

所属类别	2020 年“华数杯”全国大学生数学建模竞赛	参赛编号
本科生		CM10669

## 基于多模型对脱贫帮扶绩效综合评价的研究系统

### 摘要

通过 Python 对附件数据进行数据预处理，在不考虑最后 10 个村庄存在缺失值的情况下，本题数据不存在缺失值、重复值与异常值，数据完整度为 100%。

**针对问题一：**首先，从客观角度考虑到：各指标总体的年复合增长率、村庄在不同指标下的排名以及不同村庄的各指标年复合增长率。**然后**，结合主观评价分析，定义了“优秀”与“退步”的标准，以指标为单位对村庄进行分类，计算 2020 年“优秀”村庄的占比。**最后**，对指标进行相关性分析，参照相关系数分析对应关系。**结论：**2015 年与 2020 年对应评分间无直接关联，但五年前好的村庄在 2020 年继续好的概率较大；居民收入与居住环境之间的相关性最强，产业发展、基础设施与文化教育之间的相关性最强。

**针对问题二：**首先，采用 CRITIC 赋权法计算各指标的权重。2015 年 SR、CY、HJ、WJ、SS 指标权重为[0.213, 0.183, 0.230, 0.187, 0.187]，2020 年权重为[0.171, 0.182, 0.220, 0.227, 0.201]。**然后**，计算相应评价指标值与相应指标权重的乘积和，即关于绩效排名的总分。**最后**，参照总分对所有村庄进行排序，得到各类型下脱贫帮扶绩效前十名的帮扶单位编号。按照不同指标，分别统计相同帮扶单位类型下村庄占前 6000 名个数与总数的比值。对比值进行排序，得到不同指标上具有较高绩效的帮扶单位类型。

**针对问题三：**通过 Python 得到帮扶单位编号相同的 160 个村庄集合。**方法一：**作差得到各集合相同指标下 2020 年与 2015 年的差值，计算集合中差值的算数平均数以及各帮扶单位各指标下指标值的中位数。以平均数与中位数相加后平均值作为帮扶单位在特定指标上帮扶业绩的参考值，排序得到各指标排名前五的帮扶单位编号。**方法二：**结合问题二的指标权重，采用万有引力算法分别计算各帮扶单位在各评价指标上的贡献值。修正贡献值后，对不同年份同一指标的贡献值作差，得到对应指标的贡献增加值。参照各指标的贡献增加值对村庄进行排序，对算数平均得到的各集合的平均排名进行排序，得到各指标排名前五的帮扶单位编号。**检验：**计算各集合中前 6000 名个数与总数的比值，对比排序结果与方法一、二的结果，发现误差较小，故选用万有引力算法结果。

**针对问题四：**首先，采用双重聚类协同过滤预测算法对 10 个村庄 2020 年指标值进行填补。**然后**，通过主成分分析法重新计算得到 2015 年与 2020 年各评价指标的权重。参考五年间指标权重的增加值，发现居民收入与居住环境对村庄评选有重要影响。**最后**，筛选各年排名前 10000 名的村庄，剔除掉出前 10000 名的村庄，并匹配新晋村庄 2015 年的总分。计算五年间总分的差值，按差值排名赋予先进称号。数据表最后 10 个村庄内有 3 个一级村庄与 5 个二级村庄。

**针对问题五：**结合数据与所得结论，以信的形式向国家扶贫办阐述在脱贫帮扶工作上的观点并提出建议。

**关键词：**CRITIC 赋权法；万有引力贡献强度算法；双重聚类协同过滤预测算法；PCA+

## 一、问题重述

### 1.1 问题背景

消除贫困、改善民生、逐步实现共同富裕，是社会主义的本质要求，是我们党的重要使命。党的十八大以来，国家把扶贫开发工作纳入“四个全面”战略布局，作为实现第一个百年奋斗目标的重点工作，摆在更加突出的位置。

为了更好的激励各帮扶单位提高扶贫效率，扶真贫，真扶贫。五年前，国家启动了脱贫帮扶绩效评价机制。某科研团队接受任务后，对全国 32165 个需要帮扶的贫困村进行了初步的贫困调查。从居民收入(记为 SR)、产业发展(记为 CY)、居住环境(记为 HJ)、文化教育(记为 WJ)、基础设施(记为 SS)等五个评价指标给出了评分。以此为依据，将被帮扶的村庄划分为 160 个集合，每个集合指定帮扶单位(标记为 0-159)进行帮扶。这 160 个帮扶单位按照单位属性(如国企还是民营企业等)标记为 0-5 等 6 个类型。

2020 年，研究团队再次进行了调研，得到了被帮扶的这些村庄居民收入、产业发展、居住环境、文化教育、基础设施等五个方面的评分数据以及总分数据。为了便于比较和研究，所有数据都进行了标准化处理(标准化后的数值越大表示评分越高)。绩效评价，不能仅以最后的得分作为依据，需要考虑各个评价指标的进步幅度。因为这样才能公正的评判帮扶的效果，就能鼓励更多的帮扶单位愿意花精力去帮助非常贫困的地区。

### 1.2 需求解的问题

**问题一：**一般人的理解是，五年前的评分与 2020 年对应的各项评分有着直接的关联，如五年前的居民收入不错，现在的居民收入也会是不错的。你认为本问题有这个规律吗？请分析各个评价指标的对应关系。

**问题二：**160 个帮扶单位帮扶着基础不同的村庄，帮扶单位帮扶工作的态度、目标、投入、帮扶干部素质等显然是有差异的。仅仅用 2020 年各村庄评分高低显然是无法真正有效的体现一个帮扶单位在脱贫攻坚提升方面所做出的努力。请你运用附件的数据，阐明什么类型的帮扶单位，哪些帮扶单位在脱贫帮扶上面有较高的绩效？请给不同类型的帮扶单位绩效排序，给出脱贫帮扶绩效前十名的帮扶单位编号。

**问题三：**每个帮扶单位在扶贫上有不同的工作特色，如有些单位在提高居民收入上效果很好，而有些帮扶单位可能在改善基础设施上帮助的效果不错。请问，哪些帮扶单位分别在居民收入、产业发展、居住环境、文化教育、基础设施等评价指标上帮扶业绩明显？请列出各单项评价指标前五名的帮扶单位编号。

**问题四：**全国计划给予 10000 个村庄“脱贫先进村庄”称号。请问，哪些因素对获得这个荣誉称号有着非常重要的影响？数据表中最后有 10 个村庄的 2020 年的评价分数被删除，请你判断他们能评上“脱贫先进村庄”称号吗？如果称号分为一级和二级（一、二级称号比例为 1:3），这 10 个村庄中谁能评上“脱贫先进一级村庄”称号？

**问题五：**依据你的研究成果，向国家扶贫办写一封 500 字左右的信，阐述你的观点和建议。

## 二、问题分析

### 2.1 数据预处理

对附件数据进行数据预处理，在不考虑最后 10 个村庄存在缺失值的情况下，查找有无缺失值、重复值与异常值。若有缺失值，则采用相应方法对其进行补缺；若有重复值或异常值，则考虑其产生原因，对其采取删除或替换的方法。

## 2.2 问题一分析

**步骤一：**从客观角度分析。首先，将各年各项指标的总指标值进行平均，得到相应年份下各指标的平均值。算出各指标总体的年复合增长率，观察指标在五年内的变化情况。然后，按照各指标的指标值对村庄进行排名，选出 2015 年各指标下分别排名靠前的村庄。对它们 2020 年对应指标的排名进行匹配，观察排名变化。最后，参照 2015 年与 2020 年对应指标的指标值，分别计算排名靠前村庄不同指标下的年复合增长率。统计正增长与负增长的比率，了解指标变化的方向。综上，若 2015 年评分与 2020 年对应评分之间的相似度低，则不存在直接关联。

**步骤二：**结合主观评价分析。以指标为单位，按照各指标值分别对村庄进行排序并选取前 10000 名进行分析。参照相同指标下 2015 年与 2020 年的排名，对“优秀”与“退步”下定义，按照规定的评判标准对村庄进行分类。若该 10000 个村庄中“优秀”的占比未达到几乎全部的程度，则五年前的评分与 2020 年对应评分之间不存在直接关联。

**步骤三：**对指标数据进行相关性分析，参照指标间的相对系数分析指标之间的对应关系。观察同类指标不同时间之间、相同时间各指标之间以及所有指标之间的相对系数，进一步分析相互的对应关系。

## 2.3 问题二分析

**步骤一：**对五个评价指标的权重进行计算，考虑到本文数据较多且数据间存在较强相对关系，采用 CRITIC 赋权法确定关于绩效排名总分中各指标的赋权。

**步骤二：**由于本题已对数据进行了标准化处理，因此在各评价指标权重求解的基础上，关于绩效排名的总分即为相应评价指标值与相应指标权重的乘积和。

**步骤三：**选取关于绩效排名总分靠前的村庄，分别统计其中不同帮扶单位类型的个数。由于不同类型的帮扶单位的总个数相差较大，因此采用排名靠前中的个数与总个数的比值来评判在脱贫帮扶上面的绩效排名。分别对不同类型下各帮扶单位进行排序，整理得到各类型下脱贫帮扶绩效前十名的帮扶单位编号。

## 2.4 问题三分析

**思路一：**首先，以帮扶单位编号相同的村庄为整体，作差得到 2020 年与 2015 年相同指标下的差值。计算整体中差值的算数平均数，得到该帮扶单位在该评价指标上的平均值。然后，计算各帮扶单位在各评价指标下指标值的中位数。最后，将平均值与中位数相加平均，以平均结果作为各帮扶单位在特定评价指标上帮扶业绩排名的参考值。进而，得到在各单项评价指标上帮扶业绩明显的帮扶单位，以及前五名的帮扶单位编号。

**思路二：**在问题二所求指标权重的基础上，采用算法分别计算各帮扶单位各年在各单项评价指标上的贡献值。在一定程度上修正所得贡献值，使得数据更贴合实际。对不同年份下同一指标修正后的贡献值作差，差值即为帮扶单位在该评价指标上的贡献增加值。参照各单项指标上的贡献增加值对各帮扶单位进行排序，算数平均得到各集合村庄的平均排名，进而求解。

## 2.5 问题四分析

**步骤一：**筛选得到与各待补村庄帮扶单位编号、类型相同的村庄集合，与待补村庄进行聚类分析，筛选得到与待补村庄相似的村庄 X。结合村庄 X 与 32155 个村庄进行聚类分析，得到与村庄 X 相似的村庄集合。计算该集合下各指标 2020 年与 2015 年差值平均值，该差值平均值加上待补村庄 2015 年对应指标值即为待补村庄 2020 年的指标值。

**步骤二：**通过主成分分析确定各指标值在总分中的权重，计算相同指标下权重变化的差值，增加较大的即为评选先进村庄的主要指标。

**步骤三：**结合指标权重，重新计算各村庄各年的总分值。筛选各年前 10000 名的村庄，剔除 2020 年掉出前 10000 名的村庄，同时匹配 2020 年新晋村庄在 2015 年的总分。

对总分进行作差，以差值对村庄进行排序。前 2500 名即为“一级脱贫先进村庄”，第 2501-10000 名即为“二级脱贫先进村庄”，分别查找数据表最后 10 个村庄的称号情况。

## 2.6 问题五分析

结合数据与前面几问求解的结论，向国家扶贫办阐述在脱贫帮扶工作上的观点并提出建议。

## 2.7 总体思路流程



图 1 总体流程图

## 三、模型假设

1. 假设研究人员对数据标准化的处理合理；
2. 假设 2015 年与 2020 年指标值的赋值标准相同；
3. 假设各个村庄相应指标五年内的指标值变化，只要受帮扶单位脱贫帮扶工作的影响；
4. 假设 10 个待补村庄对 32155 个完整村庄总体的数据分析处理影响不大，可以在不考虑待补村庄的基础上，对其余村庄进行分析；

## 四、符号说明

符号	说明
$Q_N$	年份为 $N$ 时关于绩效排名的总分
$w_j^N$	年份为 $N$ 时评价指标 $j$ 的权重
$\sigma_j^N$	年份为 $N$ 时的对比强度
$c_j^N$	年份为 $N$ 时的冲突性
$M_j^N$	年份为 $N$ 时评价指标权重计算的中间值
$P_j$	帮扶单位在评价指标 $j$ 上的平均值
$Z_j$	帮扶单位在评价指标 $j$ 上的中值
$M$	表示待测评价指标的指标值
$m$	表示对应年份总分指标值
$r_{ij}$	评价指标 $i$ 与对应年份 $j$ 总分之间的距离
$K_{ij}$	帮扶单位在各单项评价指标上的贡献值

## 五、模型的建立与求解

### 5.1 数据预处理

将附件数据导入 Python, 对数据中的缺失值、重复值以及异常值进行检索。本题中, 重复值为村庄编号以及各指标值均相同的情况。在不考虑最后 10 个村庄存在缺失值的情况下, 其余数据不存在缺失值、重复值与异常值, 数据完整度为 100%。

### 5.2 问题一的求解

#### 5.2.1 多种客观分析下的指标变化

##### 1. 各指标总体的年复合增长率

年复合增长率是一项投资在特定时期内的年度增长率, 等于总增长率百分比的年数次方根减一后的差。结合本题数据, 公式转化为:

$$\text{年复合增长率} = \left( \frac{\text{2020 年指标值}}{\text{2015 年指标值}} \right)^{1/5} - 1 \quad (1)$$

由于数据表中最后 10 个村庄的数据不完整, 因此在求解中不考虑这 10 个村庄的数据。对 2015 年与 2020 年的五类指标值与总分 (各有 32155 个值) 分别进行平均, 得到相应年份下各指标的平均值。代入公式 1 得到对应指标总体的年复合增长率, 见表 1。

表 1 各指标总体的年复合增长率

	SR	CY	HJ	WJ	SS	总分
2015 平均值	0.057539	0.223842	0.111885	0.233446	0.225351	0.213066
2020 平均值	0.032438	0.024571	0.005158	0.004557	0.020135	0.015433
复合增长率	-10.83%	-35.72%	-45.96%	-54.49%	-38.31%	-40.85%

由表 1 数据可知, 每个指标整体均呈明显下降趋势。下降原因可能有以下几点:

- (1) 受 2020 年疫情的影响, 国民经济发展趋缓, 居民收入与产业发展有所下降;
- (2) 居住环境、文化教育与基础设施受到宏观经济的影响, 没有得到大力改善。

##### 2. 各指标值下的排名变化

为方便数据的统一处理与选择, 将附件数据按照村庄编号升序排列。分别计算 2015 年与 2020 年不同指标值下所有村庄的排名情况, 选取各指标 2015 年排名前十的村庄以及它们 2020 年时相同指标值下的排名情况 (见表 2)。

表 2 各指标值下的村庄排名

村庄	2015SR	2020SR	村庄	2015CY	2020CY	村庄	2015HJ	2020HJ
10044	1	9119	10216	1	90	10235	1	1098
18853	2	6846	10257	1	13	10553	1	1952
18893	3	10319	10355	1	871	10724	1	8294
18911	4	14204	15756	1	2091	10948	1	1639
35693	4	1231	25429	1	29	19013	1	5935
村庄	2015WJ	2020WJ	村庄	2015SS	2020SS	村庄	2015 总分	2020 总分
10006	1	2035	10001	1	860	18911	1	511
10009	1	44	10003	1	1631	18635	2	1289
10015	1	1532	10004	1	3024	15537	3	465
10016	1	1224	10006	1	2847	18739	3	320
10022	1	8085	10016	1	317	10808	5	263

观察发现: 1. 2015 年在某方面排名靠前的村庄 2020 年排名不一定靠前, 即不同指

标下的村庄排名并非固定不变的；2.2015 年排名靠前的村庄虽然在 2020 年有所下降，但大多数的排名在 32155 个村庄中依旧靠前。

3.不同村庄的各指标年复合增长率

参照 2015 年各项指标的指标值分别对 32155 个村庄进行排序，选取各项指标下排名前 10000 名的村庄，参照 2020 年对应指标的指标值算出该项指标下的年复合增长率。对不同指标下 10000 名村庄的年复合增长率进行整理，统计年复合增长率的正负个数，并算出正负比值。

表 3 各指标前 10000 名的复合增长率统计情况

指标	正复合增长率占比	负复合增长率占比	比值
SR	28.27%	71.73%	0.39
CY	34.89%	65.11%	0.54
HJ	54.42%	45.58%	1.19
WJ	38.11%	61.89%	0.62
SS	42.50%	57.50%	0.74
总分	44.36%	55.64%	0.80

由表 3 数据可知，除文化教育指标下前 10000 名村庄的正增长个数多于负增长以外，其余指标下负增长多于正增长。

5.2.2 结合主观评价的指标变化

对按照各指标值分别对 32155 个村庄进行排序，选取 2015 各个指标下排名前 10000 名的村庄。以指标为单位，分别匹配选取的前 10000 个村庄 2020 年时相应指标下的排名情况。判定该村庄在该指标上为“优秀”的条件如下：

- 1.相应指标下，2020 年的排名相对 2015 年有所下降，若差值小于 6000 且 2020 年排名仍处于前 10000 名；
- 2.若 2020 年的排名相对 2015 年有所进步，即两者均在前 10000 名内。

不满足上述评判条件，则该村庄在该指标上为“退步”。具体流程见图 2。

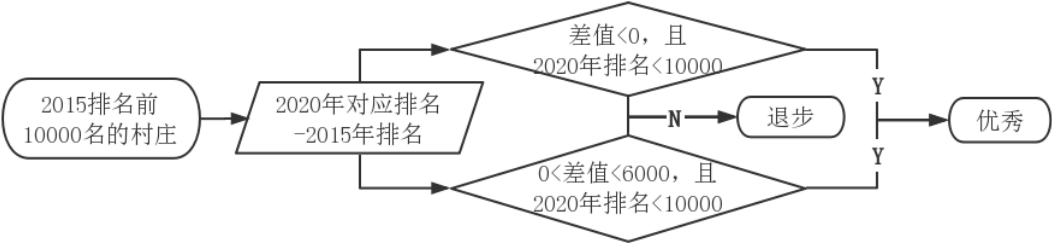


图 2 主观评价流程图

对结果进行统计，得到表 4。观察可知所以指标中优秀的占比均大于退步的占比，由此得出：2015 年的指标值好，2020 年的指标值不一定好；2015 年的指标值好，2020 年指标值继续好的概率大于退步的概率。

表 4 主观评价下的指标变化

指标	优秀占比	退步占比	比值
SR	53.10%	46.90%	1.13
CY	59.70%	40.30%	1.48
HJ	68.60%	31.40%	2.18
WJ	61.40%	38.60%	1.59
SS	59.60%	40.40%	1.48
总分	73.50%	26.40%	2.78

5.2.3 评价指标的相关性分析

相关性分析可以适当描述客观事物相互关系的密切程度，通过相关系数  $r$  可以表示两个指标间的相关程度。 $r$  的绝对值越接近 1，则两个变量之间的关联程度越强。

为探究各个评价指标之间的关系，将附件数据导入 Python，利用 pandas 科学分析库<sup>[1]</sup>对各指标进行数据与相关性分析，具体代码见附录一。

表 5 各指标的相关性分析

	2015SR	2015CY	2015HJ	2015WJ	2015SS	2015 总	2020SR	2020CY	2020HJ	2020WJ	2020SS	2020 总
2015SR	1	0.514	0.573	0.519	0.525	0.732	0.534	0.478	0.567	0.439	0.445	0.559
2015CY	0.514	1	0.566	0.683	0.669	0.837	0.475	0.640	0.555	0.561	0.535	0.646
2015HJ	0.573	0.566	1	0.562	0.585	0.817	0.530	0.546	0.737	0.477	0.481	0.632
2015WJ	0.519	0.683	0.562	1	0.734	0.858	0.502	0.638	0.578	0.636	0.603	0.698
2015SS	0.525	0.669	0.585	0.734	1	0.856	0.499	0.633	0.579	0.589	0.594	0.678
2015 总	0.732	0.837	0.817	0.858	0.856	1	0.615	0.717	0.740	0.660	0.648	0.784
2020SR	0.534	0.475	0.530	0.502	0.499	0.615	1	0.600	0.648	0.557	0.569	0.729
2020CY	0.478	0.640	0.546	0.638	0.633	0.717	0.600	1	0.685	0.758	0.738	0.890
2020HJ	0.567	0.555	0.737	0.578	0.579	0.740	0.648	0.685	1	0.664	0.670	0.850
2020WJ	0.439	0.561	0.477	0.636	0.589	0.660	0.557	0.758	0.664	1	0.758	0.894
2020SS	0.445	0.535	0.481	0.603	0.594	0.648	0.569	0.738	0.670	0.758	1	0.874
2020 总	0.559	0.646	0.632	0.698	0.678	0.784	0.729	0.890	0.850	0.894	0.874	1

由表 5 数据可知，任意两个评价指标间均呈正相关。采用色阶对相关系数进行整理，采用不同色系表示相关性的强弱。蓝色系的表明相应指标间的相关性较低，白色系表明相应指标间的相关性一般，粉色系表明相应指标间的相关性较强。

在上述基础上，排除相关系数为 1 的自身，五个指标均与总分相关性最强。因此，在不考虑自身与总分指标的情况下，分别对五个评价指标之间的相对关系强弱进行整理，得到表 6。

表 6 评价指标的对应关系

评价指标	最强	第二强	最弱	评价指标	最强	第二强	最弱
2015SR	2015HJ	2020HJ	2020WJ	2020SR	2020HJ	2020CY	2015CY
2015CY	2015WJ	2015SS	2020SR	2020CY	2020WJ	2020SS	2015SR
2015HJ	2020HJ	2015SS	2020WJ	2020HJ	2015HJ	2020CY	2015CY
2015WJ	2015SS	2015CY	2020SR	2020WJ	2020CY	2020SS	2015SR
2015SS	2015WJ	2015CY	2020SR	2020SS	2020WJ	2020CY	2015SR

#### 5.2.4 评价指标关系小结

由 2015 年及 2020 年各指标平均值的年复合增长率结果可知，2020 年各指标相较 2015 年均有明显下降。参照各指标值下的村庄排名状况可知，特定指标下 2015 年排名靠前的村庄在 2020 年的排名不一定靠前，但 2015 年排名靠前的村庄在 2020 年排名依旧靠前的概率较大。结合 2015 年特定指标值下排名前 10000 名村庄该指标在 2015 年至 2020 年间的复合增长率可知，除文化教育指标外，其余指标下前 10000 名村庄的负增长个数多于正增长。

综上，本问不存在如“五年前的居民收入不错，现在的居民收入也会是不错的”的规律，但 2015 年好的村庄在 2020 年继续好的概率较大。结合评价指标的相关性分析，可知：

(1) 居民收入与居住环境之间的相关性最强，产业发展、基础设施与文化教育之间的相关性最强；

(2) 2015 年各指标相关性最弱的指标为 2020 年的其余指标，同理，2020 年各指标相

关性最弱的指标为 2015 年的其余指标。

### 5.3 问题二的求解

分别对五个评价指标的权重进行计算，令  $w_1^N$ 、 $w_2^N$ 、 $w_3^N$ 、 $w_4^N$ 、 $w_5^N$  分别为年份为  $N$  时居民收入(记为 SR)、产业发展(记为 CY)、居住环境(记为 HJ)、文化教育(记为 WJ)、基础设施(记为 SS)这五个评价指标的权重。为了便于比较和研究，本题已对所有数据进行了标准化处理。因此，对应指标值与对应指标权重的乘积和即为关于绩效排名的总分  $Q$ 。

$$Q_N = w_1^N SR_N + w_2^N CY_N + w_3^N HJ_N + w_4^N WJ_N + w_5^N SS_N \quad (2)$$

其中：

$$w_1^N + w_2^N + w_3^N + w_4^N + w_5^N = 1, w_1^N, w_2^N, w_3^N, w_4^N, w_5^N \geq 0。$$

#### 5.3.1 CRITIC 赋权法<sup>[2]</sup>求评价指标权重

常见的指标赋权方法主要分为主观赋权法和客观赋权法。主观赋权法容易受赋权人员意愿等主观因素的影响，客观赋权法则依据数据本身之间的内在联系进行权重求解。应用较多的主观赋权法主要有熵权法与主成分分析法，但由上文可知本题指标见得相关性较强，且上述两种方法无法在不删除数据的情况下有效避免。因此，本文采用 CRITIC 赋权法对评价指标进行赋权，该方法能同时考虑指标之间的对比强度及冲突性。

##### 1. 计算对比强度 $\sigma_j^N$

对比强度有各指标的标准差决定，计算题中已标准化处理后的各评价指标值的标准差，得到表 7。

表 7 各评价指标的标准差

指标	标准差	指标	标准差	指标	标准差	指标	标准差
2015 SR	0.950	2015 WJ	0.843	2020 SR	0.934	2020 WJ	0.980
2015 CY	0.806	2015 SS	0.814	2020 CY	0.965	2020 SS	0.980
2015 HJ	0.900	2015 总分	0.844	2020 HJ	0.977	2020 总分	0.967

##### 2. 计算冲突性 $c_j^N$

指标间的冲突性由相关系数决定，具体公式如下：

$$c_j^N = \sum_{k=1}^p (1 - \text{corr}_{k,j})$$

其中：  $\text{corr}_{k,j}$ ——指标  $k$  与指标  $j$  之间的相关系数

各指标之间的相关系数在上文中已有展示，具体见表 5。参照公式 3 计算得到各评价指标间的冲突性，见表 8。

表 8 各指标间的冲突性

指标	冲突性	指标	冲突性	指标	冲突性	指标	冲突性
2015 SR	5.115	2015 WJ	3.991	2020 SR	4.741	2020 WJ	4.008
2015 CY	4.319	2015 SS	4.058	2020 CY	3.677	2020 SS	4.085
2015 HJ	4.494	2015 总分	2.735	2020 HJ	3.727	2020 总分	2.765

##### 3. 计算权重 $w_j^N$

根据评价指标的对比强度与冲突性，参照公式 4 对个评价指标的权重进行重新计算，得到表 9。具体代码见附录二。

$$w_j^N = \frac{P_j^N}{\sum_{j=1}^n M_j^N} \quad (4)$$

其中：  $P_j^N = \sigma_j^N \times c_j^N, j = 1, 2, 3, 4, 5。$



表 9 各评价指标的权重计算中间值及结果

	SR	CY	HJ	WJ	SS
$P_j^{2015}$	0.227	0.195	0.246	0.200	0.200
$w_j^{2015}$	0.213	0.183	0.230	0.187	0.187
$P_j^{2020}$	0.191	0.204	0.247	0.254	0.225
$w_j^{2020}$	0.171	0.182	0.220	0.227	0.201

### 5.3.2 参照关于绩效排名的总分求解

在已知各评价指标值与各评价指标权重的基础上，参照公式 2 即可得到重新计算后各村庄在 2015 年与 2020 年的总分。在此，定义村庄 2020 年总分与 2015 年总分之间的差值为相应帮扶单位在脱贫帮扶上的绩效值。即：

$$\text{绩效值} = Q_{2020} - Q_{2015}$$

对绩效值进行排序，取绩效值排名前 6000 的村庄为样本，分别统计 6 个类型帮扶单位的个数。由于各类型帮扶单位的个数不相等，因此结合 32155 个村庄中对应类型帮扶单位的个数，求出两者之间的比值（表 10）。从而，判断在脱贫帮扶方面具有较高绩效的帮扶单位类型。

表 10 具有较高绩效的帮扶单位类型

帮扶单位类型	前 6000 内的个数	总数	占比	排名
0	452	2483	18.20%	4
1	2046	12492	16.38%	6
2	455	2131	21.35%	3
3	1977	11279	17.53%	5
4	11	29	37.93%	1
5	1059	3741	28.31%	2

由表 10 可知，帮扶单位类型为 4、5 的帮扶单位在脱贫帮扶上面有显著的绩效。对比图见图 3。

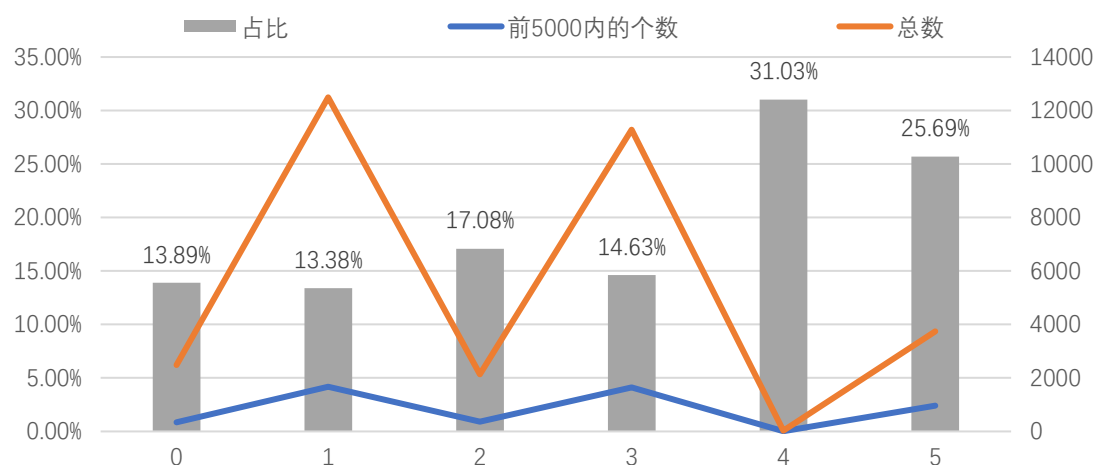


图 3 具有较高绩效的帮扶单位类型

分别对不同类型下各帮扶单位有关绩效排名的总分进行排序，整理得到各类型下脱贫帮扶绩效前十名的帮扶单位编号，见表 11、表 12。

表 11 0-2 类型下脱贫帮扶绩效前十名的帮扶单位编号

排名	单位 类型	村庄 编号	单位 编号	单位 类型	村庄 编号	单位 编号	单位 类型	村庄 编号	单位 编号
1	0	21086	42	1	50568	146	2	23188	54
2	0	24620	62	1	17234	25	2	35297	83
3	0	16237	22	1	22871	52	2	33739	77
4	0	15746	21	1	47961	138	2	47309	137
5	0	15685	21	1	48432	138	2	23170	54
6	0	10312	0	1	27792	96	2	23165	54
7	0	16085	22	1	13539	12	2	33760	77
8	0	10060	0	1	50745	146	2	23223	54
9	0	21099	42	1	44297	129	2	35298	83
10	0	18908	33	1	47956	138	2	23267	54

表 12 3-5 类型下脱贫帮扶绩效前十名的帮扶单位编号

排名	单位 类型	村庄 编号	单位 编号	单位 类型	村庄 编号	单位 编号	单位 类型	村庄 编号	单位 编号
1	3	38047	113	4	33524	75	5	38737	115
2	3	46610	131	4	20823	46	5	52794	153
3	3	38191	113	4	33425	75	5	46409	134
4	3	38211	113	4	20867	46	5	46248	134
5	3	46533	131	4	33439	75	5	46783	133
6	3	38012	113	4	20882	46	5	46200	134
7	3	38055	113	4	33513	75	5	46984	133
8	3	46588	131	4	20817	46	5	38657	115
9	3	45562	136	4	33364	75	5	46696	132
10	3	38109	113	4	20865	46	5	38708	115

## 5.4 问题三的求解

以帮扶单位的单位编号为筛选条件，通过 Python 对 32155 个村庄进行筛选分类，得到编号分别为 0-159 的 160 个村庄集合。

### 5.4.1 结合统计指标求排名

平均数是表示一组数据集中趋势的量数，但它容易受数据中心极端数值的影响。中位数不受分布数列中极大或极小值的影响，常被用于描述数据的集中趋势。因此，本文综合运用这两个统计指标对不同编号帮扶单位在帮扶上的业绩进行分析。

首先，以帮扶单位编号相同的村庄为整体，在相同评价指标下，将 2020 年的指标值与 2015 年的指标值作差，算术平均后得到各单位编号下帮扶单位在该评价指标上的平均值  $P$ 。

$$P_j = \frac{\sum_{x=1}^n (j_x^{2020} - j_x^{2015})}{n} \quad (5)$$

然后，以帮扶单位编号相同的村庄为整体，在相同评价指标下，参照公式 6 计算各单位编号下帮扶单位在该评价指标上的中位数  $Z$ 。

$$Z_j = \begin{cases} \frac{j_{\frac{n+1}{2}}^N}{2}, & n \text{ 为奇数} \\ \frac{j_{\frac{n}{2}}^N + j_{\frac{n+1}{2}}^N}{2}, & n \text{ 为偶数} \end{cases} \quad (6)$$

最后，将平均数与中位数相加，平均得到不同单位编号的帮扶单位在特定评价指标上的得分，如公式 7 所示。在特定评价指标下，按照该评价指标上的得分对 160 个不同帮扶单位编号的村庄集合进行排序，筛选得到该指标下帮扶业绩明显的前五名，见表 13。

表 13 各指标下帮扶业绩明显的前五名

排名	单位 编号	SR 得分	单位 编号	CY 得分	单位 编号	HJ 得分	单位 编号	WJ 得分	单位 编号	SS 得分
1	115	1.569	149	1.496	149	1.925	113	1.483	158	1.799
2	134	1.302	113	1.393	75	1.866	132	1.326	153	1.698
3	54	1.121	158	1.210	153	1.758	115	1.119	113	1.460
4	131	1.068	115	0.865	46	1.627	131	0.999	149	1.380
5	149	1.031	114	0.632	113	1.436	158	0.996	131	1.350

#### 5.4.2 基于万有引力算法的修正后差值求排名

计算平均数与中位数的平均值，进而求排名的方法过于简易及粗略。因此，本文基于万有引力算法，从另一角度对排名进行求解。

##### 1. 万有引力算法

万有引力算法<sup>[3]</sup>是在牛顿万有引力定律以及粒子之间相互作用的基础上被提出的智能优化算法。结合万有引力的定义以及物理学中的相关知识，可知万有引力是任意两个物体或两个粒子间的距离与其质量乘积相关相互作用的吸引力。因此，它的大小与物体的质量以及两个物体之间的距离有关。物体的质量越大，之间的万有引力就越大，距离越远，之间的万有引力就越小。

这种关系同样可以应用到其他方面，用来衡量自变量与因变量之间贡献强度的联系。将对应年份下各评价指标得指标值与总分指标的指标值代入模型，可以求出对应自变量指标贡献强度的大小。参考万有引力在城市经济方面的运用模型<sup>[4]</sup>，对其进行改进。原模型公式 7 如下：

$$K_{ij} = \frac{\sqrt{M}\sqrt{m}}{r_{ij}^2} \quad (7)$$

由于原模型涉及到根号，但本题标准化后的数值较小且存在负值，因此本文对模型进行改进：将开平方改为开立方， $r_{ij}^2$ 变成 $r_{ij}$ 。改进后的模型如公式 8 所示，其求解既符合原理又保证数值结果不失真。

$$K_{ij} = \frac{(M*m)^{\frac{1}{3}}}{r_{ij}} \quad (8)$$

其中： $M$ ——表示待测评价指标的指标值；

$m$ ——表示对应年份总分指标值；

$r_{ij}$ ——评价指标  $i$  与对应年份  $j$  总分之间的距离；

$K_{ij}$ ——帮扶单位在各单项评价指标上的贡献值。

对上文采用 CRITIC 赋权法求出的各评价指标权重进行整理，定义其倒数为评价指标与对应年份总分之间的距离，即  $r_{ij}$ 。具体见表 14。

表 14 各评价指标权重及其倒数 $r_{ij}$

	SR	CY	HJ	WJ	SS
$w_j^{2015}$	0.213	0.183	0.230	0.187	0.187
$r_{ij}$	4.695	5.464	4.348	5.348	5.348
$w_j^{2020}$	0.171	0.182	0.220	0.227	0.201
$r_{ij}$	5.848	5.495	4.545	4.405	4.975

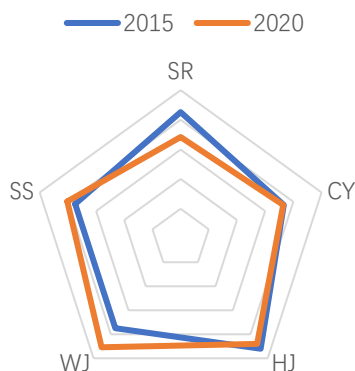


图 4  $w_j^N$  雷达图

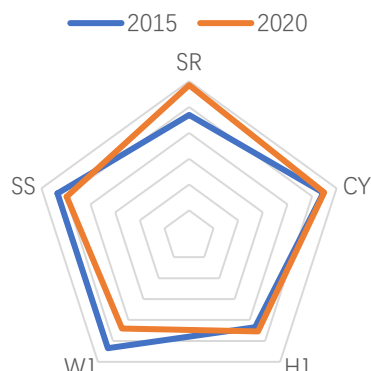


图 5  $r_{ij}$  雷达图

图 4 代表面积雷达图观察得出 SR 的影响力在下降,但是 WJ、SS 的重要性在提高。图 5 代表距离雷达图, 观察发现: SR 在偏离总分, WJ、SS 在靠近总分

## 2. 修正万有引力算法输出值 $K_{ij}$

由于原有评价指标的指标值与总分指标值均存在正负值, 万有引力算法的部分运行结果偏离实际, 对结果的影响较大。具体有如下情形:

(1) 指标值为正、总分指标值为负, 此时的输出结果为负数, 表明正指标值对总分存在负面影响;

(2) 指标值为负、总分指标值为负, 表明负指标值对总分存在正面影响。

上述两种情形的结论明显偏离实际, 因此本文对万有引力算法输出值 $K_{ij}$ 进行修正。首先, 对相应评价指标值与总分指标值的符号进行判别。然后, 参照如下修正标准对输出值进行修正:

(1) 若评价指标值与总分指标值同为负, 则输出  $-K_{ij}$  ;

(2) 若评价指标值为正, 总分指标值为负, 则输出 $-K_{ij}$ ;

(3) 其余符号情况统一输出 $K_{ij}$ 。

最后, 得到修正后的  $K_{ij}$  。以居民收入 (SR) 这一指标为例, 对修正前后的 $K_{ij}$ 进行展示, 见表 15。结合原始指标值, 对比发现对 $K_{ij}$ 的修正合理。

表 15 修正前后值对比

村庄 编号	2015 SR	2015 总分	2015 修正前	2015 修正后	2020 SR	2020 总分	2020 修正前	2020 修正后
46576	-2.1168	-2.4	0.37	-0.37	0.83201	-1.2057	-0.17	0.17
46517	-2.4788	-3.0517	0.42	-0.42	0.28506	-1.1657	-0.12	0.12
38079	-0.88563	-2.947	0.29	-0.29	1.1055	-1.665	-0.21	0.21
46984	-4.0721	-2.0742	0.43	-0.43	0.19391	-0.18716	-0.06	0.06
45684	-2.0443	-1.9345	0.34	-0.34	0.6497	-1.0559	-0.15	0.15
22551	-1.465	-1.6086	0.28	-0.28	1.1055	-1.3055	-0.19	0.19
.....	.....	.....	.....	.....	.....	.....	.....	.....

## 3. 对不同年份下同一指标的 $K_{ij}$ 作差

在评价指标值修正的基础上, 将评价指标 2020 年时的值减去相应指标在 2015 年时的值, 得到的各评价指标之间的差值即为帮扶单位在各评价指标上的贡献增加值, 即为该帮扶单位在具体评价指标上帮扶业绩的参考值。

在此, 以居民收入 (SR) 这一指标为例, 对修正前后的排名进行分析。

表 16 修正前后排名变化

村庄编号	修正 $K_{ij}$ 前差值排名	修正 $K_{ij}$ 后差值排名
46576	32154	1
46517	32155	2
38079	32153	3
46984	32151	4
45684	32150	5
22551	32149	6
.....	.....	.....

4. 各评价指标上帮扶业绩明显的单位编号

参照同一指标不同年份下修正后 $K_{ij}$ 的差值，按照特定评价指标下的差值分别对 32155 各村庄进行排序。分别统计 160 个村庄集合中各村庄的排名值，计算集合内村庄的平均排名。对各集合下的平均排名进行排序，得到单项评价指标前五名的帮扶单位编号，见表 17。

表 17 各指标下帮扶业绩明显的前五名

排名	单位 编号	SR 平均 排名	单位 编号	CY 平均 排名	单位 编号	HJ 平均 排名	单位 编号	WJ 平均 排名	单位 编号	SS 平均 排名
1	153	5923.36	149	2665.19	0	7123.23	75	516.92	149	6076.96
2	115	6153.75	158	2897.35	149	7442.02	132	916.47	158	6724.17
3	113	6236.75	113	3264.95	116	8254.92	113	1037.41	153	7614.15
4	149	6812.60	153	3921.97	21	8639.32	66	1168.22	113	7767.86
5	54	6985.24	115	5657.89	89	8840.78	72	1451.18	155	8573.79

5.4.3 检验

参照各指标值下贡献增加值的大小，对 32155 个村庄进行排序。取前 10000 名为样本，分别统计各评价指标下 160 个集合中村庄在前 10000 名内的个数。与集合内的村庄总数相除，即可得到不同评价指标下该集合内排名在前 10000 名的村庄占比。依据占比对 160 个村庄集合进行排序，得到各帮扶单位在各单项评价指标上帮扶业绩的排名。

以居民收入（SR）、产业发展（CY）指标为例，对结果进行检验。由检验结果（表 18）可知，三者前五名单位编号的误差较小且相似度高。对 SR 指标下万有引力算法排名的结果进行展示，见图 6。

表 18 检验结果

SR 排名	统计	万有引力	检验	CY 排名	统计	万有引力	检验
1	115	153	153	1	149	149	158
2	134	115	115	2	113	158	153
3	54	113	113	3	158	113	149
4	131	149	149	4	115	153	113
5	149	54	131	5	114	115	115

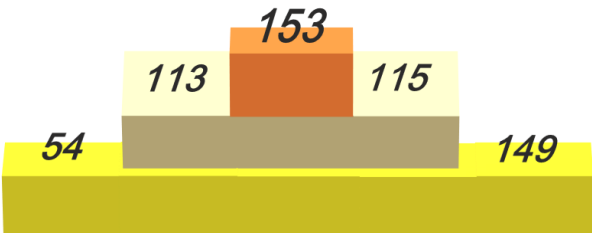


图 6 SR 指标下万有引力算法排名结果

5.5 问题四的求解

5.5.1 双重聚类协同过滤<sup>[5]</sup>预测算法补全空缺指标值

1. 一次聚类

定义 2015 年与 2020 年指标值均完整的村庄集合为 A。在村庄集合 A 中,利用 Python 分别筛选得出与 10 个待补村庄具有相同帮扶单位编号、帮扶单位类型的村庄,达到群体过滤的效果。定义这些村庄集合为 B,村庄集合 B 为一次聚类的聚类因子群。在已知待补村庄 2015 年各项指标值的基础上,结合村庄集合 B 中各村庄 2015 年各评价指标值进行聚类分析。

参照 SPSS 聚类分析生成的谱系图,分别计算得到村庄集合 B 中与各待补村庄相似度最为接近的村庄,记该村庄为村庄 X。

2. 二次聚类

在 32155 个村庄中对 2015 年-2020 年所有指标进行二次聚类,参照谱系图找到与村庄 X 各年各指标水平相近且拥有共同发展趋势的聚类因子群。进而完成双重聚类协同过滤,具体流程见图 7。

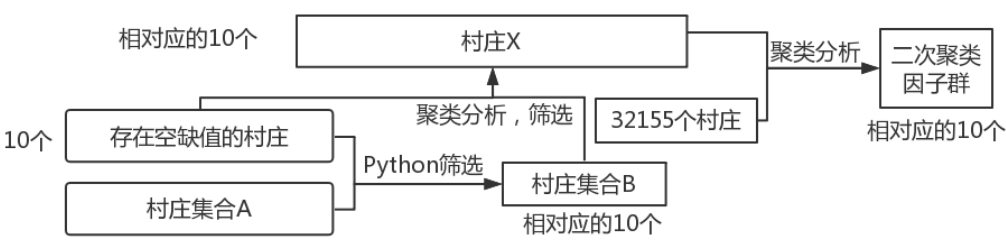


图 7 双重聚类协同过滤流程图

3. 补全空缺指标值

在二次聚类得到的聚类因子群中,计算各村庄 2020 年各评价指标与 2015 年对应指标值的差值,对差值取平均值。令该平均值为待补村庄五年扶贫的指标增加值,与 2015 年对应指标值相加得到 2020 年相应指标值。补全结果见表 19。

表 19 待补村庄补全结果

村庄 编号	单位 编号	帮扶单位 类型	2020 SR	2020 CY	2020 HJ	2020 WJ	2020 SS	2020 总分
39257	116	1	2.0120	0.8246	1.4531	0.6204	0.7820	1.1106
25149	89	1	1.9852	1.4296	1.6336	1.6164	1.6344	1.6525
12722	7	2	0.3977	0.4352	0.9860	0.9411	0.8442	0.7467
12916	10	2	1.2116	1.5724	1.2909	1.6573	1.4283	1.4391
21570	47	1	-0.4864	0.3552	0.8943	0.7199	0.1891	0.3797
22096	48	1	0.6665	1.6079	0.7619	0.6572	0.0338	0.7299
47883	138	1	1.2735	0.7395	1.5515	0.4704	0.7023	0.9410
34208	78	1	0.7489	0.7546	1.2884	0.1849	0.8712	0.7655
34276	78	1	0.6764	1.3478	1.3311	1.1221	1.1554	1.1398
52436	151	3	0.2238	-0.4115	-0.2922	0.0136	-0.2569	-0.1496

5.5.2 主成分分析法求权重<sup>[6]</sup>确定主要指标

在上文补全 10 个待补村庄空缺值的基础上,通过 SPSS 分别对 32165 个村庄 2015 年、2020 年的五个评价指标进行主成分分析。将主成分分析得到的成分矩阵及各指标的成分根拷到 Excel,各指标线性组合中的系数即为标准化数与对应主成分特征根平方根的商。计算综合得分模型中的系数,最后对所有指标权重进行归一化处理,使得权重和为 1。各指标权重结果见表 20,绘制雷达图得到图 8。

表 20 主成分分析下各指标权重及变化

	SR	CY	HJ	WJ	SS
2015	0.1827	0.2046	0.1937	0.2090	0.2100
2020	0.2074	0.2051	0.2051	0.2002	0.1822
权重变化	0.0247	0.0005	0.0114	-0.0088	-0.0278

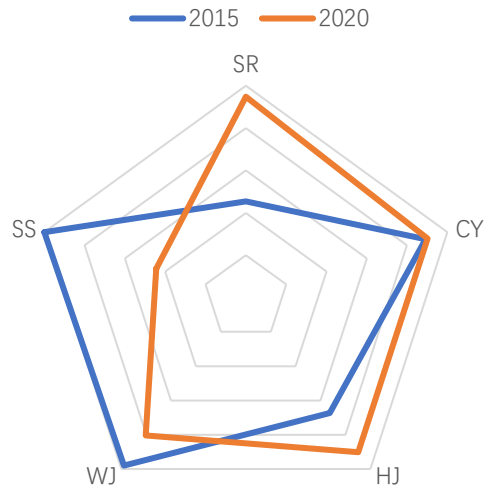


图 8 各指标权重变化雷达图

由表 20 及图 8 可知，居民收入与居住环境对村庄脱贫的总分有较大的关系。因此。这两指标对村庄评选“脱贫先进村庄”有着非常重要的影响。

5.5.3 脱贫先进村庄的评选

参照主成分分析确定的各指标权重，分别计算各指标值与对应权重相乘后的乘积和，该乘积和即为对应村庄 2015 年与 2020 年的总分。按照总分进行排序，分别选出 2015 年、2020 年排名前 10000 名的村庄，剔除 2015 年前 10000 名但 2020 年掉出前 10000 名的村庄。剩余村庄中筛选得到 2020 年在前 1000 名但 2015 年不在前 10000 名的村庄，分别匹配它们在 2015 年的总分。

分别计算剩余村庄 2020 年总分与 2015 年总分的差值，按照差值对它们进行排序，前 2500 名即为“一级脱贫先进村庄”，第 2501-10000 名即为“二级脱贫先进村庄”。其中，原始数据表中最后 10 个村庄的排名与称号情况如表 21 所示。

表 21 原始数据表中最后 10 个村庄的排名与称号情况

级别	村庄编号	排名
脱贫先进一级村庄	25149	1813
	12916	2334
	47883	2338
脱贫先进二级村庄	34208	3178
	34276	3369
	22096	3725
	39257	3798
	12722	9338
-	52436	-
	21570	-



5.5 问题五的书信

尊敬的国家扶贫局领导们：

你们好！

脱贫帮扶是我国战略中极为重要的一部分，为支持与参加脱贫帮扶工作，我们采用多种算法对全国 32165 个需要帮扶的贫困村数据进行了分析。

在已知各年五个评价指标值与总分的基础上，我们从主客观两个角度、共采用四种方法对五年间指标值的变化进行了计算，同时通过相关性分析了解指标间的对应关系。我们发现：

- 1.五年前与五年后的指标值之间无直接关系，但保持原有优劣水平的概率较大；
- 2.居民收入与居住环境之间的相关性最强，产业发展、基础设施与文化教育之间的相关性最强。

因此，脱贫帮扶工作的开展可以从“保优助劣”出发，居民收入的提升可以改善居住环境，产业发展可以带动基础设施与文化教育的进步。

通过对各评价指标的分析，分别得到不同评价指标上帮扶绩效显著的帮扶单位类型以及编号。如：单位类型为 4、5 的帮扶单位，分别有 37.93%、28.31%的村庄占总排名的前 6000 名；单位编号为 149 的帮扶单位在产业发展与文化教育上的帮扶业绩显著……

单位类型 4 下的村庄仅有 29 个，但其帮扶绩效最为显著。因此，可以增加该单位类型下所述的村庄数，重复发挥其在脱贫帮扶上的作用。

通过对主成分分析，发现居民收入与居住环境指标对“脱贫先进村庄”有着重要影响。因此，村庄的脱贫帮扶可以加大这两方面的举措。同时，可参考这些指标表彰荣誉村庄，从而激励其余村庄。

此致  
敬礼！

六、模型拓展

6.1 完善评价脱贫帮扶的指标库

题目中所给出的评价指标主要集中在绩效层面，查阅文献发现脱贫帮扶需要从多方面进行综合评价。结合文献，本文给出较完整的评价指标库为脱贫帮扶的后续研究提供参考。其中，公共财政支出包含了题中的部分指标。

表 22 评价指标库

指标类型	指标名称	单位
推广指标	第一产业从业人数	人
	农作物播种面积	亩
	农业机械总动力	千瓦
	财政扶贫资金	万元
	年末金融机构贷款约	万元
初步指标	第一产业增加值	万元
	年末脱贫人数	人
	农村居民人均可支配收入	万元
环境因素	地区生产总值	万元
	公共财政收入	万元



绩效因素	公共财政支出	万元
	耕地面积	亩
	城镇化水平	%

## 6.2 脱贫帮扶绩效识别系统的建立

### 步骤一：设计数据收集表

表 23 数据收集表

村庄编号：		帮扶单位：		单位类型：		识别评价结果：	
指标名称	上年得分	本年得分	变化	本年位次	变化	指标贡献强度	变化
公共财政支出							
教育支出							
基础设施支出							
医疗社保支出							
环境治理支出							
年末脱贫人数							
第一产业增加值							
企业帮扶事迹 年度汇报	A 级						
	B 级						
	C 级						
总体得分							

### 步骤二：导入系统

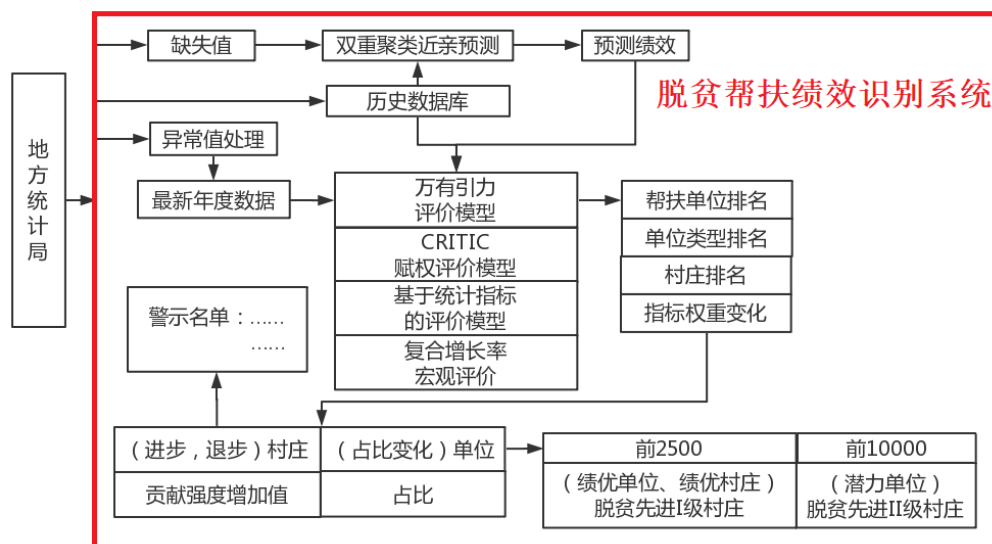


图 9 脱贫帮扶绩效识别系统

- 1.挖掘缺失值、异常值分模块处理。
- 2.缺失值进入结合历史数据库的指标预测模块，异常值处理完成后提前进入评价模块。（万有引力评价模型、CRITIC 赋权评价模型、统计指标评价模型、复合增长率宏观评价）
- 3.汇出不同类别排名（帮扶单位排名、单位类型排名、村庄排名、指标权重变化）
- 4.对比识别历史排名汇总村庄进退情况、贡献强度增加值、村庄前 10000 名不同帮

扶单位占比。

5.分析出结论，点赞前 2500 名的绩优村庄（脱贫先进 I 级村庄）、2501-10000 潜力村庄（脱贫先进 II 级村庄）;批评并激励排名靠后一万名村庄，攻坚克难，齐力奔康。

七、稳健性及灵敏度分析

7.1 稳健性分析

在双重聚类协同过滤预测算法中，预测模型结果的好坏在一定程度上影响最后统计方式，上文中采取的统计方式为差值。为了明确不同统计方式对模型稳健性的影响，在此通过复合增长率预测待补村庄 2020 年的各评价指标值。以编号为 39257 的村庄为例，利用 excel，进行模型预测结果的对比分析，见表 24。

表 24 不同统计方式的稳健性说明

差值补充法	2.01	0.82	1.45	0.62	0.78	1.11
复合增长率反推法	2.33	0.84	1.78	0.64	0.66	1.22

由表 22 可知，不同统计方式对模型影响很小，模型稳健性较大。因此，本文基于双重聚类协同过滤预测算法选择差值补充法的求解模型准确且有效。

7.2 灵敏度分析

(1) 为了进一步研究上述方法中万有引力模型的实用性和可推广性，则需要深入讨论其组成元素对模型的敏感性。则以 10001 村庄万有引力模型求解中的 SR 为例，分析模型灵敏性。

表 25 模型灵敏性

指标	2015 得分	变动	2015 贡献强度	变化	2020 得分	变动	2020 贡献强度	变化
M	1.5794	5%	0.28	4%	1.2565	5%	0.22	-
	1.5493	3%	0.27	-	1.2326	3%	0.22	-
	1.5042	-	0.27	-	1.1967	-	0.22	-
	1.4591	-3%	0.27	-	1.1608	-3%	0.22	-
	1.4290	-5%	0.27	-	1.1369	-5%	0.22	-

结论：当M在[-5%-5%]区间内波动时，其对贡献强度的影响微乎其微，表明模型对M的敏感度非常小。

(2) 分析万有引力模型中 $K_{ij}$ 和 $r_{ij}$ 的图像关系

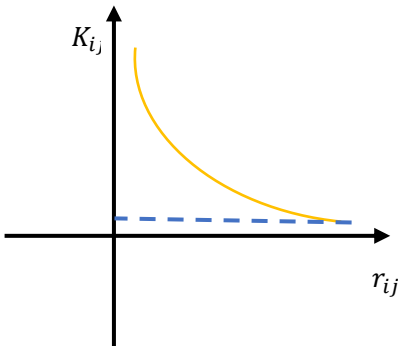


图 10 万有引力模型中 $K_{ij}$ 和 $r_{ij}$ 的图像关系

由图 10 发现，二者关系为典型反比例函数，意味着当 $r_{ij}$ 增大到一定是水平时，输出结果 $K_{ij}$ 的变化趋于平缓，逐渐逼近一条渐近线。因此模型结果在一定范围内对 $r_{ij}$ 不显著敏感。

## 八、参考文献

- [1]刘子桦,马若炎.基于 Python 关于世界自杀率影响因素的分析以及机器学习预测[J].电子技术与软件工程,2019(17):150-151.
- [2]王杨,阮妹.基于 CRITIC 赋权法的全球企业创新能力分析[J].华东科技,2019(12):56-63.
- [3]马力,刘丽涛.万有引力搜索算法的分析与改进[J].微电子学与计算机,2015,32(09):76-80.
- [4]徐宁,陈春,刘权.基于多模型的上海对长三角经济圈影响力计量分析[J].河北北方学院学报(自然科学版),2017,33(09):44-49.
- [5]郑健. 基于用户与项目双重聚类的协同过滤算法的研究与分析[D].北京工业大学,2017.
- [6]任雪娇.基于主成分分析法的挖掘机工作类型权重的计算[J].建设机械技术与管理,2016,29(09):78-81.

## 九、附录

### 附录一：

```
import pandas as pd
cc1=pd.read_csv("huashubei.csv",encoding="gbk")

cc1.isnull().count()
cc1.drop_duplicates().count()

cc1.describe()

cc1["2015-2020 sr"]=cc1["2020 SR"]-cc1["2015 SR"]
cc1["2015-2020 CY"]=cc1["2020 CY"]-cc1["2015 CY"]
cc1["2015-2020 HJ"]=cc1["2020 HJ"]-cc1["2015 HJ"]
cc1["2015-2020 WJ"]=cc1["2020 WJ"]-cc1["2015 WJ"]
cc1["2015-2020 总分"]=cc1["2020 总分"]-cc1["2015 总分"]
cc1

#相关性分析
new_data = cc1[["2015 SR","2015 HJ","2015 WJ","2015 SS","2015 总分"]]
import numpy as np
cc5=new_data.corr()
cc5.to_csv("xiangguanxin.csv")
cc5
new_data1 = cc1[["2020 SR","2020 HJ","2020 WJ","2020 SS","2020 总分"]]
import numpy as np
cc6=new_data1.corr()
cc6.to_csv("xiangguanxin1.csv")
cc6

new_data1 = cc1[["2015 SR","2015 HJ","2015 WJ","2015 SS","2015 总分","2020 SR","2020
HJ","2020 WJ","2020 SS","2020 总分"]]
import numpy as np
cc7=new_data1.corr()
cc7.to_csv("xiangguanxin1.csv")
```

### 附件二：

```
#找权重
import numpy as np
import pandas as pd

def critic(X):
    n,m = X.shape
    Z = standard(X)# 标准化 X，去量纲
```

```

R = np.array(pd.DataFrame(Z).corr())
delta = np.zeros(m)
c = np.zeros(m)
for j in range(m):
    delta[j] = Z[:,j].std()
    c[j] = R.shape[0] - R[:,j].sum()
C = delta * c
print(C)
w = C/sum(C)
return np.round(w,4)

def min_best(X):
    for i in range(len(X)):
        X[i] = max(X)-X[i]
    return X

def standard(X):
    xmin = X.min(axis=0)
    xmax = X.max(axis=0)
    xmaxmin = xmax-xmin
    n, m = X.shape
    for i in range(n):
        for j in range(m):
            X[i,j] = (X[i,j]-xmin[j])/xmaxmin[j]

    return X

if __name__ == '__main__':
    tt=cc1[["2015 SR","2015 CY","2015 HJ","2015 WJ","2015 SS"]]
    X=tt.as_matrix()

    print(critic(X))

```