数据挖掘方法应用于调查数据的抽样权重问题

——基于放回比例抽样的再抽样方法

谢佳斌　金勇进　谢邦昌

内容提要：在将数据挖掘方法应用于抽样调查数据时，会遇到抽样权重的处理问题。本文提出采用放回的、与样本单元权数大小成比例的再抽样方法，简称 PPWWR 再抽样，来实现“事后”自加权设计。实现“事后”自加权设计后的子样本可忽略掉样本权数，直接采用常规的图示方法和数据挖掘算法进行分析。随后，基于 2007 中国公民科学素质调查贵州省数据，通过模拟分析讨论了 PPWWR 再抽样子样本的样本量问题，发现 max（n，5%N）是一个比较合适的样本量。这一结论可能为其他大型复杂抽样调查数据的数据挖掘实施问题提供借鉴。

关键词：调查数据；抽样权重；数据挖掘；PPWWR 再抽样
中图分类号：C811　　文献标识码：A　　文章编号：1002－4565(2009)04－0101－04

The Study on Handling Sampling Weights Associated with the Survey Data When Applying Data Mining Methods

——Based on the Method of Re-sampling with PPWWR

Xie Jiabin　Jin Yongjin　Xie Bangchang

Abstract: The problem of how to deal with sampling weights appears when applying data mining methods to survey data. We suggest the method of re-sampling with probability proportional to the weights with replacement (PPWWR) to achieve post self-weighting design. Then, some ordinary statistical graphics and data mining algorithms can be used directly, ignoring the sample weights. Next, based on the survey data of GuiZhou Province from the survey of public understanding of science 2007, we discussed the sample size problem of the PPWWR re-sampling method by simulation and find max(n, 5% N) is an appropriate sample size. This conclusion might be useful for the implementation of data mining on other large and complex survey data.

Key words: Survey data; Sampling weights; Data mining; PPWWR re-sampling

一、样本权重问题

数据挖掘本质上作为一类数据分析方法，和统计学有着共同的目标：发现数据中的结构[1]。因而，基于数据挖掘的视角，对抽样调查数据采用一些数据挖掘的方法进行分析，是可行的，文献[2]就提供了一个范例。然而，将数据挖掘方法应用于抽样调查数据，有一个问题通常无法回避，那就是样本数据所对应的权数如何处理。

一般而言，数据挖掘问题常常针对总体数据，例如关于一个公司的所有职工数据，银行信用卡中心数据库的所有客户数据，一家大型超市一个季度以来的所有顾客购买记录等。在这种情形下，每一条记录都是总体数据中的一个单元，得到的观察值可以直接计算总体参数，无需进行统计推断。

但数据挖掘方法也越来越多地应用于抽样调查数据。与总体数据不同的是，抽样调查当中，每个样本单元的观测值都是有权数的，权数表示的是每个样本单元代表了总体中一定数目的单元，所以整个样本就“代表”了整个总体。样本单元的权数取决于抽样设计。例如，对于分层抽样，有

$$t_{str} = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}$$

其中抽样权重 $w_{hj}=N_h/n_h$ 可以看作样本观测值 y_{hj} 所代表的总体中观测值的数目，其值为该样本单元入样概率的倒数。

倘若调查采用的是自加权设计，则各样本单元的抽样权重是相等的。在不考虑非抽样误差的情况下，可以认为自加权样本完全代表了总体，因为每个样本单元都代表了总体中相同数目的单元。此时，可以忽略掉抽样权重，直接采用一些简单的图形实现对数据的探索性分析，进而在对数据进行了充分理解和准备的基础上，直接调用相关算法进行挖掘。

然而，基于一些原因，部分大规模抽样调查并不采用自加权设计，这使得各样本单元对应的抽样权重大小不一。并且，在大型复杂抽样调查中，为使得调查得到的样本结构尽可能与总体结构相一致，在处理样本数据时，还通常采用基于多变量辅助信息等的校准加权方法对样本结构进行加权调整，以减少样本结构与总体结构的差异性。也就是说，根据入样概率求得样本单元的初始权数 w_i，再利用辅助信息进行加权调整便得到每个样本单元的最终权数 w_i^*。从而，即便调查采用的是自加权设计，加权调整后各样本对应的最终权数也大都是不相同的。在这种情况下，一方面，通常用于描述简单随机样本的统计图形在描述权数不一的样本数据时，往往会产生错误，因为没有考虑不等的权数问题；另一方面，如果忽视权数问题，直接调用相关算法对收集上来的样本数据实施挖掘，所得到的结果可能是误导性的，或者很难解释。

二、解决思路

部分文献[7][8]提出用气泡图(bubble plots)来展示复杂调查数据的信息，图 1 展示了美国 1988 年全国母亲和婴儿健康调查中 30－39 岁母亲的出生体重和女儿的生出体重之间的关系。图中，每个圆圈对应一条样本数据，每个圆圈的面积与样本的权重成正比。

相比普通散点图，此类气泡图的优点是将样本数据对应的权重信息也展现出来，避免了普通散点图误导性的视觉效果。但当样本数据较多，或者个别样本权重差异过大时，气泡图会显得非常混乱。另外，气泡图只是对普通散点图的改进，我们需要寻找一种方法，既能够适用于大部分图形，同时又考虑了样本的权重信息。

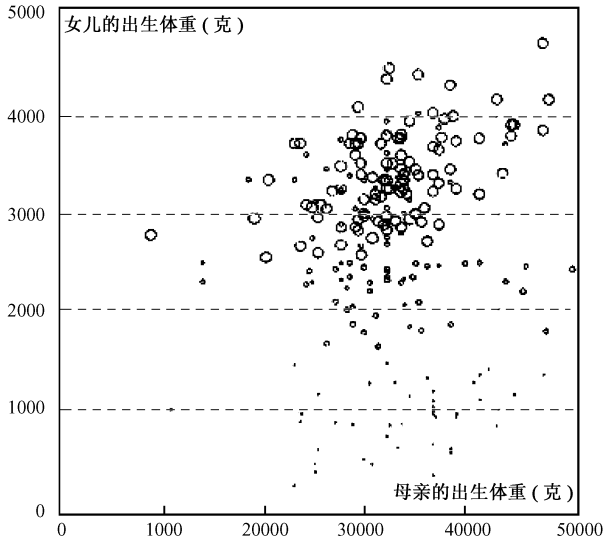

图 1　母亲和女儿出生体重关系的泡泡图

考虑到如前所述自加权样本的优越性，借鉴文献[4]和文献[6]的思想，本文提出放回的、与样本权数大小成比例的再抽样方法(re-sampling with probability proportional to the weights with replacement)，简称 PPWWR 再抽样，来实现“事后”自加权设计。具体如下：

假设样本量大小为 n，对于样本 i，$i=1,2,\cdots,n$，其对应的抽样权重为 w_i，经加权调整后的最终权数为 w_i^*。其中，$\sum_{i=1}^{n} w_i^* = N$，N 为总体单元个数。在原样本内，按权数 w_i^* 的大小采用有放回比例抽样的方法抽取一个大小为 n' 的子样本。可以证明，实施这样的重抽样后，对于子样本 n'，每个样本单元的权数都相同。证明如下：

1. 由于为放回的与样本权数大小成比例的概率抽样，因此，原样本中，样本单元 i 每次被抽中的概率 $Z_i = \dfrac{w_i^*}{\sum_{i=1}^{n} w_i^*} = \dfrac{w_i^*}{N}$；

2. 原样本中第 i 个样本单元被重复抽中的期望次数 $m_i = n'Z_i = n'\dfrac{w_i^*}{N}$；

3. 从大小为 n 的原样本中抽取大小为 n' 的子样本的过程，可看作是将第 i 个单元的权数 w_i^* 平均分解到子样本中对应的 m_i 个样本单元的过程。因而，子样本 n' 中，每个样本单元对应的新权数

$$w_j = \frac{w_i^*}{m_i} = \frac{w_i^*}{\frac{n' w_i^*}{N}} = \frac{N}{n'}。$$

由此, 可以把子样本 n' 看作是一个自加权样本。于是可以忽略掉样本权数, 直接采用常规的图示方法对数据进行初步的探索分析, 并调用算法对该子样本 n' 实施挖掘。

当然, 从样本 n 到样本 n', 这个过程存在信息丢失。Murthy(1965)证明了将子样本 n' 的总和乘上一个常数后, 便得到 $\sum_{i=1}^{n} w_i^* y_i$ 的一个无偏估计量。并且, 在这一重抽样阶段所增加的方差为

$$E_f\left\{\frac{1}{n}\left[\left(\sum_{i=1}^{n} w_i^*\right)\left(\sum_{i=1}^{n} w_i^* y_i^2\right) - \left(\sum_{i=1}^{n} w_i^* y_i\right)\right]\right\}$$

其中, E_f 代表初始样本 n 范围内的期望值。

接下来的问题是, 子样本 n' 的数量需要多大, 才能保证后面的推断有比较好的效果。在 n' 大小的确定上, 以子样本 n' 不遗漏原有样本 n 为度, 原则上可以是[n, N] 中的任何一个值。当 n' 取 n 时, 由于 n 中的部分样本在重抽样后不再出现, 因而可能有相对的信息丢失; 而 n' 当 N 取时, 如果 N 的数值过大, 则可能造成重抽样及后续分析过程计算量过于庞大, 面临不经济的问题。因而, n' 应该有一个在[n, N] 之间的最优取值, 该取值将在信息丢失和计算量之间达到一个平衡。本文将通过模拟的方法来尝试确定 n' 的合适水平。

三、关于 n' 的模拟分析

由于从理论上难以直接论证 n' 的最优水平, 我们采用模拟的方法进行讨论。本模拟的分析数据取自于 2007 中国公民科学素质调查, 为贵州省的数据。中国公民科学素质调查是通过全国性的抽样调查, 来了解分析我国 18—69 周岁的公民对科学的理解及对科学技术的态度等与公民科学素质相关问题的状况。调查内容包括三个主要方面, 即: 公民对基本科学知识的了解程度; 公民获取科技知识和科学技术发展信息的渠道与方法; 公民对科学技术的态度。调查的指标体系由背景变量和各分级指标组成。背景变量包括: 地区、城乡、性别、年龄、文化程度、职业、民族、重点人群等。调查采用分层三阶不等概抽样方法, 以全国为总体, 兼顾样本在各省级区域的分配。

在进行抽样设计时, 为满足对本地区公民科学素质状况进行推断的需求, 部分省份在全国样本的基础上, 进行了追加样本设计。以贵州省为例, 落在贵州省的全国样本量为 310, 对该地区追加的样本量为 1660, 总样本量为 1970。

在对贵州省的调查数据进行整理时, 首先通过计算每个样本单元的入样概率, 确定了各样本单元的初始权数; 其次, 通过校准加权调整, 得到各个样本单元的最终权数。最终数据由 1970 名受访者的八个背景信息变量、公民获取科技信息的渠道变量、公民科学素质四个方面的测试变量、公民对科技及其发展的态度和看法变量以及每条样本数据对应的最终权数构成。

在尝试对贵州省 2007 年公民科学素质调查数据实施数据挖掘时, 便遇到无从选择现有统计图形对数据进行描述和样本单元的最终权数与数据挖掘算法的衔接问题。而如果采用本文提出的 PPWWR 再抽样方法, 则能较好地解决上述两个问题, 而不用考虑更改现有统计图形或调整已有挖掘算法。

为了确定实施 PPWWR 再抽样方法时子样本 n' 的最合适大小, 这里通过模拟的方法比较 n' 取不同值时子样本 n' 的各辅助变量取值状况与贵州省真实数据之间的差距, 进而确定 n' 的最佳取值。对大小 n=1970 的原始样本按 PPWWR 再抽样的方法分别抽取大小为 n、0.01%N、0.10%N、1%N、5%N、10%N 的子样本, 其中 N 为贵州省的适龄总人口数。并对于每种样本量, 重复抽取 10 次, 比较这 10 次抽取结果中各辅导变量取值的波动性。模拟结果见表 1。

由模拟结果可知, 当按 PPWWR 再抽样, 子样本 n' 的大小定为 5%N 时, 10 次重复抽样中子样本 n' 的各辅助变量取值的均值与贵州省的真实情况基本一致, 并且 10 次重复抽取样本中各辅助变量取值的方差在精确到小数点第二位的情况下为 0。由此, n' 不用取值到 N, 当 $n' \geq 5\%N$ 时, 事后自加权子样本的性别、城乡、年龄和教育程度结构与贵州省的真实情况几乎没有差别。因此, 对于此例子, n' 的最合适大小应为 $\max(n, 5\%N)$。

四、小结

在数据挖掘问题中, 数据的收集方法和分析方法应该是两个不可分割的部分, 是一个整体, 分析方法必须和数据收集时的抽样设计相匹配。

表 1　　再抽样子样本 n' 分别取 n、0.01%N、0.10%N、1%N、5%N、10%N 时的情形(%)

		贵州省真实值	n		0.01%N		0.10%N		1%N		5%N		10%N	
			均值	方差	均值	方差	均值	方差	均值	方差	均值	方差	均值	方差
性别	男性	51.4	51.3	0.49	51.3	0.39	51.4	0.07	51.4	0.01	51.4	0.00	51.4	0.00
	女性	48.6	48.7	0.49	48.7	0.39	48.6	0.07	48.6	0.01	48.6	0.00	48.6	0.00
城乡	乡	63.8	63.7	1.85	63.7	1.51	63.8	0.02	63.7	0.01	63.8	0.00	63.8	0.00
	城	36.2	36.3	1.85	36.3	1.51	36.2	0.02	36.3	0.01	36.2	0.00	36.2	0.00
年龄	18—29	23.7	24.4	0.44	23.8	1.43	23.7	0.05	23.7	0.01	23.7	0.00	23.7	0.00
	30—39	30.4	29.8	1.31	30.0	1.24	30.5	0.12	30.5	0.01	30.4	0.00	30.4	0.00
	40—49	21.0	21.0	0.15	21.2	0.6	21.0	0.08	21.0	0.00	21.0	0.00	21.0	0.00
	50—59	18.1	18.1	0.46	18.1	0.87	18.1	0.06	18.1	0.01	18.1	0.00	18.1	0.00
	60—69	6.8	6.7	0.54	6.9	0.06	6.8	0.01	6.8	0.01	6.8	0.00	6.8	0.00
教育	文盲	16.9	16.8	0.93	17.4	0.23	16.9	0.06	16.9	0.01	16.9	0.00	16.9	0.00
	小学	44.5	44.5	0.93	44.3	0.56	44.6	0.06	44.4	0.02	44.5	0.00	44.5	0.00
	初中	26.6	26.7	0.84	26.4	0.98	26.5	0.05	26.7	0.00	26.6	0.00	26.6	0.00
	高中或中专	7.5	7.6	0.38	7.3	0.28	7.5	0.02	7.6	0.00	7.5	0.00	7.5	0.00
	大专	2.9	2.8	0.16	2.9	0.06	2.9	0.02	2.9	0.00	2.9	0.00	2.9	0.00
	大学及以上	1.6	1.5	0.04	1.7	0.06	1.6	0.01	1.6	0.00	1.6	0.00	1.6	0.00

为解决将数据挖掘方法应用于抽样权重问题，本文提出采用放回的、与样本权数大小成比例的再抽样方法，简称 PPWWR 再抽样，来实现“事后”自加权设计。实现“事后”自加权设计后的子样本可忽略掉样本权数，直接采用常规的图示方法和数据挖掘算法进行分析。随后，基于 2007 年中国公民科学素质调查贵州省数据，本文通过模拟分析讨论了 PPWWR 再抽样子样本的样本量问题，发现 $\max(n, 5\%N)$ 是一个比较合适的样本量。这一结论可能为其他大型复杂抽样调查数据的数据挖掘方法的实施提供借鉴。

参考文献

[1] David J. Hand. Statistics and Data Mining: Intersecting Disciplines [J]. SIGKDD Explorations, 1999(1): 16—19.

[2] 何海鹰、朱建平、谢邦昌. 证券投资意识调查分析——基于数据挖掘的视角[J]. 统计研究, 2008(9): 49—53.

[3] M. N. Murthy and V. K. Sethi. Randomized Rounded-Off Multipliers in Sampling Theory[J]. Journal of the American Statistical Association, 1961(5): 328—334.

[4] M. N. Murthy and V. K. Sethi. Self-Weighting Design at Tabulation Stage[J]. SANKHYA, 1965(2): 201—210.

[5] Susan Hinkins, H. Lock Oh and Fritz Scheuren. Inverse Sampling Design Algorithms[J]. Survey Methodologh, 1997(6): 11—21.

[6] 谢邦昌. 抽样调查的理论及其应用方法[M]. 北京: 中国统计出版社, 1998. 315—329.

[7] Edward L. Korn and Barry I. Graubard (1998). Scatterplots with Survey Data[J]. The American Statistician, 1998(1): 58—69.

[8] Sharon L. Lohr. 抽样: 设计与分析[M]. 北京: 中国统计出版社, 2002. 221—249.

[9] 金勇进、蒋妍、李序颖. 抽样技术[M]. 北京: 中国人民大学出版社, 2002. 99—103.

[10] 滕广青、毛英爽. 国外数据挖掘应用研究与发展分析[J]. 统计研究, 2005(12): 68—70.

作者简介

谢佳斌，男，1985 年 11 月生，籍贯江西，中国人民大学统计学院 2007 级博士研究生，研究方向为抽样调查及数据挖掘。

金勇进，男，1953 年 5 月生，籍贯北京，中国人民大学统计学院院长，博士生导师、教授，研究方向为抽样调查。

谢邦昌，男，1962 年 10 月生，籍贯湖南，1991 年毕业于中国台湾大学获生物统计专业博士学位，现任中国台湾辅仁大学统计信息学系、应用统计所教授、中华资料采矿协会(台湾)理事长，兼任厦门大学经济学院客座教授，博士生导师，中国人民大学统计学院客座教授。主要研究方向数据挖掘与商业智能。

(责任编辑：程　晞)