调研世界
The World of Survey and Research
ISSN 1004-7794,CN 11-3705/C

《调研世界》网络首发论文

题目：基于机器学习聚类的无回答权数调整研究
作者：金勇进，刘晓宇
DOI：10.13778/j.cnki.11-3705/c.2020.10.002
网络首发日期：2020-10-13
引用格式：金勇进，刘晓宇．基于机器学习聚类的无回答权数调整研究[J/OL]．调研世界．https://doi.org/10.13778/j.cnki.11-3705/c.2020.10.002

基于机器学习聚类的无回答权数调整研究

金勇进　刘晓宇

内容摘要：在实际调查工作中，由于客观条件的限制，难以完全避免无回答情况的出现，需在数据分析阶段弥补无回答对估计产生的负面影响。本文尝试通过机器学习中的聚类算法进行无回答权数调整，以突破可忽略性的限制，着重考察在不可忽略的无回答机制下的估计效果。实证研究根据 2015 年中国综合社会调查（CGSS）数据进行，结果表明，无论无回答机制是否可忽略，基于机器学习聚类算法进行的权数调整，均能有效控制无回答偏差、得到变异性小的最终权数和性质优良的目标变量估计。

关键词：非随机缺失；不可忽略的无回答机制；权数调整；聚类分析

中图分类号：C81　　**文献标识码：**A　　**文章编号：**1004-7794(2020)10-0011-09

DOI: 10.13778/j.cnki.11-3705/c.2020.10.002

一、引言

调查中的无回答是指由于种种原因没有能够对被抽出的样本单元进行计量，从而没有获得有关这些单元的数据。由于客观条件的限制，在实际调查中难免会产生一定数量的无回答。无回答是造成缺失数据的基本原因，它不仅致使有效样本量减少、样本信息难以真实反映总体情况，还导致估计量有偏且方差增大，从而影响了统计数据质量。

当出现无回答情况时，通常采用插补法和加权法来弥补无回答偏差。插补法利用观察数据构造预测分布，根据插补值构造完整数据集，从而进行统计推断。加权方法以某种方式将缺失单元的权数分解到观测值上，增加观测数据的权数，用有回答样本代表无回答样本，从而减小潜在的无回答偏差。加权法对权数的调整效果取决于对回答概率的估计优劣，对回答概率的估计越准确，调整效果越好。然而，这两种方法以随机缺失为前提，实际中遇到的缺失多属于非随机缺失，即不可忽略的无回答机制，此时目标变量 y 的作答情况与 y 的具体值有关，贸然使用基于随机缺失假定的处理方法可能带来极大的估计风险[1-2]。目前，针对不可忽略机制下的无回答的研究还很有限，主要基于模型进行，但由于其局限性难以在实际中应用，其不足表现在以下三个方面：一是这类方法属于参数方法，需在较强的模型假设下建模，且对模型的错误识别非常敏感；二是由于数据的缺失，无法评估模型的好坏、验证模型是否适用；三是由于这些方法依赖迭代算法进行估计，考虑到矩阵计算中的可逆性、估计量的收敛性等问题，算法通常较复杂且计算成本较大。此外，单元的作答情况往往受各种因素的影响，识别并量化其影响因素并不容易，若判断错误可能会对估计造成负面影响。

基于上述问题，本文试图去除对无回答可忽略性的要求，采用基于机器学习的聚类分析算法调整无回答权数，通过聚类将样本分为不同类，类内的回答样本可在一定程度上代表无回答样本，由此可分类进行权数调整。本文回顾了无回答情况下的权数调整和基于聚类算法的权数调整，并采用 2015 年中国综合社会调查（CGSS）

数据，设置不同的无回答机制模式、聚类算法和权数调整方法，对基于机器学习聚类算法的无回答权数调整效果进行实证研究。

二、无回答情况下的权数调整

由于无回答情况的存在，无回答单元的权数缺失，若忽略无回答单元，仅基于回答者进行推断，会导致对总体规模的低估，进而导致估计的偏差[2]。校准法和倾向得分调整法是弥补回答误差的常用方法。

（一）校准法

校准加权最早由 Deville et al（1992）提出，是一种系统利用辅助信息对权数进行调整的方法[3]。假设已知辅助变量的总体总值 T_x，并通过调查得到了样本单元的辅助变量 $x_i(i\in S)$。用设计权数 $d_i=1/\pi_i$ 对辅助变量加权估计时，可能会出现 $\sum_{i\in S}d_i x_i \neq T_x$ 的情况。校准法由此产生：根据辅助变量总体总值，寻找使得 $\sum_{i\in S}w_i x_i=\sum_{i\in U}x_i$ 成立的权数 w_i。该约束条件称为校准方程，满足校准方程的权数称为校准权数，总体总值的加权估计量 $\hat{Y}_{CAL}=\sum_{i\in S}w_i y_i$ 称为校准估计量。Deville et al（1992）提出最小化损失函数的方法求解最优权数，不同的损失函数，对应不同的校准估计方法。Zieschang（1990）将线性超总体模型与校准估计相结合，提出了广义回归加权法[4]。该方法假定目标变量与辅助变量之间存在线性关系，采用最小平方距离，在辅助变量 X 的约束下最小化目标函数 $\sum_{i\in S}(w_i-d_i)^2/d_i$，得到总体总值的估计为：

$$\hat{Y}_{GREG}=\sum_{i\in S}w_i y_i=\hat{Y}_{HT}+(T_x-\hat{X}_{HT})^T\hat{\beta}$$

其中，$\hat{X}_{HT}=\sum_{i\in S}x_i/\pi_i$ 是辅助变量总体总值的 HT 估计量；T_x 是辅助变量的总体总值；$\hat{\beta}=\left(\sum_{i\in S}d_i x_i x_i^T\right)^{-1}\sum_{i\in S}d_i x_i y_i$ 是回归系数的最小二乘估计。广义回归加权法得到调整后的权数为：

$$w_i=d_i\left[1+x_i^T\left(\sum_{i\in S}d_i x_i x_i^T\right)^{-1}(T_x-\hat{X}_{HT})\right]$$

校准权数实现了样本结构对总体结构的还原，得到的估计量具有渐进无偏性和设计一致性[5]。Lundström et al（1999）首先将校准估计运用于无回答问题[6]。已有研究表明，校准法不仅可以减小估计标准误，还可以校正调查样本无回答、抽样框覆盖不全或重复而导致的误差。

（二）倾向得分调整法

样本匹配最早基于观察性研究提出，是一种根据背景变量找到与处理组相似的对照组再进行分析的方法，可以减少处理效应所造成的估计偏差。利用倾向得分法进行无回答权数调整，即找到与未回答单元相似的回答单元，实现回答单元对未回答单元的代表作用。用于无回答权数调整的倾向得分，可定义为在给定背景变量 X_i 的条件下，个体 i 无回答的条件概率，记作 $p(X_i)=p(D_i=1|X_i)$。匹配的作用在于，找到与无回答单元 i 相似的回答单元 j，使得单元 i 和单元 j 的倾向得分尽可能相似，即 $p(X_i)\approx p(X_j)$。倾向得分调整一般需要三个步骤。

1.估计倾向得分。

估计倾向得分，即对 $p(X_i)=p(D_i=1|X_i)$ 进行估计，可采用参数估计或非参数估计方法。常用的参数估计方法有 Logistic 回归和 Probit 回归等，对背景变量 X 和示性变量 D 的关系进行建模，以 Logistic 回归为例，构建模型 $\log\left(\frac{p(X_i)}{1-p(X_i)}\right)=X_i\beta$，可得到估计值 $\hat{p}(X_i)=\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$。

2.进行样本匹配。

为将无回答单元与回答单元匹配，需根据 $\hat{p}(X_i)$ 评价单元间的相似程度，评价方法有很多，从而产生不同的匹配方法，如最近邻匹配、分层匹配和局部线性匹配等。例如，单一有放回最近邻匹配是指根据 $\hat{p}(X_i)$，从回答单元中寻找与无回答单元最接近的单元，并与之匹配，回答单元可重复使用，一个无回答单元匹配一个回答单元，但一个回答单元可以匹配到多个无回答单元。

3.加权调整。

倾向得分可与不同的权数调整方法相结合，

形成不同的倾向得分调整法。无回答样本集合记作 S_{nr}，最终的匹配样本集合记作 S_M，未匹配样本集合记作 S_{UM}。以倾向得分逆加权调整法为例，在进行了倾向得分估计和样本匹配后，可以得到无回答单元 i 在回答样本中的倾向得分估计为 $\hat{\pi}_i = 1 - \hat{p}(X_i)$。采用单一有放回最近邻匹配，对匹配到的回答单元的权数进行调整，使其可以代表对应的无回答单元，调整后的权数为 $d_{M_i} / \hat{\pi}_i$，$i \in S_M$，得到目标变量均值的倾向得分逆加权估计为 $\hat{\bar{Y}}_{IW} = \dfrac{\sum_{S_M} d_{M_i} Y_{M_i} / \hat{\pi}_i + \sum_{S_{UM}} d_{UM_i} Y_{UM_i}}{\sum_{S_M} d_{M_i} / \hat{\pi}_i + \sum_{S_{UM}} d_{UM_i}}$。

三、基于聚类算法的权数调整

将聚类分析用于无回答情况下的权数调整，聚类根据辅助信息进行，其作用是将样本划分成不同类别，假定每类中的样本具有相似的回答概率，再在聚好的类中进行无回答权数调整。不同的聚类算法和不同的权数调整方法相结合，形成基于机器学习聚类的不同无回答权数调整方法。

（一）常用的聚类算法

聚类分析的目的是自动将数据划分成类，试图从观测全体中寻找同质子类，使得划分后同类别内的观测相似度高，不同类别中的观测相对差异大。与分类规则不同，进行聚类前既不知道将要划分成几类、每类具有什么特征，也不知道依据何种规则区分样本空间。聚类分析的关注点并不在于某一类是什么，其目标仅是把相似的观测聚为一类。因此，聚类分析适用于探索已有数据的特征，不能用于预测，属于无监督的机器学习方法。

假设现有 n 个观测，将其聚为 K 个无交集的类，$D(C_k)$表示第 k 类的类内差异，聚类的目标可表达为最小化目标函数 $\sum_{k=1}^{K} D(C_k)$，即将观测分配到 K 个类中，使得类内差异尽可能小。n 个观测共有 K^n 种分配方式，除非 K 和 n 都很小，否则，K^n 将会非常大，难以遍历所有可能出现的情况。为实现更高效准确的聚类，各种算法应运而生。下文的讨论以观测为基础进行聚类，对一些经典的聚类方法进行回顾。

1. *K-means* 聚类。

K-means 是一种通过不断地聚类、划分，将数据集分成 K 个不重复的类的方法，其基本思想是随机给定 K 个点作为初始类中心，按照就近原则把待划分的样本点分到各个类，然后按平均值重新计算各个类的中心，一直迭代，直到类中心的移动距离小于某个给定的值。

K-means 算法采用欧氏距离作为距离度量，其过程如下：第一步，对于某一数据集，随机选取 K 个点作为类中心；第二步，寻找与每个观测距离最近的类中心，并划分为同一类，此时数据集被划分为 K 个类；第三步，对这 K 个类，计算各类别中所有观测的坐标平均值，并将这个平均值作为新的类中心；第四步，重复执行第二步和第三步，直到类中心不再大范围改变或达到指定聚类次数，聚类完成。

通过 *K-means* 算法的过程，可知该方法具有以下特点：（1）需首先确定所要划分的类数 K；（2）聚类的结果与初始点的选取有关、对离群值较敏感；（3）结果可能收敛到局部最小值，而不是全局最小值；（4）只能处理数值型数据，分类型数据需要进行转换，方可执行该算法。

对于所要划分的类数 K，可根据“肘部法则”和轮廓系数确定。“肘部法则”根据畸变程度确定 K，畸变程度是每个类的类中心与类内样本点的平方距离误差和，畸变程度越低，代表类内样本越紧密，反之，代表类内样本越松散。畸变程度随着类别的增加而降低，在达到某个临界点时得到明显改善，随后，改善程度减缓，这个临界点对应的类别个数即为所要划分的类数 K。轮廓系数用于评价类的密集与分散程度，计算公式为 $S=(b-a)/max(a,b)$，其中 a 代表类别 A 内样本之间距离的均值，类别 B 是与类别 A 距离最近的类，b 代表类别 A 到类别 B 内样本距离的均值。S 越大，表明类之间的离散性越好，聚类效果好。由此可根据轮廓系数确定所要划分的类数 K。

2. *K-modes* 聚类。

K-modes 算法是 *K-means* 算法的扩展，主要解决 n 个分类型属性的观测的分类问题。该算法

采用 0-1 匹配法计算不同属性的相异度，从而衡量不同属性值间的距离，相异度越小则距离越小，相比 *K-means* 算法，该算法计算更方便且时间复杂度更低[7-8]。

K-modes 算法的思想与 *K-means* 算法类似，首先随机选取初始类中心，根据相异度划分初始类，不断地进行类中心的计算和类的划分，直至类中心不再发生变化或准则函数收敛，聚类结束。定义数据对象集合 $X=\{X_1,X_2,\cdots,X_n\}$，对象的属性集合 $A=\{A_1,A_2,\cdots,A_m\}$，聚类过程可数学表达为最小化目标函数 $F(W,Z)=\sum_{l=1}^{K}\sum_{i=1}^{n}w_{li}d(Z_l,X_i)$，其中 $w_{li}=\{0,1\}$，$\sum_{l=1}^{K}w_{li}=1,0<\sum_{i=1}^{n}w_{li}<n$；$d(Z_l,X_i)=\sum_{j=1}^{m}\delta(z_j,x_i)$，若 $x_i\neq z_j$，$\delta(z_j,x_i)=1$，否则，$\delta(z_j,x_i)=0$；$1\leqslant l\leqslant K$，$1\leqslant i\leqslant n$。

经典 *K-modes* 算法的过程如下：第一步，在数据集中随机选取 K 个点作为初始类中心，记作 $Z^{(1)}$，计算 $W^{(1)}$使得目标函数 $F(W,Z^{(1)})$最小，此时 $t=1$；第二步，计算 $Z^{(t+1)}$使得目标函数 $F(W^{(t)},Z^{(t+1)})$最小；第三步，计算 $W^{(t+1)}$使得目标函数 $F(W^{(t+1)},Z^{(t+1)})$最小；第四步，重复执行第二步和第三步，直到目标函数 F 不再减小，聚类完成。

K-modes 算法是分类型数据的针对性方法，其优点体现在以下两点：（1）可直接用于原始数据是分类型时的聚类，而 *K-means* 算法在使用前需对原始数据进行转换；（2）能够有效地处理大规模的数据。*K-modes* 算法的缺点与 *K-means* 算法类似：初始类中心的选取不唯一且会影响聚类结果；有可能收敛到局部最优而不是全局最优。经典 *K-modes* 算法采用 0-1 相异度度量距离，没有体现类中观测间的区别，对算法的改进可从此着手。

3.层次聚类。

层次法是一种基于凝聚或分裂的聚类算法。凝聚层次算法初始阶段将每个样本点分别作为一类，不断合并这些类直至达到预期的类别个数或者其他终止条件；分裂层次算法则在初始阶段将所有的样本点当作同一类，不断进行分裂，直至达到预期的类别个数或者其他终止条件。这两种方法没有优劣之分，在实际应用中，根据数据特点和所要划分的类数 K 进行选择。层次聚类中的测量类间距离的方法有最短距离法、最长距离法、中间距离法、类平均法等[9]。

以基于最小距离的凝聚层次聚类为例，其过程如下：第一步，每个点看作一类，计算两两之间的距离；第二步，将距离最近的两个类合并成一类；第三步，重新计算新类与所有类之间的距离；第四步，重复执行第二步和第三步，直到所有类合并成一类。

层次聚类的优点主要体现在：易于实施，不需要预先确定聚类数；可以发现类之间的层次关系。缺点主要体现在：计算复杂度过高；离群值会对聚类结果产生较大影响；可能将数据聚类成链状。

4. *PAM* 聚类。

PAM 算法，又称 K 中心点算法，是一种围绕中心点的聚类方法，其基本思想为：先为每个类随机选择一个中心点，其他观测根据其与中心点的相异度或距离进行聚类，然后反复地用非中心点来替换中心点，直到聚类效果无法提高[10]。聚类效果用代价函数评价，用于反映某一非中心点是否是现有中心点的优良替代。代价函数由每一个非中心点 O_j 的代价 C_{ijh} 求和得到，现用函数 $d(O_i,O_j)$表示 O_i、O_j 两点间距离，C_{ijh} 的计算方式如下。

情况 1：O_j 当前属于中心点 O_i 所在的类，除中心点外，O_{j2} 是距离 O_j 最近的点，且 $d(O_j,O_{j2})\leqslant d(O_j,O_h)$，此时，$O_h$ 替换 O_i 成为中心点，O_j 将会被划分到 O_{j2} 所在的类，对 O_j 而言，代价 $C_{jih}=d(O_j,O_{j2})-d(O_j,O_i)$。

情况 2：O_j 当前属于中心点 O_i 所在的类，除中心点外，O_{j2} 是距离 O_j 最近的点，且 $d(O_j,O_h)<d(O_j,O_{j2})$，此时，O_h 替换 O_i 成为中心点，O_j 将会被划分到 O_h 所在的类，对 O_j 而言，代价 $C_{jih}=d(O_j,O_h)-d(O_j,O_i)$。

情况 3：O_j 当前不属于中心点 O_i 所在的类，

O_{j2} 是 O_j 所属类的中心点，且 $d(O_j, O_{j2}) \leqslant d(O_j, O_h)$，此时，$O_h$ 替换 O_i 成为中心点，O_j 仍被划分到 O_{j2} 所在的类，对 O_j 而言，代价 $C_{jih}=0$。

情况 4：O_j 当前不属于中心点 O_i 所在的类，O_{j2} 是 O_j 所属类的中心点，且 $d(O_j, O_h)<d(O_j, O_{j2})$，此时，$O_h$ 替换 O_i 成为中心点，O_j 将会被划分到 O_h 所在的类，对 O_j 而言，代价 $C_{jih}=d(O_j, O_h)-d(O_j, O_{j2})$。

由 $n-K$ 个非中心点的代价函数得到总代价 $T_c=\sum_{j=1}^{n-k} C_{jih}$。*PAM* 算法通过最小化代价函数 T_c 进行，其过程如下：第一步，在数据集中随机选取 K 个点作为初始类中心；第二步，将剩余 $n-K$ 个点分配到与其距离最近的类中心所在的类；第三步，随机选择一个非中心点 O_j，根据中心点 O_i 和非中心点 O_h 计算代价 C_{jih}，求和得到总代价 T_c，找出总代价 T_c 最小的聚类方式，若 $minT_c<0$，则用 O_h 替换 O_i；第四步，重复执行第二步和第三步，直到 T_c 不再减小，且 $minT_c \geqslant 0$，聚类完成。

PAM 算法的优点主要体现在：该算法较稳健，对离群值不敏感；能够处理不同类型的数据。但该算法的过程较复杂，执行成本较高。

（二）基于聚类的无回答权数调整

无回答权数调整是指当出现样本无回答时，将缺失单元的权数分解到有回答样本单元身上，以减少缺失数据的影响。在聚类算法将样本分为 K 类后，假定每类中的样本具有相似的回答概率，权数的调整分类进行，具体方法根据实际情况选取。现从以下两种不同角度出发进行权数调整。

1.估计回答概率。

对无回答权数的调整可通过估计回答单元的回答概率对其设计权数 d_i 调整进行，得到调整后的权数 $w_i=d_i \cdot \varphi_i=(\pi_i \cdot p_i)^{-1}$，其中 p_i 是回答概率，$\varphi_i=1/p_i$ 是调整因子。对回答概率 p_i 有不同的估计方法，从而形成不同的无回答调整法。

例如，回答频率可作为回答概率的估计，得到调整后权数 $w_i=d_i \cdot n/m$，其中 n 为总样本量，m 为回答单元样本量。还可根据辅助变量 z_i 对回答概率建模，得到响应模型 $p_i=p\left(z_i^T\gamma\right)=1/h(z_i^T\gamma)$，其中 γ 是未知参数，$h(\eta)$ 是一个单调、二阶可微函数，可采取不同形式，如 $h(\eta)=1+\eta$、$h(\eta)=1+\exp(-\eta)$、$h(\eta)=\exp(\eta)$ 等，得到调整后权数 $w_k=d_k h\left(z_k^T g\right)$，其中 g 是未知参数 γ 的估计。

通过估计回答概率对权数进行调整，效果取决于对回答概率的估计优劣，若模型识别错误，不仅无法弥补无回答误差，还可能得到较差的估计。

2.直接分配。

由于聚类算法已通过辅助变量将相似的样本归为一类，可直接将无回答单元的权数通过某种方式分配给回答单元，此处考虑平均分配和成比例分配两种方式。

在类别 k 中 $k \in \{1,2,\cdots,K\}$，n_{kr} 表示回答样本量，$\sum_{j\in k} d_{nrj}$ 表示无回答单元的原权数和，$\sum_{j\in k} d_j$ 表示回答单元的原权数和，d_i 表示回答单元的原权数，Y_i 表示回答单元的目标变量值。平均分配是指在第 k 类中，将无回答单元的权数按简单平均的方式分配给回答单元，得到回答单元调整后的权数为 $d_i+\frac{\sum_{j\in k} d_{nrj}}{n_{kr}}$，目标变量均值的加权估计为 $\hat{\bar{Y}}_a=\frac{\sum_{k=1}^{K}\sum_{i\in k}\left(d_i+\frac{\sum_{j\in k} d_{nrj}}{n_{kr}}\right)Y_i}{\sum_{k=1}^{K}\sum_{i\in k}\left(d_i+\frac{\sum_{j\in k} d_{nrj}}{n_{kr}}\right)}$。成比例分配是指在第 k 类中，将无回答单元的权数根据回答单元的权数分配给回答单元，得到回答单元调整后的权数为 $d_i+\sum_{j\in k} d_{nrj}\cdot\frac{d_i}{\sum_{j\in k} d_j}$，得到目标变量均值的加权估计为 $\hat{\bar{Y}}_w=\frac{\sum_{k=1}^{K}\sum_{i\in k}\left(d_i+\sum_{j\in k} d_{nrj}\cdot\frac{d_i}{\sum_{j\in k} d_j}\right)Y_i}{\sum_{k=1}^{K}\sum_{i\in k}\left(d_i+\sum_{j\in k} d_{nrj}\cdot\frac{d_i}{\sum_{j\in k} d_j}\right)}$。

直接分配的思想操作简洁，且与聚类的作用相契合，在聚好的类中易于实现回答单元对无回答单元的代表作用。与简单平均分配相比，成比例分配利用原权数将回答单元区分开，能够体现出不同回答单元间的区别。

四、实证研究

（一）数据来源与整理

本文采用 2015 年 CGSS（China General Social Survey）调查数据，进行无回答情况下的权数调整研究。CGSS 调查始于 2003 年，是我国最早的全国性、综合性、连续性学术调查项目，调查的目标总体范围涵盖了全国 31 个省、自治区、直辖市（不含中国港澳台地区）的所有城市、农村家庭户，并通过分层三阶段抽样的方式获取了全国层面的代表性样本。

收入属个人隐私，收入水平不同的人群回答率是不同的，这是典型的不可忽略的无回答机制。本文选取个人年收入作为目标变量进行均值估计，将个人年收入的均值简称为人均年收入；用于权数调整的背景变量有性别、民族、宗教信仰、政治面貌、父母受教育程度以及本人受教育程度等。

数据整理的目的是得到完整数据集作为总体情况的代表，并以此为基础构造无回答样本。本文将个人收入变量选择“无法回答”“拒绝回答”“不知道”“不适用”和“其他”的个案视为无回答；剔除了个人年收入完全为 0 和“不适用”的个案；剔除了背景变量和设计权数缺失的个案。最终获得 7677 个有效个案，得到虚拟总体的人均年收入 $\bar{T}=37042.92$。

（二）无回答的构造与聚类方法的设置

校准法和倾向得分调整法是进行无回答情况下权数调整的常用方法。为探究聚类分析算法在不可忽略的无回答机制下的估计优势，实证研究分别采用分别校准法、倾向得分调整法、*K-means* 聚类、*K-modes* 聚类、层次聚类和 *PAM* 聚类进行权数调整。现从无回答样本的构造以及聚类方法的有关设置两方面对实证过程进行说明。

1.无回答样本的构造。

假定单元是否作答服从二项分布。实证中，对于可忽略的无回答机制，根据受教育程度 x_1 和性别 x_2 计算 $p_{nri}=a+b\cdot10^{-10}\cdot x_1-c\cdot x_2+\varepsilon_i$，其中 $a\sim U(0.2,0.5)$，$b\sim U(0.1,0.5)$，$c\sim U(0.01,0.1)$，$\varepsilon_i\sim N(0,0.01)$，再根据 p_{nri} 由二项分布随机生成无回答单元，得到的样本无回答率约 30%。对于不可忽略的无回答机制，个人年收入为目标变量，其回答情况与收入本身密切相关，假定收入越高无回答率越高，无回答概率采用线性函数和二次函数两种形式，具体如下。

假定单元无回答概率 p_{nri} 与收入 y 具有线性关系：根据收入计算 $p_i=a+b\cdot10^{-10}\cdot y+\varepsilon_i$，其中 $a\sim U(0,1)$，$b\sim U(0.1,1)$，$\varepsilon_i\sim N(0,1)$；将 p_i 调整到 0~0.4 的范围内，得到各单元的无回答概率 $p_{nri}=0.4\times\frac{p_i-min(p_i)}{max(p_i)-min(p_i)}$；根据 p_{nri} 由二项分布随机生成无回答单元。由于 $b>0$，符合收入越高无回答率越高的假定，据此得到的样本无回答率约 20%。

假定单元无回答概率 p_{nri} 与收入 y 具有二次函数关系：$p_{nri}=a+b\cdot10^{-10}\cdot y+c\cdot10^{-14}\cdot y^2+\varepsilon_i$，其中 $a\sim U(0.1,0.5)$，$b\sim U(0.1,1)$，$c\sim U(0.1,1)$，$\varepsilon_i\sim N(0,0.01)$，根据 p_{nri} 由二项分布随机生成无回答单元。由于该二次函数开口向上且对称轴为负，符合收入越高无回答率越高的假定，据此得到的样本无回答率约 25%。

实证中的样本无回答率较高，有必要进行权数的调整以提高估计效果。

2.调整方法设置。

所有无回答权数调整方法均根据 7 个背景变量：性别、民族、宗教信仰、政治面貌、父母受教育程度以及本人受教育程度等进行；倾向得分调整法采用最近邻匹配的原则进行逆加权调整；根据“肘部法则”和轮廓系数，得到 *K-means* 算法的最佳分类数为 $K=4$；为方便对比不同聚类算法间的估计效果，所有聚类算法的分类数设置相同；为探究不同分类数对聚类方法的影响，分别采用 $K=4$、10、20 进行研究。

在完成样本的聚类后，将每类中无回答单元的权数分解到回答单元上，采取简单平均分配和根据回答单元的设计权数成比例分配两种方式

进行。

分别在可忽略的无回答机制、基于线性函数和二次函数的不可忽略的无回答机制下，利用校准法和倾向得分调整法进行无回答权数调整；聚类算法下的权数调整方法设置如表 1 所示。

表 1　　无回答权数调整方法设置

		可忽略的无回答机制	不可忽略的无回答机制			
			线性函数			二次函数
聚类调整的分类数 K		K=4	K=4	K=10	K=20	K=4
平均分配	*K-means* 聚类	✔	✔	✔	✔	✔
	K-modes 聚类	✔	✔	✔	✔	✔
	层次聚类	✔	✔	✔	✔	✔
	PAM 聚类	✔	✔	✔	✔	✔
成比例分配	*K-means* 聚类	✔	✔	-	-	-
	K-modes 聚类	✔	✔	-	-	-
	层次聚类	✔	✔	-	-	-
	PAM 聚类	✔	✔	-	-	-

注：“✔”表示在对应的聚类数和权数分配方式下进行实证，“-”表示未进行。

通过前述的实证设置，可以在不同无回答机制下，对比各权数调整方法的相对优劣，并探究不同分类数 K 和不同权数调整方式对基于聚类算法的权数调整效果的影响。

（三）结果与分析

本文共进行 500 次重复试验，按照上文所述的构造方式随机生成无回答单元，分别采用校准法、倾向得分调整法、*K-means* 聚类、*K-modes* 聚类、层次聚类和 *PAM* 聚类进行权数调整和参数估计。方法的评价准则采用相对偏差、标准误（*SE*）、均方误差根（*RMSE*）和权效应（*deff*）的平均值，计算公式为：相对偏差=$\sum_{i=1}^{R}\frac{\hat{\bar{t}}_r-\bar{T}}{\bar{T}}/R$；标准误=$\sqrt{\frac{\sum_{i=1}^{R}(\hat{\bar{t}}_r-\sum_{s=1}^{R}\hat{\bar{t}}_r/R)^2}{R-1}}$；均方误差根=$\sqrt{Bias^2+SE^2}$；权效应 $deff\approx \mathrm{CV}(\omega)^2+1$。其中 $\hat{\bar{t}}_r$ 表示一次重复试验对人均年收入的估计；R=500 表示重复试验次数；$\mathrm{CV}(\omega)^2$ 表示权数的变异系数。

1.可忽略机制下的无回答权数调整。

在可忽略机制下的无回答中，不同方法的无回答权数调整效果见表 2。

由表 2 可知，在所有的权数调整方法中，倾向得分调整法的相对偏差、标准误、均方误差根和权效应平均值最大，校准法次之。无论是平均分配还是根据回答单元设计权数成比例分配的聚类方法，得到的调整后权数变异性更小，估计量的相对偏差、标准误、均方误差根更小，估计效果更优。

表 2　　可忽略机制下的调整效果对比

		相对偏差	*SE*	*RMSE*	*deff* 平均值
倾向得分调整法（逆加权）		–0.5676	8086.4230	8087.4133	1.3316
校准法		0.0424	1518.2197	1530.8832	1.3741
平均分配	*K–means* 聚类	0.0132	1339.6635	1353.5991	1.1403
	K–modes 聚类	0.0152	1367.4316	1381.1144	1.1397
	层次聚类	0.0133	1337.4872	1351.4467	1.1406
	PAM 聚类	0.0065	1302.5800	1316.8144	1.2922
成比例分配	*K–means* 聚类	0.0072	1446.2562	1459.0984	1.2919
	K–modes 聚类	0.0064	1317.5673	1331.6395	1.2921
	层次聚类	0.0046	1459.3214	1472.0165	1.2923
	PAM 聚类	0.0053	1361.3409	1374.9497	1.2922

2.不可忽略机制下的无回答权数调整。

以线性函数和二次函数为基础，构造不可忽略的无回答机制，此时，不同方法进行无回答权数调整的具体效果见表 3。

由表 3 可知，基于线性函数构造无回答样本，无论权数分配方式如何，各聚类方法的相对误差明显小于倾向得分调整法和校准法，各聚类方法的标准误小于倾向得分调整法、与校准法相差不大。各聚类方法的均方误差根明显小于倾向得分调整法、与校准法相差不大，但各聚类方法对应的权效应平均值更小，即各聚类方法得到的调整后权数变异性更小。基于二次函数构造无回答样本，各聚类方法的相对误差明显小于倾向得分调整法和校准法，各聚类方法的标准误和均方误差根小于倾向得分调整法，大于校准法。但各聚类方法对应的权效应平均值更小，得到的调整后权数变异性更小。总体而言，基于聚类算法的权数调整效果更好。

表 3　　不可忽略机制下的调整效果对比

			相对偏差	*SE*	*RMSE*	*deff* 平均值
线性函数	倾向得分调整法（逆加权）		–0.2923	1499.4975	1508.2140	1.3055
	校准法		0.0283	986.7768	1005.8918	1.4352
	平均分配	*K–means* 聚类	0.0064	1039.2562	1057.0398	1.1874
		K–modes 聚类	0.0078	1043.7326	1061.4664	1.1876
		层次聚类	0.0064	1037.6157	1055.4273	1.1876
		PAM 聚类	–0.0002	1022.7025	1040.6521	1.2920
	成比例分配	*K–means* 聚类	–0.0005	1021.1428	1039.1129	1.2922
		K–modes 聚类	0.0012	981.7825	1000.4926	1.2921
		层次聚类	–0.0004	1002.2969	1020.6019	1.2919
		PAM 聚类	–0.0012	1039.0754	1056.7285	1.2921
二次函数	倾向得分调整法（逆加权）		–0.5300	9696.4030	9697.3007	1.3252
	校准法		0.0360	1444.2449	1457.4699	1.3847
	平均分配	*K–means* 聚类	–0.0132	1951.9661	1961.3071	1.1487
		K–modes 聚类	–0.0108	1979.5848	1988.8181	1.1492
		层次聚类	–0.0131	1947.5234	1956.8868	1.1490
		PAM 聚类	–0.0215	1803.7893	1813.8092	1.2922

3.影响因素探究。

对于聚类方法效果的影响因素探究，无回答样本根据线性函数构造；权数分配方式采取平均分配和根据回答单元设计权数成比例分配；聚类数分别择 4、10 和 20。不同权数分配方式的影响见表 2 及表 3 上半部分，不同聚类数的影响见表 4，其中 K=4 与表 3 对应部分相同，为方便对比再次列出。

表 4　　不可忽略机制下的调整效果对比

聚类数 K		相对偏差	*SE*	*RMSE*	*deff* 平均值
K=4	*K–means* 聚类	0.0064	1039.2562	1057.0398	1.1874
	K–modes 聚类	0.0078	1043.7326	1061.4664	1.1876
	层次聚类	0.0064	1037.6157	1055.4273	1.1876
	PAM 聚类	–0.0002	1022.7025	1040.6521	1.2920
K=10	*K–means* 聚类	0.0076	1006.2581	1024.6360	1.1892
	K–modes 聚类	0.0088	1012.2929	1030.5858	1.1891
	层次聚类	0.0082	1009.5315	1027.8615	1.1886
	PAM 聚类	0.0013	1000.2332	1018.6062	1.2921
K=20	*K–means* 聚类	0.0053	1039.9870	1057.7389	1.1909
	K–modes 聚类	0.0063	1042.9759	1060.6961	1.1907
	层次聚类	0.0052	1042.4079	1060.1180	1.1906
	PAM 聚类	–0.0009	1028.2686	1046.1096	1.2920

由表 2 可知，在可忽略的无回答机制下，根据回答单元设计权数成比例分配相比简单平均分配，相对误差明显减小、标准误和均方误差根变化不大，权效应的平均值有所增加，其中 *PAM* 聚类方法的效果变化最小。由表 3 可知，在基于线性函数的不可忽略的无回答机制下，根据回答单元设计权数成比例分配相比简单平均分配，除 *PAM* 聚类外，其他聚类方法的相对误差明显减小、标准误和均方误差根变化不大，权效应的平均值有所增加，*PAM* 聚类方法的相对偏差有略微增加，但仍很小，标准误和均方误差根及权效应的变化不大。

由表 4 可知，随着聚类数 K 的增加，各聚类方法的标准误、均方误差根变化不大，相对偏差的变化没有明显规律。

总体而言，根据回答单元设计权数的成比例分配能够更好地反映回答单元之间的区别，估计的相对偏差有所降低，虽然权效应有所增加，但仍在可接受的范围内，无论是简单平均分配还是根据回答单元设计权数的成比例分配均能有效控制权数的变异性，得到性质优良的估计。此外，聚类数 K 的增加不会对权数调整效果造成明显影响。具体可根据实际情况进行选择。

（四）评价

在可忽略的无回答机制，尤其是现实中更一般的不可忽略的无回答机制下，基于聚类方法的无回答调整方法能够有效控制无回答偏差，获得精度更高的估计量和变异性更小的调整后权数。

基于聚类方法的无回答调整方法的优势主要体现在以下两点：一是适用于不可忽略的无回答机制；二是聚类方法相对稳健、易于执行，可与不同的权数调整方法结合，可根据实际情况设置聚类数 K。

五、总结与讨论

当样本存在无回答时，直接利用回答样本进行推断势必是有偏且效率低的，尤其当无回答机制不可忽略时，样本是否作答与目标变量有关，潜在的无回答偏差不容忽视。无论缺失数据机制如何，基于模型的方法对模型假设和模型识别具有较强的要求，当实际数据不满足假设或模型识别错误时，可能造成估计失效。

本文采用基于机器学习的聚类算法，将样本分为不同类，在每类中进行无回答权数调整。在聚类前不对每类样本的特征加以限制，假定聚类后每类样本具有相似的无回答概率，聚类过程未考虑目标变量与参加聚类的样本背景信息是否有关、目标变量是否与其他变量有关。相比传统方法对无回答（或回答）概率进行建模调整，聚类方法的度量相对“模糊”，因此，其效果受无回答机制的影响相对更小。

本文采用 2015 年 CGSS 调查数据进行了实证研究。结果表明，无论是可忽略的无回答机制，还是不可忽略的无回答机制，与校准法和倾向得分调整法相比，使用聚类算法进行无回答权数调整可以得到变异性更小的最终权数，有效降低了估计偏差。同时，聚类方法更为灵活、稳健，在不同的权数调整方法和聚类数 K 下的调整效果保持更优。

本文采用聚类分析的方法有效解决了不可忽略无回答机制下的权数调整和目标变量的估计问题。进一步研究可将单元回答概率的估计纳入聚类算法，具体操作需另文讨论。

参考文献

[1] Rubin D B. Inference and Missing Data[J]. Biometrika, 1976, 63(3): 581-592.

[2] 金勇进, 刘晓宇. 抽样推断中权数的调整与检验——以美国 ECLS-K (2011) 调查为例[J]. 调研世界, 2020(1): 25-33.

[3] Deville J C, Särndal C E. Calibration Estimators in Survey Samping[J]. Journal of the American Statistical Association, 1992, 87(418): 376-382.

[4] Zieschang K D. Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey[J]. Journal of the American Statistical Association, 1990, 85(412): 986-1001.

[5] Kott P S. Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors, Survey Methodology 32[J]. Survey Methodology, 2006.

[6] Lundström S, Särndal C E. Calibration as a Standard Method for the Treatment of Nonresponse[J]. Journal of Official Statistics, 1999, 15(2): 305-327.

[7] Zhexue Huang. Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Know ledge Discovery, 1998, 2(3): 283304.

[8] Ng M K, Li M J, Huang J Z, et al. On the Impact of Dissimilarity Measure in K-modes Clustering Algorithm[J]. Ieee Transactions on Pattern Analysis & Machine Intelligence, 2007.

[9] 段明秀. 层次聚类算法的研究及应用[D]. 中南大学, 2009.

[10] 陈志强, 刘钊, 张建辉. 聚类分析中 PAM 算法的分析与实现[J]. 计算机与现代化, 2003(9): 1-3,6.

作者简介：

金勇进，男，1953 年生，北京人，现为中国人民大学统计学院教授、博士生导师，研究方向为抽样技术。

刘晓宇，女，1995 年生，山西运城人，现为中国人民大学统计学院在读博士研究生，研究方向为抽样技术。

(责任编辑：黄煌)