【大数据应用案例】

# 融合统计思想的大数据算法*

李 扬 张 长 朱建平

内容提要: 海量化的数据规模作为大数据的第一个特征,带来了计算上的首要挑战。大规模样本不一定能够完全替代总体,因此大数据分析的算法设计不仅要考虑精简计算成本,还要考虑如何刻画估计结果的不确定性。本文以分治自助算法和子集双重自助算法为例讨论了兼具计算效率提升和不确定性评价的可并行计算的大数据统计算法设计,通过比较分析探讨设计思想与未来的研究方向。

关键词: 自助法; 不确定性; 大规模数据; 并行计算

DOI: 10.19343/j.cnki.11-1302/c.2018.07.011

中图分类号: O213 文献标识码: A 文章编号: 1002-4565(2018)07-0125-04

## Statistical Algorithms for Big Data

Li Yang Zhang Zhang Zhu Jianpping

**Abstract:** The large volume of massive dataset is the key feature of Big Data which brings the main challengers for computing. The dataset with large sample size cannot always stands for the population, therefore the algorithms design for Big Data should consider how to reduce computing cost and how to characterize the uncertainty of the estimated results. In this paper, we study the design of statistical algorithms for massive dataset considering both computing efficiency and uncertainty assessment. Both the Bag of Little Bootstrap (BLB) and Subsampled Double Bootstrap (SDB) algorithms are discussed as illustrative examples. Additionally, a comparison of BLB and SDB is discussed with conclusions of future work.

**Key words:** Bootstrap; Uncertainty; Massive Data; Parallel Computing

## 一、案例背景

数据规模“大”仍是大数据的第一个特征。海量的数据规模给计算带来双重挑战: 一方面,大规模数据集要求计算设备具有更大的内存容量; 另一方面,数据规模的快速增长会直接导致计算效率的大幅下降。虽然基于分布式的算法设计可以解决计算效率的问题,但这类方法得出的结果只有在大规模数据是总体(或对总体有代表性)时才有意义。大多数情况下,大规模样本因为代表性的问题并不能替代总体,反而可能会因为有偏部分的大样本量夸大局部作用而带来估计的偏倚[1]。因此,大数据分析的算法设计不仅要考虑计算成本,还要考虑如何刻画估计结果的不确定性[2]。虽然这种不确定性会随着数据量的增大而降低,但给定具体的大规模数据集时,研究者仍需要通过比较估计量的不确定性来选择最适合的模型[3]。从统计的角度看,大数据的算法设计应

* 本文为中国人民大学科学研究基金(中央高校基本科研业务费专项资金资助)“生物医学大数据的统计方法基础研究”(15XNI011)的阶段性成果。

兼顾如下三方面功能: 第一,通过减小计算复杂度降低运算负荷; 第二,通过并行计算提升运算效率; 第三,通过度量不确定性刻画估计的有效性。

以 m - n 自助法[4]、子集抽样自助法[5]为代表的方法虽然通过抽样降低数据量级实现了计算速度的提升,但同时降低了数据的变异性。另一方面,基于快速双重自助法[6-8]的算法设计虽然可以同时提升运算效率与估计精度,但两次自助抽样的计算成本仍远高于传统自助法。本文以 Kleiner 等[3]提出的分治自助算法( Bag of Little Bootstrap,BLB) 和 Sengupta 等[2]提出的子集双重自助算法( Subsampled Double Bootstrap,SDB) 为例讨论兼具计算效率和不确定性评价的大数据并行统计算法。

## 二、实现过程

自助法通过重复有放回抽样得到一组对总体有代表性的经验样本来构造估计量的经验分布。由于基于每个经验样本的估计过程是独立的,自助法具备进行并行计算的条件。然而,在传统的自助法中每个经验样本平均包含总样本中 63% 的样本单元,当样本量 n 很大时计算复杂度仍为 $O(n)$。虽然子集抽样自助法[7]基于总样本子集构造经验样本的方法可以实现降低计算复杂度的目的,但数据变异性的相应降低导致其结果依赖于子样本的选取。Bickel 和 Sakov[9]提出了基于数据的子样本确定方法,但该方法的计算量限制了其在大规模数据上的应用。而分治自助算法和子集双重自助算法通过针对子样本的自助法实现数据变异性的调整,在降低计算成本的同时实现了数据变异性的还原,具有较好的估计不确定性度量。

### (一) 分治自助算法

分治自助算法设计如表 1,包含三个关键步骤: 第一,针对总样本无放回地抽取样本子集以降低计算复杂度; 第二,针对子样本通过自助法实现数据变异性的还原; 第三,基于蒙特卡洛样本刻画估计量的经验分布。

在第一步中,该算法从样本量为 n 的总样本中无放回地抽取 s 个样本量为 b 的子样本。虽然第二步中通过自助法在每个子样本中有放回地抽取 r 个容量为 n 的蒙特卡洛样本以实现数据变异性的还原,但每个蒙特卡洛样本实际上是 b 个不同样本单元的加权组合,因此其计算复杂度仍为 $O(b)$。当子样本量 b 远小于总样本量 n 时,每个蒙特卡洛样本的计算成本可以大大降低。理论上,对于任意固定的子样本数 s,当总样本量 n 和子样本量 b 趋于无穷时,分治自助算法得到的不确定性度量具有一致性[3]。

在实证研究中,总样本量 n 由大数据本身决定,子样本量 b、子样本数 s 和蒙特卡洛自助样本数 r 需要由研究者确定。参数 b 和 r 的组合共同决定并行计算中单个任务的计算成本 $O(b) \times r$,而 s 决定了分治自助算法在该设定下的精度。加州大学伯克利分校教授 Michael Jordan 曾提出一个针对大数据推断的关键问题 "在给定计算成本时,随着数据量的增大,如何保证推断的精度?"[10] 严格来讲,分治自助算法虽然从理论和计算两方面回答了"如何在大数据量下保证推断精度"的问题,但没有给出其在"给定计算成本"情况下的答案。譬如,纵使 Kleiner 等[3]在理论上给出了该算法对不确定性估计的一致性,但当子样本量 b 趋于无穷时,算法再次面临计算成本过高的困境。即使子样本量 b 远小于总样本量 n,计算成本仍与蒙特卡洛样本数 r 强正相关。因此,在给定计算成本时,如何得到可以保证推断精度的参数组合,仍是分治自助算法研究需要讨论的开放问题。

### (二) 子集双重自助算法

与分治自助算法类似,子集双重自助算法[2]也包含三个关键步骤: 第一,针对总样本无放回地抽取子集以降低计算复杂度; 第二,针对子样本通过自助法实现数据变异性的还原,但每个子样本

表 1　　分治自助算法

输入: 数据 $X_n = (X_1, \cdots, X_n)$;
子样本的大小 b; 子样本的个数 s; 评价准则 $\xi(\cdot)$; 蒙特卡洛自助样本数 r
过程:
　for　j←1 to s
　从数据 $X_n = (X_1, \cdots, X_n)$ 中无放回抽取样本量为 b 的子样本数据 $X_{j,b}$
　　for　k←1 to r
　从数据 $X_{j,b}$ 中利用自助法有放回抽取样本量为 n 的蒙特卡洛样本 $X_{j,b}^k$
　计算样本 $X_{j,b}^k$ 的参数估计 $\hat{\theta}_{j,b}^k$
　　end
　　计算 $\hat{\theta}_{j,b} \leftarrow r^{-1} \sum_{k=1}^{r} \hat{\theta}_{j,b}^k$
　　计算 $\xi_j \leftarrow \xi(\hat{\theta}_{j,b})$
　end
计算 $\hat{\theta} \leftarrow s^{-1} \sum_{j=1}^{s} \hat{\theta}_{j,b}$
计算 $\xi(\hat{\theta}) \leftarrow s^{-1} \sum_{j=1}^{s} \xi_j$
输出: 加权估计 $\hat{\theta}$; 不确定性度量 $\xi(\hat{\theta})$

仅抽取一个蒙特卡洛样本; 第三,基于蒙特卡洛样本刻画估计量的经验分布。

对比表 1,子集双重自助算法可以看作分治自助算法在蒙特卡洛样本数 r = 1 时的特例。当给定计算成本时,分治自助算法把运算资源消耗在少量子样本的 r 个蒙特卡洛样本的重复计算中,而子集双重自助算法则把运算资源放在更多的子样本个数 s 上。换言之,子集双重自助算法在给定计算成本时通过覆盖更多的总样本单元来提升估计不确定性的精度,且这种优势随着总样本量 n 趋于无穷而逐渐明显。因此,子集双重自助算法在计算资源有限时对大数据实证研究者更具实际意义。特别地,Sengupta 等[2]证明了该算法的条件弱一致性,保证了子集双重自助算法在大数据下对不确定性度量的理论性质。

## 三、案例反思

上述算法拥有各自的优势与不足: 分治自助算法对不确定性的估计具有更好的一致性,但当计算资源有限时会陷入样本覆盖率相对较低的困境,适用于计算资源相对于样本量充足的情况; 子集双重自助算法虽然一致性较弱但最大程度上减少了蒙特卡洛样本对计算资源的消耗,在计算资源有限时通过提升样本覆盖率优化估计,适用于计算资源相对数据量不足的情况。

这类研究仍留给后续研究者很大的改进空间:

首先,Kleiner 等[3]和 Sengupta 等[2]虽然通过数值分析讨论了不同参数设定下分治自助算法和子集双重自助算法的估计效果,但均未深入探讨子样本量 b 的选择问题。大数据的实证研究者需要根据实际情况权衡精度和计算成本: 一方面,算法估计的精度随子样本量 b 的增大而提升; 另一方面,计算复杂度 $O(b)$ 也会随子样本量 b 的增大而增长。该问题的研究需要后续研究者从理论或计算的角度展开深入讨论。

其次,高维化也带来越来越大的挑战。以“国家重点研发计划”中“精准医学研究专项”为例,该项目“以我国常见高发、危害重大的疾病及若干流行率相对较高的罕见病为切入点”来“构建百万人以上的自然人群国家大型健康队列和重大疾病专病队列”。上述研究势必涉及与疾病相关的动辄上万甚至上百万维的高通量分子生物大数据分析。研究者应考虑同时在样本层面和变量层面自助抽样的思路[11],设计兼具变量选择和不确定性度量的大数据统计算法。

第三，子样本层面的有放回抽样是分治自助算法和子集双重自助算法设计的关键，通过在子样本中重复抽取容量为 n 的蒙特卡洛样本实现数据变异性的还原。不同的自助抽样方式在不同条件下具有各自的优势与不足，尤其是在利用自助法解决高维变量选择问题时。因此，选择合适的自助抽样很重要。

## 参考文献


[1] 林存洁，李扬. 大数据分析仍需要统计思想——以 ARGO 模型为例[J]. 统计研究，2016，33(11)：109－112.

[2] Sengupta S，S Volgushev and X Shao. A Subsampled Double Bootstrap for Massive Data[J]. Journal of the American Statistical Association，2016，111(515)：1222－1232.

[3] Kleiner A，A Talwalkar，P. Sarkar，et al. A Scalable Bootstrap for Massive Data[J]. Statistical Methodology，2014，76(4)：795－816.

[4] Bickel P J，F Götze and W R van Zwet. Resampling Fewer than n Observations：Gains，Losses，and Remedies for Losses[J]. Statistica Sinica，1997，7：1－31.

[5] Politis D N，J P Romano and M Wolf. Subsampling[M]. New York：Springer，1999.

[6] Davidson R and J G MacKinnon. Improving the Reliability of Bootstrap Tests with the Fast Double Bootstrap[J]. Computational Statistics & Data Analysis，2007，51(7)：3259－3281.

[7] Giacomini R，D N Politis and H White. A Warp-speed Method for Conducting Monte Carlo Experiments Involving Bootstrap Estimators[J]. Econometric Theory，2013，29(3)：567－589.

[8] Chang J and P Hall. Double-bootstrap Methods that use a Single Double-bootstrap Simulation[J]. Biometrika，2015，102(1)：203－214.

[9] Bickel P J and A Sakov. On the Choice of m in the m out of n Bootstrap and Confidence Bounds for Extrema[J]. StatisticaSinica，2008，18(3)：967－985.

[10] Jordan M I. On Statistics，Computation and Scalability[J]. Bernoulli，2013，19(4)：1378－1390.

[11] Fang K and S Ma. Analyzing Large Datasets with Bootstrap Penalization[J]. Biometrical Journal，2017，59(2)：358－376.


## 作者简介


李扬，男，2010 年毕业于中国人民大学统计学院，获经济学博士学位，现为中国人民大学应用统计科学研究中心研究员，中国人民大学统计学院副教授、博士生导师，国际统计学会推选会员，中国人民大学统计咨询研究中心主任。研究方向为决策与预测、生物医学大数据方法。

张长，男，现为中国人民大学统计学院在读硕士研究生。研究方向为函数型数据与机器学习算法。

朱建平，男，2003 年获南开大学理学博士学位，现任厦门大学管理学院 MBA 中心教授、博士生导师，厦门大学数据挖掘研究中心主任。研究方向为数理统计、数据挖掘。


（责任编辑：倪立行）