所属类别	2020 年“华数杯”全国大学生数学建模竞赛	参赛编号

基于数据分析与统计学方法解决脱贫帮扶绩效评价问题

摘要

2020 年是脱贫攻坚战的决胜之年，全面建成小康社会，是中国共产党对中国人民的庄严承诺。自 2015 年以来，我国脱贫攻坚工作取得了卓越的成绩，各级扶贫单位在脱贫攻坚岗位上取得了不可磨灭的政绩。然而，在评价扶贫情况时，评价标准不够公平，评判指标不够综合等问题接踵而至，这些问题最终将影响到脱贫攻坚战的部署和扶贫工作的质量和效率。基于以上问题，本文从数据分析的角度入手，利用统计学方法，对相关扶贫单位和村庄在不同年份的数据和指标进行建模，对相关问题提出了如下解决方案：

针对问题一，本文将 2015 年和 2020 年某一村庄的某一单项指标数据组合成为一个二维点，通过对每一个村庄的扶贫情况进行如上处理，得到一个二维点图。利用最小二乘拟合，得到了各个指标的线性方程，通过观察系数则可以直接得到该指标数据在 2015 年和 2020 年的相关性情况，发现相关性均大于 0.5，进而判断出 2015 年的 2020 年对应评分存在相关性，符合现实发展规律。不仅如此，本文利用上述方法，对五项指标分别拟合得到的方程两两之间进行相关性分析，总结出各个指标之内，均存在一定关联和规律的结论。

针对问题二，本文考虑对某一帮扶单位及某一类型的所有帮扶单位，在两年的五项评价指标分别进行特征提取，利用主成分分析（PCA）方法进行数据投影，得到其投影后的降维数据和因子得分情况，并对两组降维后的一维数据作差，分别拟合出不同的帮扶单位个体及不同类型的帮扶单位在总体绩效上的进步幅度，进而对进步幅度和评分进行加权求和，得到其最终的绩效排序。

针对问题三，本文对各帮扶单位中，各个指标组合过后的二维数据进行 K-means 聚类分析，得到其对应指标的聚类中心点，并对应其坐标，得到该单位在 2015 年及 2020 年该指标数据的新的投影坐标，利用对两坐标值作差的方法得到其当前指标的帮扶业绩，并对其进行排序，进而分析出各帮扶单位排名。

针对问题四，本文考虑构建神经网络进行数据预测，对神经网络进行有监督训练。由于该数据集中的数据均为浮点数，数据精密，且标准化之后的数据差距较小，因此本文采用高斯-伯努利受限玻尔兹曼机（GBRBM）对一个三层神经网络进行分层初始化，再采用反向传播算法对网络进行微调，得到一个训练好的神经网络。基于以上原理，本文采用各指标数据和问题二得到的各村庄绩效训练新的神经网络，利用神经网络中的两个权值矩阵，将二者相乘，得到所有指标对绩效评分的影响权重。同时，作为一种数据预测模型，我们将最后 10 个村庄的数据输入神经网络中，将输出的数据作为其 2020 年的五项指标，并利用之前标准化的矩阵迹作为各指标权值，对所有指标进行加权求和，进而得到各村庄对应总分，混入所有村庄中依次排列，按照 1：3 的比例选取一级村庄。

最后，针对问题五，我们总结了对五项指标的分析结论，对脱贫攻坚和扶贫工作提出建议，并向国家扶贫办写了一封信，希望以数据科学的力量，助力脱贫攻坚。

关键词：绩效评价 统计学方法 主成分分析 最小二乘拟合 聚类分析 神经网络

一、 概述

1. 背景描述

贫困问题始终是困扰世界各国发展的难题，也是我国如期建成全面小康社会的关键。党的十八大以来，以习近平同志为核心的党中央把扶贫开发工作纳入“五位一体”总体布局和“四个全面”战略布局，全面打响了脱贫攻坚战，农村贫困人口大幅减少，区域性整体减贫成效明显，贫困群众生活水平大幅提高，贫困地区面貌明显改善，我国脱贫攻坚取得了决定性进展，走出了一条中国特色扶贫开发道路，取得了举世瞩目的扶贫成就，创造了人类减贫史上的中国奇迹。当前，脱贫攻坚已经进入决胜阶段，对每一个贫困户的进步考核方法也成为了当前亟待解决的一项重要问题，建立以扶贫成效为导向的考核机制，能够引导各村各单位在精准施策上出实招，在精准推进上下实功，在精准落地上见成效。有鉴于此，对扶贫工作中的业绩进行有效分析已经成为了摆在领导干部面前的一项难题。

2. 问题分析

本题旨在利用数学建模的方法，基于各村庄在 2015 年和 2020 年五个指标数据（居民收入、产业发展、居民环境、文化教育、基础设施），以及评估总分，对脱贫攻坚过程中的扶贫工作评估和绩效考核问题提出考核方案。

问题一要求对单一指标寻找其中是否存在时间层面上的关联和规律，并分析五个评价指标之间是否存在对应关系。

问题二要求对脱贫帮扶的帮扶单位和不同类型的所有扶贫单位的绩效进行评估，并对不同单位绩效进行评估排序。

问题三要求从单一指标的角度，利用有效的数学模型，对各个指标层面上表现最佳的单位进行评估，并将对应的帮扶单位列举排序。

问题四要求对影响“脱贫先进村庄”荣誉评选的因素进行评估，并通过数学建模，对一些数据被删除的贫困村庄的数据进行评估，以判定其是否能够获得此项荣誉。

问题五则要求利用以上问题得到的研究结果，向国家扶贫办以书信的形式阐述自己的观点和建议，向国家提出自己的扶贫帮扶策略。

二、 模型假设与符号说明

1. 模型假设

(1) 所有标准化之后的数据之间仍可直接进行运算，即不受原始数据分布的均值和标准差影响；

(2) 假定村庄脱贫评价仅考虑居民收入、产业发展、居住环境、文化教育和基础设施五项指标，且相对于任何其他因素独立；

(3) 假定数据集内所有数据均真实、有效。具有统计学意义。

2. 符号说明

符号	描述
r	线性相关系数
$\boldsymbol{H_{SR}}$	居民收入指标
$\boldsymbol{H_{FZ}}$	产业发展指标
$\boldsymbol{H_{HJ}}$	居民环境指标
$\boldsymbol{H_{JY}}$	文化教育指标
$\boldsymbol{H_{SS}}$	基础设施指标
2015/2020 SR	2015/2020 年居民收入
2015/2020 CY	2015/2020 年产业发展
2015/2020 HJ	2015/2020 年居民环境
2015/2020 WJ	2015/2020 年文化教育
2015/2020 SS	2015/2020 年基础设施
2015/2020 ZF	2015/2020 年总分
$\boldsymbol{\alpha}$	公式(12)中的系数矩阵，包含α_1，$1-\alpha_1$
$\boldsymbol{e}$	熵值

三、 建立成对数据相关性及线性回归模型进行指标分析

1. 模型建立

1.1 单指标成对数据相关性判断

该部分所建模型，旨在探索 2015 年与 2020 年的某项指标是否存在某种关联和规律，以及各项评价指标之间是否存在对应关系。从统计学的角度分析，考察两项指标的关联情况可以考虑采用相关性分析[1]，对这些指标进行检验，以确认其间是否存在规律。因此，如何对同一指标在两年之间进行相关性分析成为了一项亟待解决的问题。

有鉴于此，本文考虑采用图像和函数拟合的方式进行相关性表征。假定对指标H进行分析，其在 2015 年的所有村庄的数据表示为向量$\boldsymbol{H^{15}}$，在 2020 年的所有村庄的数据表示为向量$\boldsymbol{H^{20}}$。对于村庄s，本文表征其指标H在 2015 年和 2020 年的数据点C为：

$$C_s = (H_s^{15}, H_s^{20}), \tag{1}$$

由此，数据集中所有村庄的数据点可构成一个坐标矩阵$\boldsymbol{C}$：

$$\boldsymbol{C} = \begin{pmatrix} H_1^{15} & H_1^{20} \\ \vdots & \vdots \\ H_n^{20} & H_n^{20} \end{pmatrix}, \tag{2}$$

其中n为数据集中所有村庄的数量，为 32165。通过这种方式，所有指标均可以根据给定数据集构建一个类似于(2)的矩阵，最终可构建出五个坐标矩阵。通过这五个坐标矩阵，可以对五个指标的数据分布情况进行绘图，得到各个指标的数据分布情况。

然而，仅仅通过画图观察其数据分布情况并不能观察出某一指标的关联和规律，更不能获得其相关性。为了获取各项指标的深层规律，模型必须对其分布的性质进行更加深刻的分析。

本文试引入线性相关系数的概念：用来度量两个变量x和y的线性相关关系的标量称作线性相关系数。在线性方程中，通常构建函数如下：

$$y = rx + b, \tag{3}$$

在本文中，r可被视作其线性相关系数。由线性函数性质可知，相关系数的值介于-1与$+1$之间，即$-1 \le r \le +1$。当$r > 0$时，表示两变量正相关，当$r < 0$时，表示两变量为负相关。当$|r| = 1$时，表示两变量为完全线性相关即函数关系。当$r = 1$时，称为完全正相关，而当$r = -1$时，称为完全负相关。当$r = 0$时，表示两变量间无线性相关关系。

基于以上定义，我们可以通过拟合某一指标分布的线性函数，观察其系数，来确定某一指标在 2015 年和 2020 年的数据之间的相关性和规律。

关于线性方程的拟合问题，由于本问题中的所有点均为离散点，满足最小二乘法的条件，故本文拟采用最小二乘法对线性方程进行拟合。给定数据集包含n个观测数据点，数据点描述为(x_i, y_i)，假设线性方程为：

$$y = f(x, a_1, a_0) = a_1 x + a_0, \tag{4}$$

其中a_0和a_1作为待定参数。最小二乘法的拟合准则是使$y_i(i = 0,1,2,\dots,n)$与$f(t_i)$的距离的平方差最小，为了寻找函数$f(x, a_1, a_0)$的参数a_0和a_1的最有估计值，对于给定n组观测数据，对目标函数L进行求解：

$$\min_{a_1,a_0} L\big(y, f(x, a_1, a_0)\big) = \sum_{i=1}^{n} [y_i - f(x_i, a_1, a_0)]^2, \tag{5}$$

最终得到L取最小值时的a_1, a_0。利用此方法可以得到各个指标在两年数据之下的拟合线性方程，最终通过观察其参数a_1得到相关性情况。

1.2 线性回归模型对各指标对应关系判断

通过上一小节所叙述的模型构建方法，可以得到五个指标经过拟合得到的对应函数关系式。而接下来需要探究的，是各个指标之间的对应关系。在线性代数中，考虑线性函数相关性可以直接建立齐次线性函数组，可以根据其系数矩阵直接确定其为线性相关或线性无关。但问题在于，先前的函数拟合已经损失了原有分布的大量信息，通过拟合得到的函数判定相关性并不严谨。合理的做法是直接基于原有数据分布进行相关性分析。

关于函数相关性的确定，由于给定数据集已经进行了标准化，故在此不再对数据进行额外标准化和归一化。对于两个指标H，h，其相关系数可计算如下：

$$\rho_{H,h} = \frac{Cov(H, h)}{\sqrt{D(H)}\sqrt{D(h)}}, \tag{6}$$

其中Cov为两指标的协方差计算，D为方差计算。通过计算两两之间的相关系数，可以清晰的展现出两组数据分布的相关性。

2. 模型检验与结果分析

2.1 探究单指标评分关联性及其规律

通过对 2015 年和 2020 年的各项指标分别进行组合，本文得到了各年各指标的二维点分布。同时利用最小二乘法，对各个分布进行线性方程拟合，得到如图 1 所示的分布和直线方程。

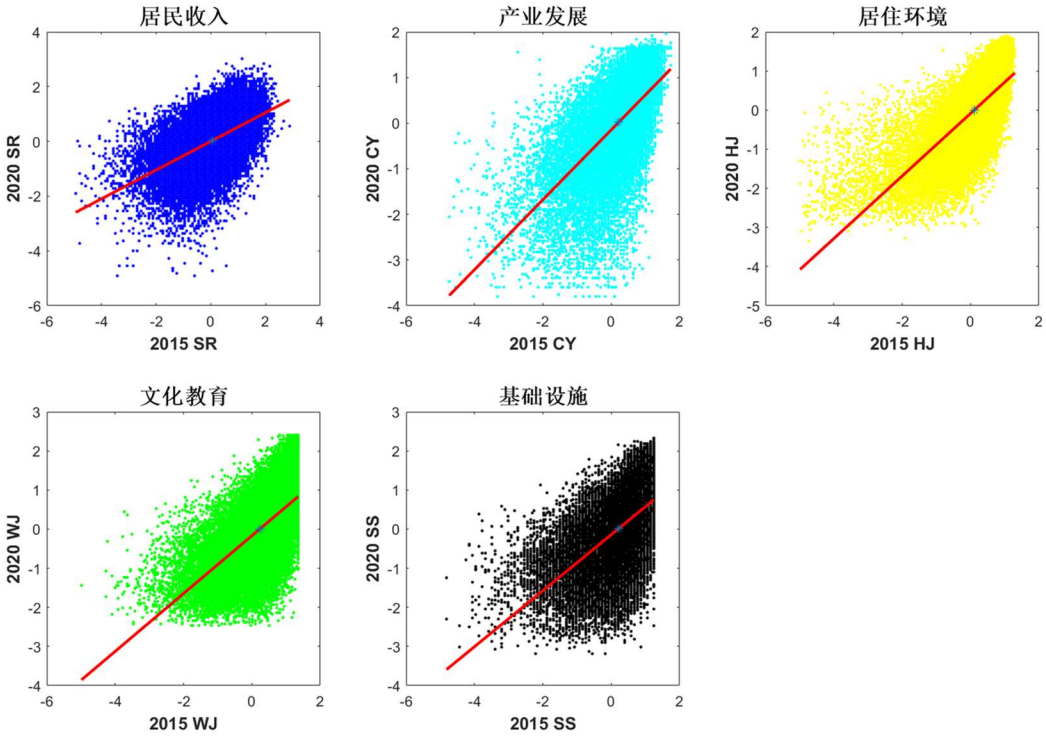

图 1 五项指标对应的二维点数据分布及拟合得到的线性方程

通过以上拟合，本文得到了五个指标相关性分析的线性方程如下：

$$H_{SR}^{20} = 0.5248H_{SR}^{15} + 0.0022,$$

$$H_{CY}^{20} = 0.7611H_{CY}^{15} - 0.1496,$$

$$H_{HJ}^{20} = 0.8H_{HJ}^{15} - 0.0844,$$

$$H_{WJ}^{20} = 0.7395H_{WJ}^{15} - 0.1681,$$

$$H_{SS}^{20} = 0.7160H_{SS}^{15} - 0.1412,$$

可见，所有方程的相关系数均大于 0，这也就意味着，对于任意一个指标，其 2015 年的数据与 2020 年的数据呈正相关，该组数据中存在题干所述规律。

2.2 探究各指标之间对应关系

接下来，本文将通过对数据进行相关性分析，探索五个指标之内是否存在对应关系。通过对五个指标在各年的数据进行相关性分析，确认各个变量之间是否存在关联。如表1，表2所示，本文分析了各年内所有指标之间的相关性。

表 1 2015 年内各指标数据之间相关性分析

指标	2015 SR	2015 CY	2015 HJ	2015 WJ	2015 SS
2015 SR	1.000	0.514	0.573	0.519	0.525
2015 CY	0.514	1.000	0.566	0.683	0.669
2015 HJ	0.573	0.566	1.000	0.562	0.585
2015 WJ	0.519	0.683	0.562	1.000	0.734
2015 SS	0.525	0.669	0.585	0.734	1.000

表 2 2020 年内各指标数据之间相关性分析

指标	2020 SR	2020 CY	2020 HJ	2020 WJ	2020 SS
2020 SR	1.000	0.600	0.648	0.557	0.569
2020 CY	0.600	1.000	0.685	0.758	0.738
2020 HJ	0.648	0.685	1.000	0.664	0.670
2020 WJ	0.557	0.758	0.664	1.000	0.758
2020 SS	0.569	0.738	0.670	0.758	1.000

根据表1，表2，所有的相关性均大于0.5，可知各因素之间存在相关性，且均为中度相关，通过以上分析可知，本模型给定的五个指标之间并非完全相互独立，两两之间均存在相关性。

四、 建立绩效模型进行脱贫工作评价

1. 模型建立

对于扶贫工作的绩效评价，只参考总分是不合理的，显然，将所有评价指标综合起来考量的模型，相较于单一参考总分，能够更加全面、客观、综合的反应扶贫工作进展和效果。这就要求建立一个能够将五项评价指标进行有效结合，并与总分进行统一的模型，来更好的对帮扶单位成绩进行有效评估[2]。

首先考虑五项指标的结合问题，由于绩效评价更倾向于考虑村庄的进步幅度，总分的进步幅度可以轻松的用两年的差值ΔQ表示，而此五项指标的进步幅度数据较为分散，

但却互有相关。对于这些并非各自独立的指标，本文首先采用主成分分析对各年的五项指标进行数据特征提取，得到因子得分最高的一维数据。

对于某一年的五项指标数据，可计算矩阵如下：

$$\boldsymbol{IN} = (\boldsymbol{H_{SR}} \quad \boldsymbol{H_{FZ}} \quad \boldsymbol{H_{HJ}} \quad \boldsymbol{H_{JY}} \quad \boldsymbol{H_{SS}}),$$

接着，对矩阵$\boldsymbol{IN}$进行处理，使得其每一列向量减去其对应的平均值，得到一个新的$\boldsymbol{IN}^*$矩阵，然后计算其协方差矩阵：

$$Cov = \frac{1}{n}\boldsymbol{IN}^*\boldsymbol{IN}^{*\boldsymbol{T}}, \tag{7}$$

得到一个5×5的协方差矩阵Cov。对于矩阵$\boldsymbol{IN}$，我们希望能够将他投影在一组新的基空间上，使之矩阵大小得到压缩，即：

$$\boldsymbol{D}_{n\times N} = \boldsymbol{IN}^*_{n\times 5} \cdot \boldsymbol{P}_{5\times N}, (N < 5), \tag{8}$$

其中$\boldsymbol{P}$为转换矩阵，$\boldsymbol{D}$转换后的矩阵，我们要做的就是将 5 个特征压缩为N个特征，对于压缩过的数据投影，其协方差和方差为：

$$Cov(\boldsymbol{D})_{NN} = \boldsymbol{D}^T\boldsymbol{D} = \boldsymbol{P}^T\boldsymbol{IN}^{*T}\boldsymbol{IN}^*\boldsymbol{P}. \tag{9}$$

接下来对转换矩阵$\boldsymbol{P}$进行求解，使得$Cov(\boldsymbol{D})$的对角线元素尽可能大，非对角线元素尽可能小。这种目的可以通过矩阵相似对角化和特征向量实现：

$$\boldsymbol{P}^T\left(\boldsymbol{IN}^{*T}\boldsymbol{IN}^*\right)\boldsymbol{P} = \boldsymbol{P}^{-1}\left(\boldsymbol{IN}^{*T}\boldsymbol{IN}^*\right)\boldsymbol{P} = \boldsymbol{\Lambda}, \tag{10}$$

其中$\boldsymbol{\Lambda}$为$\boldsymbol{IN}^{*T}\boldsymbol{IN}^*$的特征值组成的对角阵。最终：

$$\boldsymbol{D}_{n\times N} = \boldsymbol{IN}^*_{n\times 5} \cdot \boldsymbol{P}_{5\times N} \tag{11}$$

得到的矩阵$\boldsymbol{D}$即为最终需要的降维矩阵。

接着对此两年的一维数据作差，得到其主要进步幅度$\boldsymbol{\Delta T}$。对于每个村中最终的绩效G，其可利用加权求和进行求解如下：

$$G = \alpha_1 \cdot \boldsymbol{\Delta T} + (1 - \alpha_1) \cdot \boldsymbol{\Delta Q}, \tag{12}$$

对于其中的权值α_1，本文采用熵权法，通过测定原始数据作差后的分布熵值，来确定权重。对于本例而言，由于熵值法的计算过程要求数值中不能存在 0 或负数，要先对数据集当中标准化过后的数据进行非负平移后，再进行处理。数据集中有n个样本和 2 个指标，定义H_{ij}为第i个样本的第j个指标的数值。首先计算第j项指标下第i个样本值占该指标的比重：

$$p_{ij} = \frac{H_{ij}}{\sum_{i=1}^{n} Y_{ij}}, \tag{13}$$

再计算第j项指标的熵值：

$$e_j = -k \sum_{i=1}^{n} p_{ij} \ln(p_{ij}) \tag{14}$$

其中：$k = \frac{1}{\ln n} > 0$。而后计算信息熵冗余度：$d_j = 1 - e_j$。此后，可以计算各项指标的权重如下：

$$\alpha_j = \frac{d_{ij}}{\sum_{i=1}^{n} d_{ij}}. \tag{15}$$

将(15)得到的权值带回到(12)中，得到每个村庄的得分情况，再对每个帮扶单位和每个类型的帮扶单位管辖的村庄进行划分，分块计算其得分平均数，并进行排列，即可区分出各个帮扶单位的绩效情况。

2. 模型检验与结果分析

首先，本文利用上述熵权法，对得分公式的权值进行计算，得到数据如下：

$$\alpha_1 = 0.4649, 1 - \alpha_1 = 0.5351$$

将以上权值代入(12)式，对所有帮扶单位管辖下的数据进行得分计算，并对各个帮扶单位下的所有村庄的评分G求平均值，得到排名前十的帮扶单位如下：

表 3 帮扶工作绩效评分排名前十的单位编号及其得分

帮扶单位	149	113	158	153	115
得分 G	1.3924	1.3574	1.2711	1.1667	0.9932
帮扶单位	131	155	54	132	45
得分 G	0.8753	0.7928	0.7777	0.7295	0.7178

利用同样的算法，对不同类型的所有帮扶单位下的村庄得分分别进行平均，得到各个类型的得分如下：

表 4 帮扶工作绩效评分的单位类型编号及其得分

单位类型	4	5	0	2	3	1
得分 G	0.4968	0.1076	−0.0397	−0.0721	−0.1384	−0.1604

通过以上两个表格可知，类型 4 的帮扶单位，在脱贫攻坚中成果显著，具有较高的绩效，并且，在全国 160 个帮扶单位中，编号为 149、113、158、115、131、155、54、132 和 45 的帮扶单位具有较好的绩效，排名在全国前十。

五、 建立二维聚类评价模型解决帮扶业绩评估问题

1. 模型建立

不同于上一个模型，该问题要求模型针对不同帮扶单位的各项指标分别进行业绩分析，并对其帮扶业绩做出评估和排序，因此，如何针对某一帮扶单位同一指标下的所有数据进行特征提取，成为了建立模型的主要方向。本文采用第一问的数据构建方法，将 2015 年和 2020 年同一指标下的数据两两组合，构成新的二维数据分布，并利用 K-means 聚类分析方法，对各个帮扶单位下的所有村庄情况进行聚类分析，得到各个指标下所有帮扶单位的聚类中心点，通过反向对应数据点，得到其大致的 2015 年和 2020 年的总体指标数据，通过作差的形式，得到其对应的帮扶业绩，并依据帮扶业绩情况进行排序，得出前五名帮扶单位编号 [3][4]。

首先，利用(1)(2)式构建二维分布，对于某一编号为m的帮扶单位而言，其所帮扶的所有村庄的数据点均为一个观测数据，其帮扶单位m可视作一个类别，而对于其他所有帮扶单位而言，其对应帮扶单位编号也同样可被视作一个类别。不妨假设需要将数据$C_s = (H_s^{15}, H_s^{20})$聚为$k$类，而这$k$个聚类的中心点为$\mu_s$，由此引入其损失函数：

$$\min L = \sum_{j=1}^{k}\sum_{s=1}^{n}(C_s - \mu_s)^2 1_{\{t_s=j\}}, \tag{16}$$

而聚类的目的则是寻找最佳的t_s，使得损失函数L最小，之后就可以对中心点μ_s=(μ_s^{15}, μ_s^{20})逐个求解了。在判断每个点所属质心的过程中，本文采用了欧几里得距离R进行计算。

在聚类分析结束，得到了所有聚类中心点坐标矩阵后，可以从二维平面上对应得到该帮扶单位两年之间在该指标上的数值，而这个数值是一个泛概念，意图对所有样本进行更加合理的“平均化”。尔后，本文采取了同样的方法，通过作差的方式得到了对应帮扶业绩A_s：

$$A_s = \mu_s^{20} - \mu_s^{15}. \tag{17}$$

由此，所有帮扶单位的帮扶业绩情况均可得到排序。

2. 模型检验与结果分析

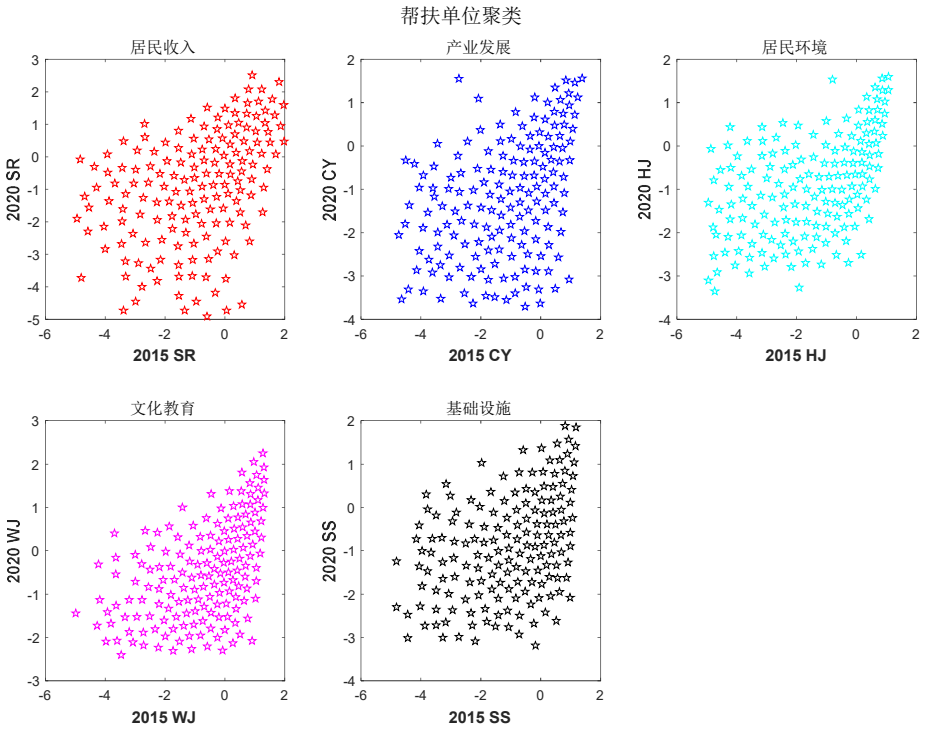

图 2 所有 160 个帮扶单位在五项指标上的聚类中心点分布

表 5 各指标排名前五的单位编号及其进步幅度

居民收入		产业发展		居民环境		文化教育		基础设施	
进步幅度差值	编号	进步幅度差值	编号	进步幅度差值	编号	进步幅度差值	编号	进步幅度差值	编号
4. 753	38	4. 287	151	4. 792	32	4. 092	14	4. 287	151
4. 054	25	4. 180	139	4. 657	139	3. 910	43	4. 180	139
4. 007	159	3. 803	23	4. 310	45	3. 543	65	3. 803	23
3. 867	155	3. 501	10	3. 987	37	3. 484	42	3. 501	10
3. 701	92	3. 242	131	3. 974	102	3. 127	148	3. 242	131

首先，如图 2 所示，本文对所有 160 个帮扶单位在五项指标上的聚类中心点进行了绘制。这些数据点的横纵坐标做差后能够直接看出其帮扶业绩的进步情况。对各个指标中所有的帮扶单位两年间的差值进行排序，得到了各指标排名前五的单位编号及其进步幅度，如表格 5 所示。

根据表 3 可知，在居民收入指标上，帮扶单位 38，25，159，155 和 92 表现最佳，在产业发展指标上，帮扶单位 151，138，23，10 和 131 表现最佳，在居民环境指标上，32，139，45，37 和 102 表现最佳，在文化教育指标上，帮扶单位 14，43，65，42 和 148 表现最佳，最后，在基础设施指标上，帮扶单位 151，139，23，10 和 131 的表现最佳，进步幅度较大。

六、 建立高斯-伯努利 RBM 神经网络模型进行数据预测

1. 模型建立

问题四要求对评选先进的重要影响因素进行分析，并对几个村庄的数据进行预测，以确定其是否能够评上荣誉称号。关于对重要影响因素的分析问题上，本文需要对数据的深层次特征进行提取，同时，更要利用已有的三万五千条数据，对给出的十个村庄的未知数据进行准确预测。显然，就本题而言，每一条数据都包含 2015 年和 2020 年的具体信息，属于有监督学习，在这种情况下进行数据预测，神经网络是不二之选。然而，常规的神经网络对于深层数据挖掘的效果十分有限，且深层神经网络的训练代价较高，对于本文来说从时间角度并不适用。

有鉴于此，如图 3 所示，本文作者有针对性的设计了一个常规的三层感知机作为神经网络预测模型[5]，并采用与感知机相同输入、相同隐层的受限玻尔兹曼机作为特征提取器，对数据深层次特征进行提取，并将特征提取后得到的权值矩阵移植到神经网络中进行初始化，最后利用反向传播算法（BP）进行第一层与第二层之间的微调和第二层和第三层之间的训练[6]，最终得到一个既可以提取深层特征，又可以进行数据预测的多层感知机。通过训练得到的输入指标与隐层神经元之间的权值矩阵W_{IH}，和隐层神经元到输出神经元的权值矩阵W_{HO}。令输入向量为$\boldsymbol{v}$，利用问题二中所涉及到的绩效评价模型获得其绩效G'，其计算公式如下：

$$G' = PCA(\boldsymbol{v} \times W_{IH} \times W_{HO})\boldsymbol{\alpha}^T,$$

其中PCA为对括号中的矩阵进行主成分分析操作，降维后取得分最高的一维数据，与(12)式当中的加权向量α相乘可得，最后得到所有村庄的得分。

对于如何筛选影响荣誉称号评选的重要因素的问题，本文将按照该模型建立方法，重新构建一个输入神经元为 10，输出神经元为 1 个的新的神经网络，通过将第二问的各村庄的绩效评分设置为期望输出，利用 2015 年和 2020 年的所有指标作为输入，训练神经网络后，将输入-隐层权值矩阵$\boldsymbol{W}_{IH}(10\times 5)$与隐层-输出神经元权值矩阵$\boldsymbol{W}_{HO}(5\times 1)$相乘，得到新的矩阵$\boldsymbol{W}^{*}$观察各个权值分布情况，其计算公式如下：

$$\boldsymbol{W}^{*}=\boldsymbol{W}_{IH}\boldsymbol{W}_{HO}$$

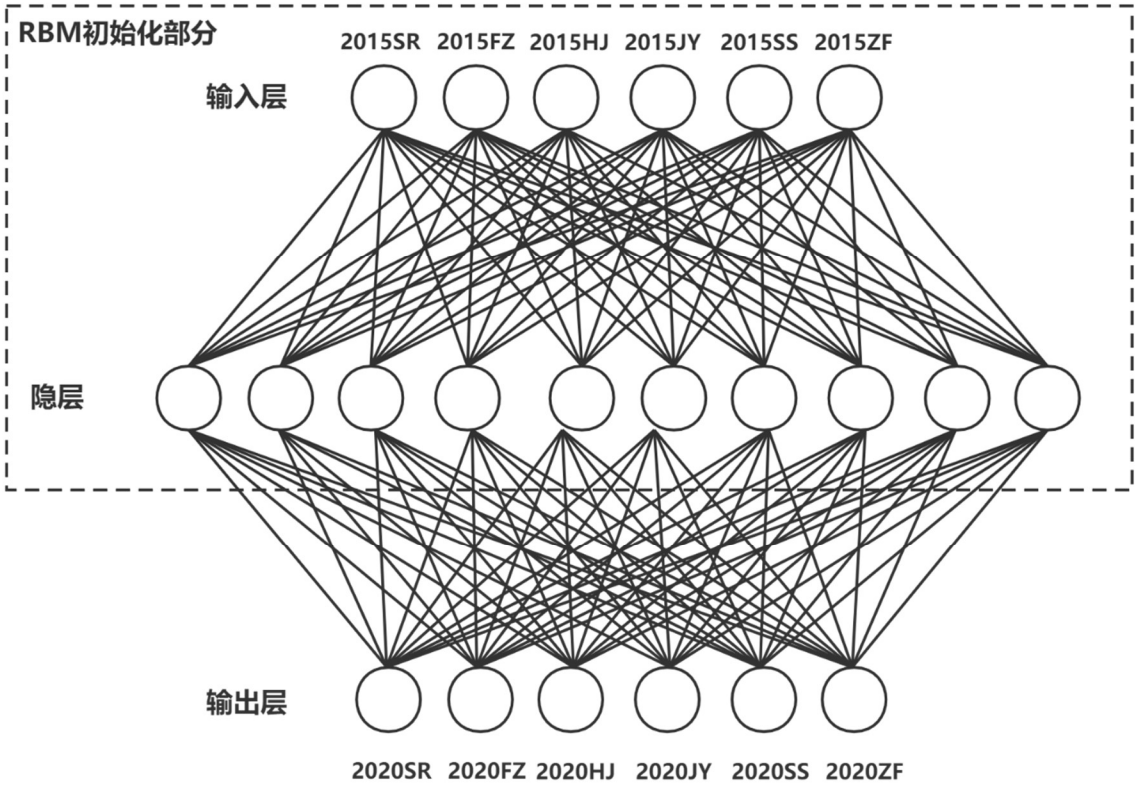

图 3 高斯-伯努利 RBM 神经网络模型

值得一提的是，通过观察数据发现，本数据集当中的数据多为离散的浮点数，不适用于传统的二值受限玻尔兹曼机模型，因此，本文对该模型进行改进，将原有隐层后验概率分布——伯努利分布改为高斯分布，称为高斯-伯努利受限玻尔兹曼机[7]，适应了数据特征。

尔后，本文将预测得到的十个村庄的数据，以第二题模型中的绩效评价方法进行判定，将其得分与全国所有村庄进行比较，排位在前 10000 名之内的，自然为脱贫先进村庄，尔后在其中，按照 1：3 的比例筛选等级。

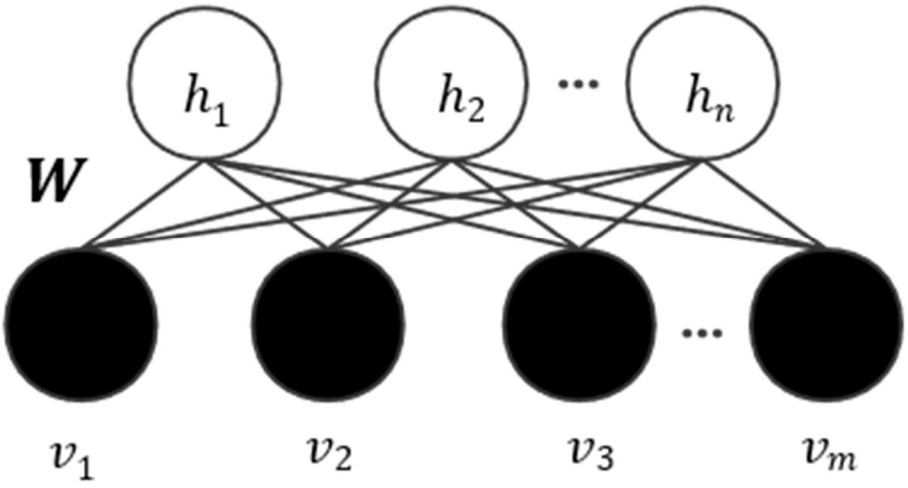

图 4 受限玻尔兹曼机的概率图模型

如图 4 所示，受限玻尔兹曼机（RBM）[8]是一种二层的随机生成模型，通过对数据的玻尔兹曼分布进行拟合，当达到热平衡时达到全局最优解，能够有效解决常规神经网络易陷入局部最小的缺点，但是其二值特性使得其应用受到了很大限制，于是，一种适应于浮点数据的高斯-伯努利受限玻尔兹曼机应运而生。其可被简化为一个概率图模型。对于本题而言，此次输入为五个指标和一个总分，共有 6 维，隐层数量为n，$\boldsymbol{v}=(v_1,v_2,\dots,v_m)^T$为输入状态向量，$\boldsymbol{h}=(h_1,h_2,\dots,h_n)^T$为隐层状态向量，其权值矩阵$\boldsymbol{W}=[w_j](i=1,\dots,n;j=1,\dots,m)$。$\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_m)^T$为输入层偏置，$\boldsymbol{\beta}=(\beta_1,\dots,\beta_n)^T$为隐层偏置。对于该模型而言，其能量函数可被描述为：

$$E(\boldsymbol{v},\boldsymbol{h})=\sum_{j=1}^{m}\frac{(v_j-\alpha_j)^2}{2\sigma_j^2}-\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}h_i\frac{v_j}{\sigma_j^2}-\sum_{i=1}^{n}\beta_i h_i\,, \tag{18}$$

其联合概率分布为：

$$P(\boldsymbol{v},\boldsymbol{h})=\frac{1}{Z}\exp\big(-E(\boldsymbol{h},\boldsymbol{v})\big), \tag{19}$$

其中，$Z=\sum_{v,h}\exp\big(-E(\boldsymbol{v},\boldsymbol{h})\big)$，为配分函数。

其边缘概率分布为：

$$P(\boldsymbol{v})=\frac{1}{Z}\exp\left(\boldsymbol{\alpha}^T\boldsymbol{v}+\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}softplus(\boldsymbol{w_i}\boldsymbol{v}+\beta_i)\right), \tag{20}$$

$$P(\boldsymbol{h})=\frac{1}{Z}\exp\left(\boldsymbol{\beta}^T\boldsymbol{h}+\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}}softplus\big(\boldsymbol{w_j}\boldsymbol{h}+\alpha_j\big)\right), \tag{21}$$

其中，$\boldsymbol{w_i}$是权值矩阵$\boldsymbol{W}$的行向量，$\boldsymbol{w_j}$是权值矩阵$\boldsymbol{W}$的列向量，$softplus = \log(1 + e^x)$。

其后验概率分布为：

$$P(h_i = 1|\boldsymbol{v}) = sig\left(\beta_i + \sum_{j=1}^{m} w_{ij} v_j\right), \tag{22}$$

$$P(v_j = 1|\boldsymbol{h}) = sig\left(\alpha_j + \sum_{i=1}^{n} w_{ij} h_i\right), \tag{23}$$

其先验概率分布为：

$$P(v_j = 1|\boldsymbol{h}) = \mathcal{N}(v_j|\alpha_j + \textstyle\sum_{i=1}^{n} w_{ij} h_i, \sigma_j^2) \tag{24}$$

$$P(h_i = 1|\boldsymbol{v}) = sig\left(\beta_i + \sum_{j=1}^{m} w_{ij} v_j\right), \tag{25}$$

其中，$\mathcal{N}(\cdot|\mu, \sigma^2)$为高斯分布，$\mu$为均值，$\sigma$为标准差，而$sig$为 sigmoid 逻辑斯蒂方程。

在本模型中，高斯-伯努利 RBM 被用来初始化如图所示的多层感知机的前两层，最后一层作为输出层，由反向传播算法进行直接训练。基于本例的实际情况，我们设计了一个$6 \times hidden \times 6$的三层神经网络模型，网络选用 S 型传递函数：

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{26}$$

通过误差反传函数（t_i为期望输出，O_i为输出）：

$$\min E = \frac{\sum_i^n (t_i + O_i)^2}{2}, \tag{27}$$

和梯度下降法，不断调节由高斯-伯努利 RBM 初始化的网络权值，使得误差函数E达到极小。

2. 模型检验与结果分析

首先，考虑 2015 年的五项指标和总分数据为训练集数据，2020 年的五项指标和总分数据作为目标数据，利用高斯-伯努利受限玻尔兹曼机模型，对输入层和隐层之间的权值矩阵进行初步训练，再建立神经网络模型，利用反向传播算法对初始化后的多层感知机进行训练。本文采用了输入层*隐层*输出层=$6 \times 12 \times 6$的网络结构，学习率设为 0.2，收敛目标值设为10^{-5}，预设迭代次数为 5000 次，训练过程中的误差如图 5 所示，梯度变化情况如图 6 所示。

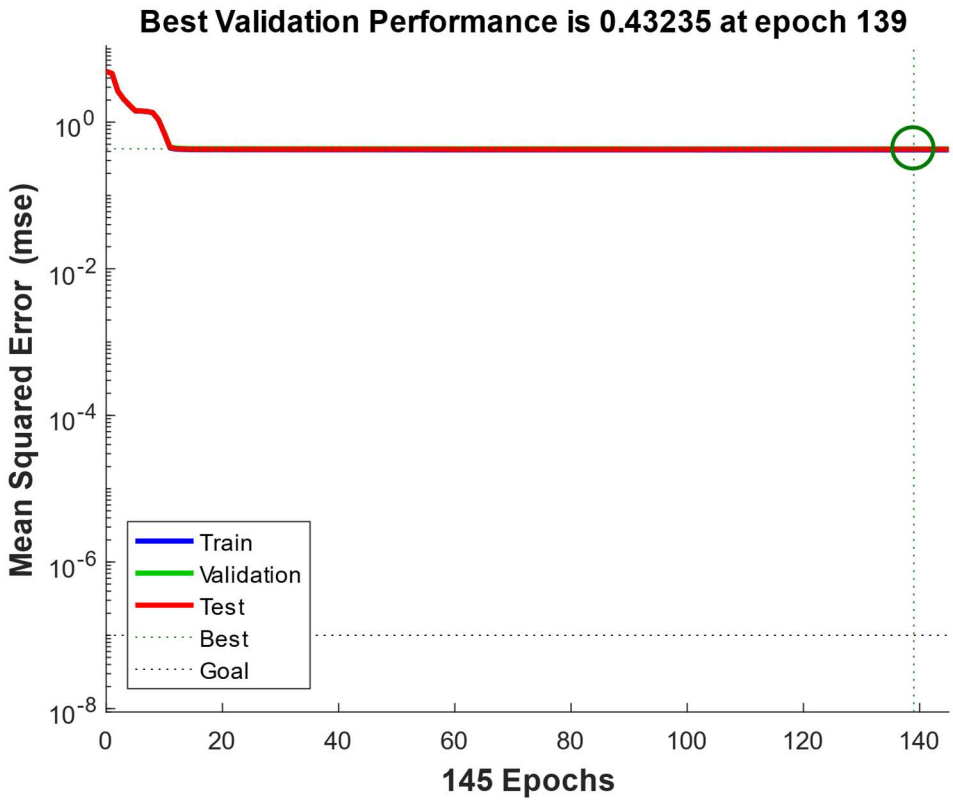

图 5 神经网络训练过程中的误差变化情况

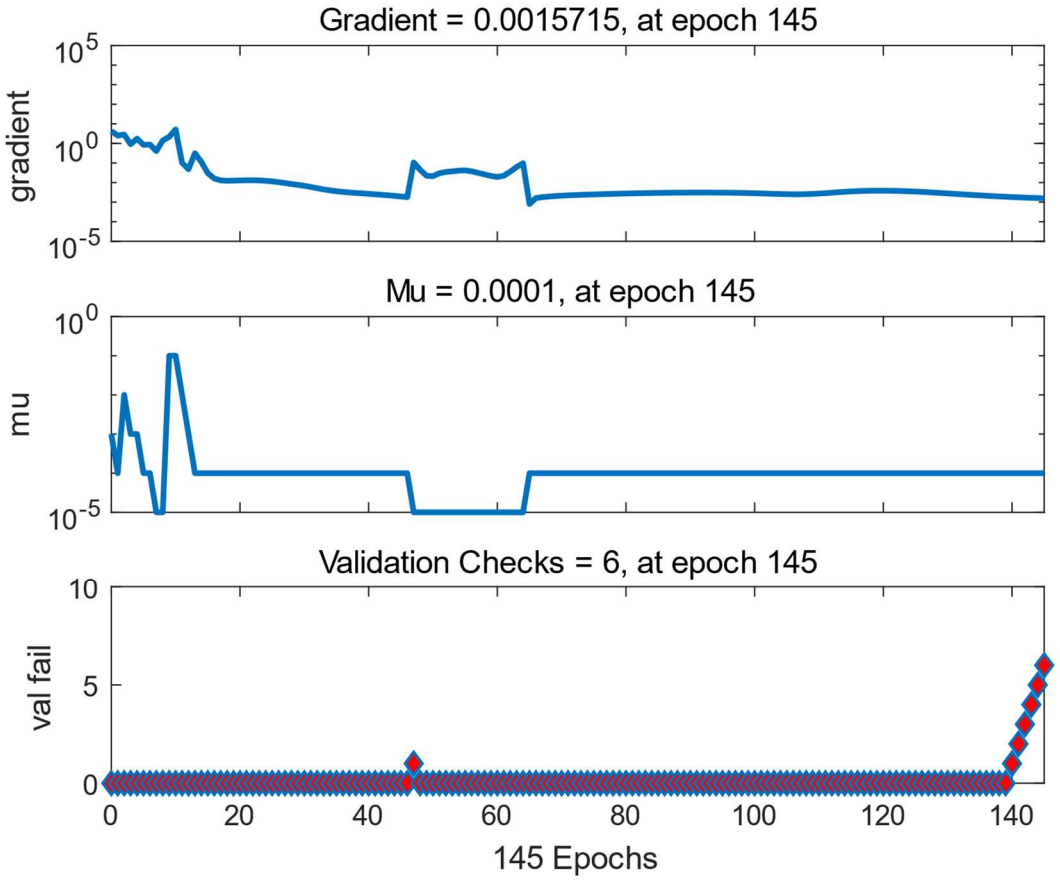

图 6 神经网络训练过程中的梯度变化情况

神经网络训练完成后，将数据被删除的十个村庄在2015年的数据输入神经网络中，得到预测数据如下：

表格 6 数据被删除的十个村庄在2020年的预测数据

指标\村庄编号	12722	12916	21570	22096	34208	34276	25149	39257	42883	12722
2020 SR	0.822	1.203	0.819	0.933	0.776	0.065	0.528	0.372	0.654	−0.212
2020 CY	0.641	1.216	0.920	0.986	0.857	0.456	0.293	0.182	0.782	−0.276
2020 HJ	0.953	1.339	0.992	1.018	1.168	−0.123	0.769	0.637	0.938	−0.192
2020 WJ	0.512	1.420	1.031	1.076	0.938	0.496	0.068	−0.114	0.744	−0.278
2020 SS	0.598	1.447	1.000	1.115	0.912	0.341	0.193	−0.014	0.730	−0.328
2020 总分	0.789	1.514	1.111	1.184	1.088	0.338	0.380	0.184	0.899	−0.306

最后，将以上数据代入到公式(12)，得到十个被删除数据的村庄的预测得分和在全国过所有村庄的排名如表7所示：

表格 7 被删除数据的村庄的预测得分及全国排名

村庄编号	12722	12916	21570	22096	34208	34276	25149	39257	42883	12722
得分	0.490	1.543	0.939	1.053	0.899	−0.196	−0.082	−0.353	0.629	−1.081
全国排名	20427	**6504**	**9774**	**2573**	**3436**	26034	29798	28160	**8300**	20427

根据题目给定条件，排名前10000名的村庄可以获得荣誉称号，且在这10000个村庄内按照1：3的比例划分一等和二等，即2500：7500，因此，编号为12916、21570、22096、34208、42883的村庄将获得二级荣誉称号。由于这些村庄里没有进入前2500名的村庄，故这十个村庄均不能获得一等荣誉称号。

接下来，本文将对影响荣誉称号评定的因素进行评价。在这里，本文建立了一个结构为$10 \times 5 \times 1$的神经网络，学习率设为0.2，收敛目标值设为10^{-5}，预设迭代次数为5000次，训练过程中的误差如图7所示，梯度变化情况如图8所示。