基于决策树的农业保险精准扶贫研究＊[1]

——以湖南省 14 地市为例

王 韧 王弘轩

(1 湖南商学院财政金融学院 湖南长沙 410205;

2 厦门大学计算机通信学院 福建厦门 361005)

【摘 要】自 2007 年实施补贴政策以来,我国政策性农业保险获得了井喷式的发展,在其实施过程中逐渐积累了大量的数据。另一方面,我国当前扶贫开发工作已经进入了“啃硬骨头,攻坚拔寨的冲刺期”,农业保险在国家精准扶贫战略中大有可为。在大数据思维下如何充分利用农业保险积累的这些数据来助力精准扶贫,成为当今可能的研究热点。作者研究发现,运用大数据技术中的决策树分类的 ID3 算法对近年来湖南省农业保险的保费补贴范围进行决策分析,通过计算信息增益确定扶贫情况数据中可以明显降低政府保费补贴上下波动较大的因素,据此将决策树算法用于创建保费补贴范围决策的决策树模型,并对其准确性及扩展性进行说明,可较为精准地估测湖南省各地区农业保险保费补贴范围,为农业保险支持精准扶贫提供有力证据。

【关键词】大数据 决策树算法 精准扶贫 农业保险 保费补贴

【中图分类号】F840.66 **【文献标识码】**A**【文章编号】**1003—7470 (2017)—11—0063 (06)

一、引言

随着国民经济的持续增长和农村市场化改革的逐步深入,我国反贫困步伐明显加快,扶贫开发正处在以解决温饱为主的阶段转入巩固温饱成果、改善生态环境、提高发展能力、缩小发展差距的关键时期。提高扶贫开发的精细程度,重视扶贫减贫质量的动态提升,是时代赋予扶贫开发的新使命。2015 年我国中央一号文件明确提出,大力推进农村扶贫开发,推进精准扶贫。众所周知,作为重要的支农惠农手段,政策性农业保险通过风险阻隔、风险补偿及要素管理,可充分发挥为贫困地区“造血扶贫”的功能。据此,2014 年颁布的《加快发展现代保险服务业的若干意见》明确指出,“大力发展三农保险,创新支农惠农方式。”这恰与当前精准扶贫战略方针高度契合。在我国安徽等地,将特色农业保险纳入“减贫扶贫”战略工程已经启动。

另一方面,以 PB (1PB=2^42)为单位的数据”时代的到来,人们生活上、工作上与思维上的变革,信息处理能力也得到了极大的提高,其技术价值在农业保险扶贫这一领域也可得到充分利用。具体而言,基于大数据技术以及算法优化的政府扶贫治理机制,可以精准提高扶贫资源配置的效率,可以切实做到“扶真贫、真扶贫、治本扶贫”,以此提高民生福利。湖南作为农业大省,近年来,在保费补贴政策的推动下,其农业保险发展迅速,势头喜人。然而,道德风险频发、补贴资源不均衡、补贴力度同质化、补贴内容缺乏特色等问题也随之出现,其支农惠农的效果也大打折扣。有鉴于此,我们拟采用基于大数据视野下的决策树算法,构建保费补贴范围决策的决策树模型,以期更精准地确定农业保险保费补贴力度,创新农业保险扶贫方式,更精准确定保费补贴范围,实

[1]＊本文系国家社会科学基金项目“基于精准扶贫视角下特色农业保险补贴机制研究”(编号：15BJY091)的研究成果。

现二者良好对接,以更好服务于农村精准扶贫事业,成为当前急需攻关的课题。

二、文献回顾及评述

农村、反贫困、农业保险补贴政策等一直是政、学、业界共同关注的热点及难点命题。对于如何将农业保险纳入精准扶贫工程,目前研究成果仍然稀少。精准扶贫战略与农业保险补贴的关联研究,对于如何将农业保险纳入精准扶贫工程,尤其是如何构建特色农险补贴体系以更好服务于精准扶贫战略,目前仍处于探索阶段,研究成果稀少。张伟等[1]从财政补贴的公平性角度出发,指出政府应该进一步优化民族地区(贫困地区)农业保险的险种结构,设计灵活的保障水平和保费补贴组合供农民自由选择,并将拨付给民族地区的部分扶贫资金转化为保费补贴的方式发放,以发挥农业保险扶贫的杠杆效应。另一方面,对于决策树算法在农业保险扶贫中的应用,目前尚无研究成果,相关文献仅集中于大数据技术应用于精准扶贫的探索。如汪三贵[2]已提出完善精准识别机制是创新精准扶贫工作机制的重要内容;郑瑞强[3]提出应运用大数据的思维方式纠正扶贫靶向机制,更除原有资源传递内耗大、中间力量的利益阻隔明显等弊病,实施普惠式扶贫向竞争式扶贫战略转变。王茜[4]认为大数据时代,技术的发展为解决我国当前扶贫对象识别不精准,扶贫过程困难等问题提供了技术支持,大数据应用能够揭示传统技术方式难以展现的关联关系,实现基于数据的精准扶贫决策,推动构建精准扶贫工作长效机制。

而在精准扶贫与农业保险关联研究成果方面,由于在农业保险制度相对完善的欧美发达国家并不存在明显的农村贫困问题,所以该研究主要见于发展中国家或落后国家。但国外并无“精准扶贫”这一概念,其成果主要在于强调农险补贴如何与其它扶贫手段配合发挥支农扶贫作用。如 Veermani etal[5]认为农业保险在稳定农户收入,尤其政府灾后偿付成本方面发挥重要作用,但政策性农业保险对农村反贫困的主要手段如农村信贷的作用非常有限。而 Jose agel Villalobos[6]指出,在发展中国家,农业保险应当与其它手段相配合,如农村信贷政策、财政支持的持续性、气象基础设施完善、数据遥感技术的应用等,才能更有效发挥扶贫支农的作用。Aditya K et al[7]的研究以印度为例,尽管农业保险对农户产量及收益无显著影响,但购买农业保险的农户在从事风险性农业生产活动及农村信贷的比例明显高于未投保的农户。同时,考虑到农作物价格存在较大波动,印度政府应考虑发展农产品价格波动损失基金,以对原有的农业保险计划作出必要补充。而就“大数据”而言,尽管这一概念早在上个世纪的 1980 年就已经出现,但关于大数据技术在农业保险精准扶贫方面,国外文献鲜有提及。

我们拟以湖南省 14 地市为例,通过决策树算法,精准保费补贴水平,以最大程度契合精准扶贫之目标,为农业保险更好与精准扶贫机制实现耦合进行初步的尝试及探索。

三、数据来源及模型构建[8][9]

1.变量选取及数据准备我们使用的数据来自湖南省各地区 2008 年~2014 年扶贫情况数据。该资料包含 710 个具体情况样本,共计 11 个字段:填报单位、指标分类、保险标的种类、承保数量、参保农户户次、保险金额、签单保费、已决赔付数量、受益农户户次、已决赔款、政府保费补贴。

(1)填报单位、指标分类与保险标的种类。这三个变量原来就是离散型,其中虽然指标分类只有 2 个属性值,但是其对保费补贴范围的影响依然不能小视,即使在同一年中,由于不同公司承保标的数量不同,同一城市同一标的的保费补贴在不同的指标分类下存在较明显的差别,因此需要将指标分类这一变量加入模型;同样的,在不同城市,同一标的同一指标分类下,保费补贴的数额也差别巨大,例如在 2008 年,同样是人保股份的指标,湖南衡阳在水稻上的保费补贴大约是长沙的 2 倍。不同的保险标的也会使得保费补贴的数量波动较大,所以此三个离散性变量都有必要加入模型中。

(2)承保数量、参保农户户次。假定农业保险补贴对于农户投保行为有正向激励作用,因此对于农户收入偏低、承保数量少、参保农户户次偏低的地区,政府应当加大补贴力度,有针对性地优化补贴效果,以鼓励农户购买农业保险,以更好地保障农户的经济利益。

(3)保险金额、签单保费。我国目前仍以实施保费补贴为主,签单保费与补贴关系的重要性不言而喻。同时,从保险金额而言,目前保成本的保障水平仍然整体偏低,保产量或会成为政府增加补贴、提升保障水平的重要方向。

(4)已决赔付数量、受益农户户次、已决赔款。已决赔付数量、收益农户户次都可以直观的认为出当地农户的受益程度,若已决赔和受益农户过少,除去灾害损失少的年份,可考虑保障程度欠缺,抑或参保人数较少,政府可考虑强化补贴力度对其加以改进。

在决策树分类算法的数据预处理中,所有的连续型数据都需要离散化处理。通过将连续型数据放入根据其大小排序所得的、分布均匀的区间中,可以大大降低连续型数据带来的复杂程度。因此,要保证输入变量在成为算法的直接输入之前,进行离散化化,分区分类,预处理后变量的部分说明见表 1。

表 1　变量预处理及其分类

数据属性	重新编号	变量值	变量类型
填报单位	city	(C1)长沙市，(C2)株洲市，(C3)湘潭市， (C4)衡阳市，(C5)邵阳市，(C6)岳阳市， (C7)常德市，(C8)张家界市，(C9)益阳市， (C10)郴州市，(C11)永州市，(C12)怀化市， (C13)娄底市，(C14)湘西自治州	类别型
农产品	Item	(11)水稻，(12)小麦，(13)大豆， (14)玉米， (15)棉花，(16)油料作物，(17)能繁母猪， (18)奶牛	类别型
签单保费(万元)	attri__D	(D—1) 0～50，(D-2) 50～100,(D—3) 100～200，(D-4) 200—300，(D—5) 300～400, (D-6) 400～700 (D-7) 700～1000,(D-8) 1000～2000，(D-9) 2000～3000,(D-10) 3000～4000，(D-11) 4000～8000	离散型

我们的分析主要关注在划定范围分类的前提下,如何决定保费补贴的合理范围。在 700 多条记录数据中,随机抽取 200 条数据作为训练数据,并将其格式化输入至算法需要的输入文件。

2.熵和信息增益:数据处理

对于以上十一个属性的字段,为了寻找分类的最优办法,需要做的工作就是让生成的分类树的深度尽量小,因此我们用信息增益值的判定方法来判定哪一个属性的数据可以将样本的混乱程度降到最低,这是一个迭代的过程,信息增益的最小值需要不停地计算和更新。

假设S是用以生成决策树的训练样本集合,它包括n个类别的样本,这些类别标记为C1,C2,……Cn,也就是之前说的在离散化中将连续数据按照范围转换成类别标记,例如,上例中,样本根据填报单位就分为了14类,一个地区一类。那么S的熵值或者期望信息就是:

$$\text{Entropy}(S) = -\sum_{i=1}^{n} p_i * \log_2 p_i \qquad (1)$$

假设属性A将样本集合S划分成m份，根据A的属性值划分的子集所得的期望信息由以下公式给出：

$$\text{Entropy}(S, A) = \sum_{i=1}^{m} \frac{|S_i|}{|S|} \text{Entropy}(S_i) \qquad (2)$$

其中Si表示根据属性A 划分的S的第i个子集;|S|和|Si|分别表示数据集合S和Si中的数据样本数目。信息增益用于衡量信息熵值的期望减少值,因此,使用属性A对S进行划分获得的信息增益为:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \text{Entropy}(S, A)$$

Gain (S,A)是指因为知道属性A 的值之后导致的熵的期望压缩。信息增益Gain (S,A)越大,说明选择测试属性A 对分类提供的信息越多,所以信息增益越大的属性可以使得样本在根据该属性分类后的纯度越高。

根据以上原理,下面是首次判断分支属性的流程。由信息熵值和信息增益的公式得出,各个属性的信息增益如表2。

表2　第一次计算信息增益结果

填报单位	指标分类	农产品	承保数量	参保农户户次
0.827584	0.144874	0. 676075	1.15927	0.863114
保险金额	签单保费	已决赔付数量	受益农户户次	已决赔款
1. 61705	2.81549	1.1828	1. 26708	1. 64774

根据以上的信息增益也可以得出排名,最重要的,也就是对当前样本分类效果最明显的就是前担保,所以选择属性“签单保费”作为根节点的测试属性,并对应每一个值(即从D-1到D-11)在根节点乡下创建分支。

四、模型实施与评估

1．模型实施说明

根据上述决策树算法模型的说明，选择“政府保费补贴”字段作为最终结果的分类属性，由于该属性为连续值，将其按照范围离散化，为了尽量精确，将其分类表 3 所见。

表 3 补贴范围分类标记

单位：万元

0～40	40 ～100	100～160	160～250	250-330	330～500
H-1	H-2	H-3	H-4	H-5	H-6
500～600	600～.800	800～1000	1000～1500	1500～2000	2000～2500
H—7	H-8	H-9	H-10	H-11	H-12
2500～3000	3000～3500	3500～4000	4000～4500	4500～5000	5000～5700
H-13	H-14	H-15	H-16	H-17	H-18

并根据此分类得出各个类别下的频率,以此计算总的熵值。建树的部分也是一样,每一次进入算法都要根据当前输入的样本集合重新计算熵值和信息增益。

决策树算法选择 ID3 算法,采用自顶向下的方法递归地构造决策树。

2.决策树构造

我们的研究使用 C++为编程语言实现以上算法,开发环境为 codeblocks,由于传统的算法输入为手动输入,但是此次研究所用数据有 200 条,于是将其输入模式修改为从文件读取。使用 excel 软件对数据进行预处理,并根据各自属性的数据范围进行数据离散化处理,以供算法读取并处理。为方便读取,特将所有属性名称转换为英文属性名并做记录,其表现形式见表 4。

表 4 属性名称转换表

结束标记	填报单位	指标分类	农产品	承保数量	参保农户户次
judge	city	Target	Item	attri__A	attri_B
保险金额	签单保费	已决赔付数量	受益农户户次	已决赔款	政府保费补姑
attri_C	attri_D	attri_E	attri_F	attri_G	attri_H

数据输入后，算法就会对这 200 条数据进行分类统计，记录下每一个属性的每一个属性值，之后对其进行信息熵值，信息增益的计算，这一计算过程是递归的，每一次节点的加入都有一次自己的信息增益计算。算法将根据信息增益的原理进行决策树创建，并将决策树以深度优先的方式输出。

3．模型结果及评价

经过算法的处理后，所输出的决策树的节点数多达 414 个，在此展示从根部开始的部分决策树：

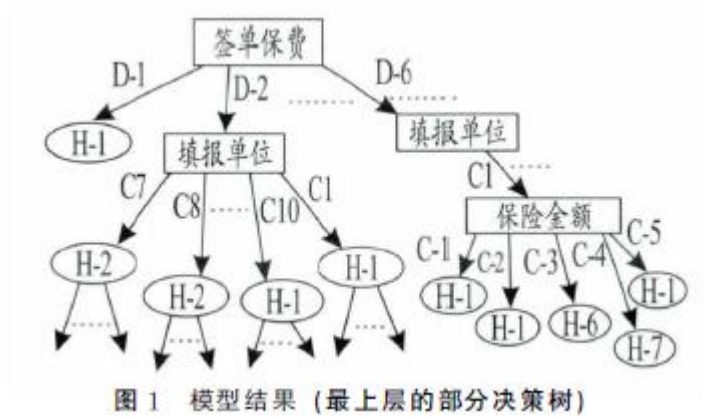

图 1　模型结果（最上层的部分决策树）

从决策树中可知,第一层决策是根据签单保费,这与先前所做的第一次信息增益计算结果对应,可以看出在输入的 11 个属性字段中，“签单保费”对“政府保费补贴”范围决策的影响最大。而第二层出现了“填报单位”，“承保数量”，“已决赔付数量”三个字段,而大部分仍然集中在“填报单位”,而在上述的第一次信息增益计算中,“填报单位”字段的信息增益并不是第二大,这样就验证了之前所说的,信息熵值与增益是要在建树过程中不断更新计算的,而更新计算的依据,就是每一次输入的数据样本集合和剩余属性集合。

对于模型结果的使用以及检验,可以改写算法中的输出路径,输出至文本文件,由于在输出的过程中,层次以“tab”键的数量来隔开,所以可以将输出文件中的内容复制进 excel 电子表格,并根据从左到右的顺序依次进行判断决策,根据以上所做的离散化,处理之后的数据见表 5。

根据 excel 表格中的输出格式,从左往右依次寻找。例如,以上关于郴州市的数据,就找到表格中 attri-D,也就是“签单保费”中的 D-3,根据图 4 就可以从左往右看到下一个考虑的属性就“已决赔付数量”,再找到“已决赔付数量”中的 E-1,就可以看到下一个属性为“农产品”,往下找到“油料作物”的 I6,然后看到下一个属性为“参保农户户次”,其中参保农户户次为 6.31,属

于 B-3,再往右看到“承保数量”,以此类推往右,往下寻找,最终根据属性“已决赔款”为 0 找到属性分类的 G-1,往右找到 H-3,即保费范围的分类决策为H-3,即政府补贴的范围在100万元到160万元之间,与输入的测试数据中政府保费补贴范围的H-3一致。

表 5 离散化后的部分训练数据

填报单位	指标分类	农产品	承保数量	参保农户户次	保险金额	签单保费	已决赔付数量	受益农户户次	已决赔款	政府保费补贴
C1	T4	11	A-6	B—5	C- 5	D-10	E-6	F-7	G— 9	H-14
C2	T4	11	A-6	B-6	C-5	D-11	E-8	F-9	G-10	H-14
C3	T4	11	A—1	B-2	C-3	D-5	E-3	F-2	G—4	H-5
C4	T4	11	A—4	B-4	C-4	D-9	E-7	F-8	G-9	H-11
C5	T4	11	A-6	B-6	C—5	D-11	E-7	F-7	G-10	H-14
C6	T4	11	A-6	B— 6	C- 5	D-11	E-8	F-8	G-10	H-13
C7	T4	11	A—7	B-6	C—5	D-11	E-8	F-8	G-10	H-17
C8	T1	16	A—1	B— 3	C-3	D-3	E-1	F-1	G-1	H-2
C9	T1	16	A-3	B-3	C-4	D-6	E-5	F-6	G-6	H-6
C10	T1	16	A-1	B— 3	C- 3	D-3	E-1	F-1	G-1	H-3
C11	T1	16	A-2	B-3	C-3	D-5	E-3	F-2	G-3	H-5
C12	T1	16	A-3	B-4	C-4	D-6	E-2	F-1	G— 2	H-6
C13	T1	16	A—1	B-2	C-2	D— 2	E-1	F-1	G-1	H-1
C14	T1	16	A-2	B-3	C-3	D— 5	E-2	F-3	G— 2	H-5

表 6 补贴预测结果

长沙市	株洲市	湘潭市	衡阳市	邵阳市	岳阳市	常德市
H-13	H-5	H 一 13	H-18	H-12	H-12	H-17
张家界市	益阳市	郴州市	永州市	怀化市	娄底市	湘西自治州
H-8	H-15	H-13	H-18	H-12	H-12	H—10

为清晰显示预测的保费补贴，将其还原成数据形式的范围见表 7 。

表 7 补贴范围预测的量化结果 单位：万元

长沙市	株洲市	湘潭市	衡阳市	邵阳市	岳阳市	常德市
2500～3000	250～330	2500～3000	5000～5700	2000～2500	2000～2500	4500～5000
张家界市	益阳市	郴州市	永州市	怀化市	娄底市	湘西自治州
600～800	3500～4000	2500～3000	5000～5700	2000～2500	2000～2500	1000～1500

根据以上说明的步骤进行检验,可见通过决策树的判定,保费补贴的范围预测均得出正确结论,说明在大量数据的基础上所得到的决策树有很高的准确性,尽管该决策树预测仍有偏差,但随着新数据的纳入,新的离散化分类也随之加入,使结果更加精准。

五、结论及对策

我们使用 ID3 算法对湖南省各地区的扶贫数据进行分析,并得出一个合理的决策树来判断对一个精确的情况如何来精确地决定补贴范围,从抽取的测试样本中也得到了正确的结果。尽管 ID3 算法的处理方式较为单一,但此次研究所用变量数量较少,所以并不会产生太大误差,检验结果也证实了这一点。

从此次研究可以看出,签单保费对保费补贴的信息增益值远远大于大部分变量,说明在扶贫工作中,最能影响补贴数额的因素,仍属签单保费。因此,对于今后的相关研究,仍应围绕保费和补贴同时展开,首先可以根据大量的农业经营过程所得的数据研究保费的合理决策方式,再进一步研究保费补贴的合理范围。同时,可以据此作为示范,运用 ID3 决策树分类算法进行数据挖掘,从海量贫困农户数据中积极探索数据甄别、数据决策、数据管理、数据考核的精准扶贫方式,可以突出精准性、体现有效性,确保克服传统扶贫中传统的层层上报机制和手工建档立卡方式产生的问题,以最大程度确保贫困对象的精准识别和动态帮扶,为今后的扶贫决策的制定和修正提供更科学的依据。

参考文献:

〔1〕张 伟,罗向明,郭颂平.民族地区农业保险补贴政策评价与补贴模式优化——基于反贫困视角[J].中央财经大学学报,2014,(08).

〔2〕汪三贵,郭子豪.论中国的精准扶贫[J].贵州社会科学,2015,(05).

〔3〕郑瑞强,曹国庆.基于大数据思维的精准扶贫机制研究[J].贵州社会科学,2015,(08).

〔4〕王 茜.大数据时代数字信息资源优化管理和应用模式研究[J].中国管理信息化,2016,(09).

〔5〕Veeramani,Venkat N,Leigh J Maynard,Jerry RSkees.Assessment of the Risk Management Potential of aRainfall Based Insurance Index and Rainfall Options in Andhra Pradesh,India [J].Indian Journal of Economics,2005,(01).

〔6〕Jose, Angel,Villalobos. Agricultural Insurance for Developing Countries: The Role of Governments [J].

Agricultural Outlook Forum,2013,(04).

〔7〕Aditya K.S,Tajuddin Khan,Avinash Kishore.Crop Insurance in India:Drivers and Impact[R].Ricultural& Applied Economics Association Annual Meeting,Boston,MA.2016.

〔8〕王 韧,邹西西,刘司晗.基于 AHP 方法的湖南省农业保险补贴政策扶贫效率评价研究[J].湖南商学院学报,2016,(02).

〔9〕张光建,黄贤英.基于最小聚类单元的聚类算法研究及其在 CRM 中的应用[J].计算机科学,2007,(33).