采用精确率指标;精确率越高,说明已有模型对贫困的识别越精确;评价数据如表 6。

表 6 基于随机森林的贫困精准识别模型精确率分析

序号	精确率/%
1	94.12
2	94.27
3	94.18
4	95.26
5	95.43
6	96.25
7	95.91
8	95.10
9	94.82
10	94.91
平均值	95.02

经分析,本文构建的基于随机森林的贫困精准识别模型的识别精确率为 95.02%;相比李佳容在《随机森林在甘肃省农村贫困户识别中的应用》一文中所构建的贫困识别模型,精确率由 71%提高到了 95.02%。这说明基于随机森林的贫困精准识别模型在贫困识别方面具有较高的精确率,能够从繁杂的贫困数据中精确地识别出贫困人口。

五、模型对比评价

为了对随机森林贫困识别模型的效果进行横向对比,本文采用了神经网络和支持向量机两种典型的分类算法对贫困人口进行了识别,并运用精确率、召回率和 F 值三个评价指标对这三个模型在贫困人口识别方面的效果进行了评价。另外,为了提高评估的准确性,采用十折交叉验证法,将 7 860 个样本数据构成的贫困数据集划分为 D1-D10 十个贫困数据子集,重复 10 次实验,求得平均精确率、召回率和 F 值作为贫困精准识别模型的评价指标。随机森林、人工神经网络和支持向量机三种算法在贫困人口识别方面的精确率、召回率和 F 值如表 7 所示。

表 7 各个模型识别效果评价

算法	评估指标		
	精确率	召回率	F 值
随机森林	0.951 1	0.962 5	0.953 8
支持向量机	0.874 1	0.864 2	0.873 8
人工神经网络	0.891 9	0.891 8	0.893 6

通过分析,基于随机森林的贫困精准识别模型在贫困人口识别方面,相比人工神经网络和支持向量机算法,具有较好的表现,精确率和召回率均在 95%以上,且 F 值较大,表明基于随机森林的贫困精准识别模型具有更好的识别效果。因此,本文构建的基于随机森林算法的贫困人口精准识别模型精确率较高,能够从繁冗的人口数据中精准地识别出贫困人口。

六、结　论

本文基于可持续生计分析框架,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六个方面建立了基于可持续生计的多维贫困指标体系,在构建可持续生计的多维贫困指标体系的基础上,采用北京大学中国社会科学调查中心的中国家庭追踪调查数据(CFPS),通过对数据进行数值转化和比例变换等数据处理,完成了人口数据向贫困向量的转化,建立了贫困信息库。本文通过分析现有的分类算法与精准识别的契合性,基于随机森林算法构建了贫困精准识别模型,利用混淆矩阵对贫困精准识别模型进行了对比验证并得出以下结论:

(1)基于随机森林的贫困精准识别模型,其决策树个数 ntree 设置为 300,决策树分裂属性个数 mtry 设置为 4 时,模型的贫困人口识别效果最好。

(2)相比人工神经网络和支持向量机算法,基于随机森林的贫困精准识别模型在贫困人口识别方面效果更好。

(3)本文构建的基于随机森林的贫困精准识别模型在贫困人口识别方面识别精确率达到了 95.02%,具有较高的精确率,能够从繁杂的贫困数据中精确地识别出贫困人口。

参　考　文　献

[1] 汪三贵,殷浩栋,王瑜.中国扶贫开发的实践、挑战与政策展望[J].华南师范大学学报(社会科学版),2017(4):18-25.

[2] 戈大专,龙花楼,屠爽爽,等.新型城镇化与扶贫开发研究进展与展望[J].经济地理,2016(4):22-28.

[3] 王剑. 夯实重庆精准扶贫、精准脱贫基层基础工作——关于广西马山扶贫事件的教训启示[J].重庆行政(公共论坛),2016(6):34-35.

[4] 朱梦冰,李实.精准扶贫重在精准识别贫困人口——农村低保政策的瞄准效果分析[J].中国社会科学,2017(9):90-112,207.

[5] LIU L L.The research on the new pattern and new approach to Accurate Poverty Alleviation in Henan based on big data analysis [C]//2017 9th international conference on measuring technology and mechatronics automation (ICMTMA). New York:IEEE, 2017:422-426.

[6] KARLAN D,THUYSBAERT B. Targeting ultra-poor households in Honduras and Peru[J].Social science electronic publishing, 2013,6(1):24-28.

[7] SANO A V D,NINDITO H. Application of K-means algorithm for cluster analysis on poverty of provinces in Indonesia[J]. ComTech:computer,mathematics and engineering applications,2016,7(2):141-150.

[8] PERMANA R A,BRATADIREDJA R R,MUNANDAR A. Classification of village poverty's status by C4.5 algorithm as a basis of determining development policy [J]. Bogor agricultural university institut pertanian bogor,2016,10(5):13-17.

[9] VIJAYA R M,LAHOTI R,SWAMINATHAN H. Moving from the household to the individual:multidimensional poverty analysis[J].World development,2013,59(3):70-81.

[10] 邓维杰.贫困村分类与针对性扶贫开发[J].农村经济,2013(5):42-44.

[11] 王瑜,汪三贵.农村贫困人口的聚类与减贫对策分析[J].中国农业大学学报(社会科学版),2015(2):98-109.

[12] 田宇,许建,麻学锋.武陵山片区多维贫困度量及其空间表征[J].经济地理,2017(1):162-169.

[13] 张传华,赵敏.大数据条件下农村精准扶贫机制与路径[J].中国统计,2018(6):14-16.

[14] DENG H,ZHANG L,SU W . Clustering the families successfully applying for minimum living standard security system based on K-means algorithm[C]//2016 12th international conference on computational intelligence & security. New York:IEEE,2016:494-498.

[15] 李琳一,李小云.浅析发展学视角下的农户生计资产[J].农村经济,2007(10):100-104.

[16] 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(2):182-192.

(责任编辑:毛成兴)