大数据统计方法综述

叶小青,汪政红,吴 浩
(中南民族大学 数学与统计学学院,武汉 430074)

摘 要 回顾大数据统计分析方法的现状,重点分析线性及非线性模型的分治算法,详细阐述 3 种抽样法,并比较其差异,归纳总结在线更新算法和基于变量选择的在线更新算法,最后展望大数据统计分析的未来.

关键词 大数据; 分治算法; 抽样法; 在线更新算法
中图分类号 O213 **文献标识码** A **文章编号** 1672-4321(2018)04-0151-06

A Survey on the Statistical Methods of Big Data

Ye Xiaoqing, Wang Zhenghong, Wu Hao
(College of Mathematics and Statistics, South-Central University for Nationalities, Wuhan 430074, China)

Abstract Reviewing the current status of statistical analysis methods of big data, the divide and conquer algorithms based on linear and nonlinear model are analyzed. Three sampling methods are described in detail and the differences between them are compared. The online update algorithms and online update algorithms based on variable selection are summarized. Finally, the main future challenges of statistical analysis of big data are discussed.

Keywords big data; divide and conquer algorithm; sampling method; online update algorithm

随着通讯和信息技术的高速发展,全球数据爆炸性地增长.面对铺天盖地的海量大数据,有效的数据分析与挖掘将推动国家、企业乃至整个社会的高效、可持续发展.值得强调的是,在大数据分析与挖掘任务中,统计分析的研究受到更为广泛的关注和重视.而大部分传统统计方法对大数据的研究存在局限性: 一是传统统计方法适合分析单个计算机存储的数据,无疑导致了数据处理和整合的困难; 二是传统统计方法难以适应大数据源的高速性和实时性等特点.因此,为了适应大数据这一新的研究对象,传统统计学必须进行改进,以更好地服务于人类.目前国内外将大数据和传统统计学相结合的研究文献并不多,使得大数据背景下统计分析的研究成为学术界关注的热点难点问题,其代表性文献集中于三大方面: 分治算法、抽样法和在线更新算法.

第一,分治算法是将初始大数据集分成适合当前计算管理能力的 K 个子集,先对每个子集做统计分析,然后综合 K 个子集的分析结果.分治算法通过对子集的平行计算可以缩减计算成本,但是,如何总结 K 个子集的估计结果,才能使最终的估计结果更有效? 为了得到最终有效的估计量,部分研究者提出了不同的估计方法,如 Li 等对大数据集的单参数回归模型提出了二阶段法,其研究结果表明二阶段法可以显著地降低计算成本,且估计量具有渐进正态性[1].Lin 和 Xi 对非线性回归方程的参数估计提出了有效的聚合估计(AEE, Aggregated Estimating Equation),研究结果显示 AEE 估计量具有一致性,而且能显著地缩减计算成本[2].在应用方面,AEE 估计方法适合于大型数据立方和数据流.Xu 等对广义回归方程提出了非参数分布核估计方法(DKR, Distributed Kernel Regression),其研究结论为: 在适当划分子样本集的条件下,DKR 估计方法具有一致性[3].Chen 和 Xie(2014) 运用惩罚似然函数估计各子集的广义线性回归方程,并利用多数表决法得到大数据集的最终估计量,研究结果表明估计量具有符号一致性[4].

第二,抽样法的基本思想是从初始大数据中随机提取子样本代替原始数据对模型进行估计、预测

收稿日期 2018-08-14
作者简介 叶小青(1979-),女,副教授,博士,研究方向: 应用统计学,E-mail: yshtim@126.com
基金项目 国家自然科学基金资助项目(11401596); 湖北省本科教研项目(JYS15012)

以及统计推断.抽样法的难点在于设计子样本的概率分布.最简单的概率分布是均匀分布,大量研究文献表明运用统计杠杆值作为子样本的概率分布优于均匀分布(Mahoney 等[5],Drineas 等[6]).Ma 和 Sun 也认为利用杠杆值作为子样本的概率分布能准确有效地提取大规模样本信息,并且从统计角度研究了基于杠杆值抽样算法估计量的性质[7].

第三,随着科学技术的不断普及,大数据的规模和程度不断地增大,具有实时更新特性.例如在银行的存款中,每天都有成千上万的客户利用自动取款机或人工服务进行交易,每一位客户对自己的银行账号进行操作,对于银行的整个数据系统来说是极小的一部分,面对这样实时在线更新的大数据,如何获得计算效率高、成本低的算法呢? Schifano 等扩展了 Lin 和 Xi 的理论方法,研究了广义非线性模型的回归系数和标准误的在线更新估计量,其研究结果显示在线更新估计量具有一致性,而且有限样本仿真模拟表明在线更新估计量具有较小的偏误[8].Wang 等基于 Schifano 研究结论首次提出了标准变量选择的在线更新模型,并根据 AIC、BIC 和 DIC 信息准则来选择最优变量[9].

从以上大数据统计分析的研究进展来看,尽管研究成果尚不丰富,仍处于起步阶段,但对现有成果的梳理与总结,可以为关注大数据统计分析的研究者、教学者提供借鉴.

1 分治算法理论

分治算法是将初始大数据集分成适合当前计算机管理能力的 K 个子集,先对每个子集做统计分析,然后综合 K 个子集的分析结果.下面详细介绍几种有效的分治算法.

1.1 二阶段法

Li 等提出了二阶段法,其基本思想:第一阶段将整个数据集划分为若干子样本,使得每个子样本都适合目前的计算机管理能力,估计每个子样本参数;第二阶段对每个子样本估计结果取平均值.

假设 $x_1,x_2,\cdots,x_n$ 是独立同分布样本,此处 x_i 可以是随机变量或随机向量.为了清晰表达二阶段估计算法,将样本表示如下:

$$\begin{bmatrix} x_{11} & \cdots & x_{1\alpha_n} \\ \vdots & \ddots & \vdots \\ x_{\beta_n 1} & \cdots & x_{\beta_n\alpha_n} \end{bmatrix},$$

这里,$x_{ji}=x_{(j-1)\cdot\alpha_n+i}$,$i=1,2,\cdots,\alpha_n$ 和 $j=1,2,\cdots,\beta_n$,β_n 是子样本数,α_n 是子样本容量,$n=\alpha_n\beta_n$.Li 等建议 $\alpha_n=O(\sqrt{n}\log\log(n))$,对每个子样本运用相同的估计方法,基于第 j 个子样本 $x_{j1},\cdots,x_{j\alpha_n}$ 的估计值标记为 $\hat{\theta}_j$.对 $\hat{\theta}_j(j=1,2,\cdots,\beta_n)$ 取其平均值得到全样本参数 θ 的估计值:$\bar{\theta}=\frac{1}{\beta_n}\sum_{j=1}^{\beta_n}\hat{\theta}_j$,$\bar{\theta}$ 具有一致正态性.

1.2 AEE 估计法

Lin 和 Xi 研究了广义非线性模型的参数估计问题,假设独立样本观测值 $\{z_i,i=1,2,\cdots,N\}$,对某个得分函数 φ 满足 $\sum_{i=1}^{N}E[\varphi(z_i,\beta_0)]=0$,其中,$\beta_0\in R^p$ 为未知参数真值.由样本方程 $\sum_{i=1}^{N}\varphi(z_i,\beta)=0$ 计算出 β_0 的估计值记为 β_N.如果 N 偏大,β_N 计算成本越大.针对此缺点,Lin 和 Xi 提出了 AEE 估计方法,主要计算步骤如下:将初始大数据集分为 K 个子样本,并假设每个子样本容量为 n,第 k 子样本集记为 $z_{k1},z_{k2},\cdots,z_{kn}$.由方程 $M_k(\beta)=\sum_{i=1}^{n}\varphi(z_{ki},\beta)=0$ 确定第 k 个子样本参数估计值为 $\hat{\beta}_{nk}$.结合 K 个子集的结果,构造的大数据样本 AEE 估计量 $\tilde{\beta}_{NK}$ 为:

$$\tilde{\beta}_{NK}=\Big(\sum_{k=1}^{K}A_k\Big)^{-1}\sum_{k=1}^{K}A_k\hat{\beta}_{nk},$$

$$A_k=-\sum_{i=1}^{n}\frac{\partial\varphi(z_{ki},\hat{\beta}_{nk})}{\partial\beta}.$$

Lin 和 Xi 的研究结果显示 $\tilde{\beta}_{NK}$ 具有一致性.

1.3 DKR 估计法

Xu 等提出了非参数分布核回归(DKR)方法来估计参数方程,该方法不仅具有普适性,而且不依赖于任何真实模型的参数假设.

设 $Y\in[-M,M]\subset R$ 是因变量,其界限 $M>0$,X 是解释变量且 $X\in\kappa\subset R^d$,κ 为 d 维空间一紧集.总体 $Z=[-M,M]\times\kappa$ 分布未知,且设 $S=\{z_i=(y_i,x_i),i=1,2,\cdots,N\}$ 是来自 Z 的 N 个独立样本观测值.记 $f:\kappa\to R$ 为 X 与 Y 之间的潜在函数关系.在大数据背景下,如何估计函数关系 f,非参数分布核回归算法具体如下:

1) 假设 S 平均随机分成 m 份,每份样本容量为 $n=N/m$,m 个子样本标识分别为 $S_1,S_2,\cdots,S_m$;

2) 记 $T_M[\cdot]$ 为阈值 M 的截取算子,基于子样本 S_j,得 f_j 估计值为:

$\hat{f}_j = T_M[f_j]$,$f_j = \arg\min_{f\in H_K}\{\frac{1}{n}\sum_{z_i\in S_j} l(f,z_i) + \lambda \|f\|^p\}$,

此处,$l(\cdot)$ 为非负的损失函数,$\frac{1}{n}\sum_{z_i\in S_j} l(f,z_i)$ 是样本风险函数,$\|\cdot\|$是范数,$\lambda \geqslant 0$ 是正则化参数. 记 K: $\kappa \times \kappa \to R$ 为连续、对称、半正定核函数, $H_k = \overline{span}\{K(x,\cdot), x \in \kappa\}$ 是由 K 产生的L^2 可积函数的 Hilbert 空间;

3) 基于各个子样本的估计值$\hat{f}_j, j = 1,2,\cdots,m$,可得全样本$f$的估计方程$\bar{f} = \frac{1}{m}\sum_{j=1}^{m}\hat{f}_j$,Xu 等的研究结果表明 DKR 估计量具有一致性.

1.4　多数表决法

Chen 和 Xie 为广义线性模型的参数估计提出了一种分治算法,其估计思想为基于子样本的似然函数,加入惩罚项,称为广义似然函数,并最大化广义似然函数估计子样本参数,最后运用多数表决法得出大样本数据的综合估计量,通过仿真和数据实例表明该方法能极大地缩减计算时间和计算存储空间.

考虑如下的广义线性模型: $E(y_i) = g(x_i^{\mathrm{T}}\beta)$,$i = 1,2,\cdots,n$,其中,y_i 为因变量,x_i 是 $p \times 1$ 维解释变量,β 是 $p \times 1$ 维未知参数,g 为关联函数.

假设在给定 $X = (x_1,\cdots,x_n)^{\mathrm{T}}$ 的情况下,$y = (y_1,\cdots,y_n)^{\mathrm{T}}$ 条件分布为典型指数分布,那么似然函数为:

$$f(y;X,\beta) = \prod_{i=1}^{n} f_i(y_i;\theta_i) = \prod_{i=1}^{n}\{c(y_i)\exp[y_i\theta_i - b(\theta_i)/\varphi]\},$$

此处,$\theta_i = x_i^{\mathrm{T}}\beta, i = 1,2,\cdots,n$,$\varphi$ 代表方差参数. 对数似然函数 $\ln f(y;X,\beta)$ 为:

$$l(\beta;y,X) = \frac{y^{\mathrm{T}}X^{\mathrm{T}}\beta - 1^{\mathrm{T}}b(X^{\mathrm{T}}\beta)}{n},$$

其中,$b(\theta) = [b(\theta_1), b(\theta_2),\cdots,b(\theta_n)]^{\mathrm{T}}$ 且 $\theta = (\theta_1,\theta_2,\cdots,\theta_n)^{\mathrm{T}}$,函数 $b(\cdot)$ 二阶可导.

假设初始大样本数据分为 K 个子集,第 k 个子集有n_k 个观测值$(x_{ki}, y_{ki}), i = 1,2,\cdots,n_k$. 记$y_k = (y_{k1},y_{k2},\cdots,y_{kn_k})^{\mathrm{T}}$,$X_k = (x_{k1}^{\mathrm{T}},x_{k2}^{\mathrm{T}},\cdots,x_{kn_k}^{\mathrm{T}})^{\mathrm{T}}$,基于第 k 个子块的似然函数为:

$$l(\beta;y_k,X_k) = \frac{y_k^{\mathrm{T}}X_k^{\mathrm{T}}\beta - 1^{\mathrm{T}}b(X_k^{\mathrm{T}}\beta)}{n_k},$$

相应地,第 k 个子块的惩罚估计量为:

$$\hat{\beta}_k = \arg\max_{\beta}\{\frac{l(\beta;y_k,X_k)}{n_k} - \rho(\beta;\lambda_k)\},$$

此处,$\rho(\beta;\lambda_k)$ 是调节参数为λ_k 的惩罚函数. 基于$\hat{\beta}_k, k = 1,2,\cdots,K$,使用多数表决法得到最后大样本的综合估计量.

记$\hat{A}^c \overset{\mathrm{def}}{=} \{j: \sum_{k=1}^{K} I(\hat{\beta}_{k,j} \neq 0) > \omega\}$ 为变量选择集,其中 $\omega \in [0,K)$ 是给定的临界值,I 为示性函数. 记 $E = \mathrm{diag}(v_1,\cdots,v_p)$ 为 $p \times p$ 的投票矩阵,当 $\sum_{k=1}^{K} I(\hat{\beta}_{k,j} \neq 0) > \omega$ 时,$v_j = 1$; 否则$v_j = 0$; Chen 和 Xie 给出大样本数据 β 综合估计量为:

$$\hat{\beta}^{(c)} \overset{\mathrm{def}}{=} A\left(\sum_{k=1}^{K} A^{\mathrm{T}}\{X_k^{\mathrm{T}}\Sigma(\hat{\theta}_k)X_k\}A\right)^{-1} \cdot \sum_{k=1}^{K} A^{\mathrm{T}}\{X_k^{\mathrm{T}}\Sigma(\hat{\theta}_k)X_k\}A\hat{\beta}_{k,\hat{A}^{(c)}},$$

其中,$\hat{\theta}_k = X_k\hat{\beta}_k$,$\Sigma(\hat{\theta}_k)$ 为方差协方差矩阵,$A = E_{\hat{A}^{(c)}}$ 为 E 的子矩阵,其维度为 $p \times \lceil \hat{A}^{(c)} \rceil$. Chen 和 Xie 的研究结果显示$\hat{\beta}^{(c)}$ 具有符号一致性.

2　抽样法

抽样法基本思想是从初始数据中提取伴随一定概率分布的子样本代替原始大数据对模型进行估计、预测以及统计推断. 抽样法的难点在于对各子样本概率分布的设计. 下面以经典线性回归模型为例,阐述几种典型的抽样法.

假设 $y = X\beta + \varepsilon$,y 是 $n \times 1$ 向量,X 是 $n \times p$ 维矩阵,包含截距项和 $p - 1$ 个解释变量,β 是 $p \times 1$ 系数向量,ε 为服从多元正态分布的残差项,系数向量 β 的 OLS 估计量为:

$$\hat{\beta}_{OLS} = \arg\min_{\beta}\|y - X\beta\|^2 = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y. \quad (1)$$

矩阵 $X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ 通常被记为 H,称为帽子矩阵,且有$\hat{y} = Hy$.

2.1　一般抽样法

抽取 r 个随机样本,样本观测值记为 $\{y_i^*, X_i^*\}_{i=1}^{r}$,相应的抽样概率为: $\{\pi_i^*\}_{i=1}^{r}$. 那么基于抽样样本的加权估计量为:

$$\tilde{\beta} = \arg\min_{\beta}(y^* - X^*\beta)^{\mathrm{T}}W(y^* - X^*\beta), \quad (2)$$

其中权重矩阵 $W = \mathrm{diag}(\{\pi_i^*\}_{i=1}^{r})$.

Ma 等的研究结果显示$\tilde{\beta}$ 具有无偏性. 特别地,如

果抽样概率忽略不同子样本分布的差异性,都设置为均匀分布,也即$\pi_i = 1/n, i = 1,2,\cdots,n$. WLS 估计便退化为 OLS 估计,其估计量为:

$$\widetilde{\beta}_{UNIF} = \arg\min_{\beta} \|y^* - X^*\beta\|^2.$$

$\widetilde{\beta}_{UNIF}$ 称为一致性估计. Meng 等发现,该一致性估计算法具有计算时间成本少、无偏性的优点,但不足的是估计量方差偏大[10].

2.2 基本杠杆抽样法

杠杆值抽样法(Basic Leverage Sampling Method(BLSM))基本思想是选取对回归线具有影响的样本点. 最初是 Weisberg 提出杠杆值的概念[11]. 杠杆值越大,其影响越大.

对第 i 个数据点(y_i, X_i),我们定义杠杆值为$\frac{\partial \hat{y}_i}{\partial y_i}$. 如果杠杆值越大,则意味着$y_i$ 小的扰动会导致$\hat{y}_i$比较大的改变. 由$\hat{y} = Hy$可知杠杆值计算公式为:$\frac{\partial \hat{y}_i}{\partial y_i} = \frac{\partial(\sum_{j=1}^{n} h_{ij}y_i)}{\partial y_i} = h_{ii}$,$h_{ii}$ 为矩阵 H 对角线上的第 i 个元素. 记$\sum_{i=1}^{n} h_{ii} = p$,定义抽样概率公式为:$\{\pi_i^{BLSM}\}_{i=1}^{n} = \left\{\frac{h_{ii}}{p}\right\}_{i=1}^{n}$,相应于子样本$\{y_i^*, X_i^*\}_{i=1}^{r}$ 的概率为$\{\pi_i^{BLSM}\}_{i=1}^{r}$,代入式(2),得到最终的加权估计量记为$\widetilde{\beta}_{BLSM}$.

Meng 等研究发现当不同维度的数据集具有不同的分布时,BLSM 方法将不利于获得高倍影响点. 为了克服这个缺点,下面提出缩减杠杆值法.

2.3 缩减杠杆值法

缩减杠杆值法(Shrinkage Leveraging Method (SLM))综合了均匀分布概率法和基本杠杆值得分法,其表达式为这两者的线性组合:

$$\pi_i^{SLM} = \alpha\pi_i^{BLSM} + (1-\alpha)\pi_i^{UNIF}; \alpha \in (0,1), i = 1,2,\cdots,n.$$

利用这种样本概率计算出的估计量记为$\widetilde{\beta}_{SLM}$. SLM 的性质与缩减指数α息息相关. 如果选择α非常靠近0或者1,将退化为一致性均匀分布样本方法或BLSM,如果 α 适当,将克服大方差问题. 文[12]中,Ma 等推荐 α 最优值为 $[0.8, 0.9]$. 通过式$\pi_i^{SLM} = \alpha\pi_i^{BLSM} + (1-\alpha)\pi_i^{UNIF}$可知,SLM 估计方法能保持杠杆值的优势,在大数据线性模型中最受欢迎.

3 在线更新算法

3.1 在线更新一般算法

Schifano 等扩展了 Lin 和 Xi 的理论方法,研究了广义非线性模型的回归系数和标准误在线累积更新估计量. 假设初始大样本数据分成若干个时点,第 k 个时点子样本集有n_k 个观测值(x_{ki}, y_{ki}),$i = 1,2,\cdots,n_k$. 为了有效利用当前时点 k 及以前的信息量,Schifano 等先给出过渡估计量递推公式:$\breve{\beta}_{n_k,k} = (\widetilde{A}_{k-1} + A_{n_k,k})^{-1}\left(\sum_{l=1}^{k-1}\widetilde{A}_{n_l,l}\breve{\beta}_{n_l,l} + A_{n_k,k}\hat{\beta}_{n_k,k}\right)$. $k = 1,2,\cdots$,$\widetilde{A}_0 = 0_p$,$\breve{\beta}_{n_0,0} = 0$ 且$\widetilde{A}_k = \sum_{l=1}^{k}\widetilde{A}_{n_l,l}$,$\widetilde{A}_{n_k,k} = [A_{n_k,k}(\breve{\beta}_{n_k,k})] = -\sum_{i=1}^{n_k}\frac{\partial\varphi(z_{ki},\breve{\beta}_{n_k,k})}{\partial\beta}$,沿用 Lin 和 Xi 的表达方式,$\varphi$ 为得分函数,$\hat{\beta}_{n_k,k}$ 为方程$M_k(\beta) = \sum_{i=1}^{n_k}\varphi(z_{ki},\beta) = 0$的解. 过渡估计量$\breve{\beta}_{n_k,k}$结合了中间估计量$\breve{\beta}_{n_l,l}(l = 1,2,\cdots,k-1)$和当前估计量$\hat{\beta}_{n_k,k}$.

基于过渡估计量$\breve{\beta}_{n_k,k}$,Schifano 等给出回归系数的累积更新估计量$\widetilde{\beta}_k$:

$$\widetilde{\beta}_k = (\widetilde{A}_{k-1} + \widetilde{A}_{n_k,k})^{-1} \cdot (a_{k-1} + \widetilde{A}_{n_k,k}\breve{\beta}_{n_k,k} + b_{k-1} + M_{n_k,k}(\breve{\beta}_{n_k,k})), \quad (3)$$

其中,

$$a_k = \sum_{l=1}^{k}\widetilde{A}_{n_l,l}\breve{\beta}_{n_l,l} = \widetilde{A}_{n_k,k}\breve{\beta}_{n_k,k} + a_{k-1},$$

$$b_k = \sum_{l=1}^{k}M_{n_l,l}(\breve{\beta}_{n_l,l}) = M_{n_k,k}(\breve{\beta}_{n_k,k}) + b_{k-1},$$

$$a_0 = b_0 = 0, \widetilde{A}_0 = 0_p, k = 1,2,\cdots.$$

为了计算$\widetilde{\beta}_k$的方差,观测到:

$$0 = -M_{n_k,k}(\hat{\beta}_{n_k,k}) \approx -M_{n_k,k}(\breve{\beta}_{n_k,k}) + \widetilde{A}_{n_k,k}(\hat{\beta}_{n_k,k} - \breve{\beta}_{n_k,k}),$$

由此可以得到:

$$M_{n_k,k}(\breve{\beta}_{n_k,k}) + \widetilde{A}_{n_k,k}\breve{\beta}_{n_k,k} \approx \widetilde{A}_{n_k,k}\hat{\beta}_{n_k,k}.$$

利用这个近似关系,式(3)左右两边取方差,得到方差的累积更新估计量为:

$$\widetilde{V}_k = (\widetilde{A}_{k-1} + \widetilde{A}_{n_k,k})^{-1} \cdot$$

$$\left(\sum_{l=1}^{k}\widetilde{A}_{n_l,l}\widehat{V}_{n_l,l}\widetilde{A}_{n_l,l}^{\mathrm{T}}\right)[(\widetilde{A}_{k-1}+\widetilde{A}_{n_k,k})^{-1}]^{\mathrm{T}}=$$

$$(\widetilde{A}_{k-1}+\widetilde{A}_{n_k,k})^{-1}(\widetilde{A}_{k-1}\widetilde{V}_{k-1}\widetilde{A}_{k-1}^{\mathrm{T}}+\widetilde{A}_{n_k,k}\widehat{V}_{n_k,k}\widetilde{A}_{n_k,k}^{\mathrm{T}})\cdot$$

$$[(\widetilde{A}_{k-1}+\widetilde{A}_{n_k,k})^{-1}]^{\mathrm{T}}, \quad (4)$$

其中,

$$\widehat{V}_{n_k,k}=var(\widehat{\beta}_{n_k,k}),k=1,2,\cdots,\text{且}\widetilde{A}_0=\widetilde{V}_0=0_p.$$

Schifano 等的研究结果表明$\widetilde{\beta}_k$ 和$\widetilde{V}_k$ 具有一致性,而且有限样本仿真模拟显示该估计量更精确,偏误更小.

相较于 Lin 和 Xi (2011),Schifano 等做了两个方面的改进: 1) 考虑了回归参数的方差更新估计量,如式(4) 所示. 基于此,可以进一步研究回归模型真实参数的统计推断问题; 2) 当新数据流来临时,更新估计量$\widetilde{\beta}_k$ 和$\widetilde{V}_k$ 不需要存储历史数据集,只需历史估计量和当前时点 k 的数据集,减小了存储空间,提高了计算效率,缩减了计算成本.

3.2　基于标准变量选择的在线更新估计量

Wang 等首次提出了基于标准变量选择的在线更新模型. 假设将大数据集分为 K 个子集,第 k 子块数据记为(Y_k,X_k),样本容量为n_k,其解释变量观测矩阵为$n_k\times(p+1)$,p 为解释变量个数,$k=1,2,\cdots,K$. 整个样本观测矩阵记为: $Y=(Y_1^{\mathrm{T}},Y_2^{\mathrm{T}},\cdots,Y_K^{\mathrm{T}})^{\mathrm{T}}$,$X=(X_1^{\mathrm{T}},X_2^{\mathrm{T}},\cdots,X_K^{\mathrm{T}})^{\mathrm{T}}$,总样本容量为 $n=\sum_{k=1}^{K}n_k$. 考虑标准线性回归方程: $Y=X\beta+\varepsilon$. 记 M 为回归模型集,因每个解释变量均有选择的可能性,所以模型维度为 2^p,$m=1,2,\cdots,2^p$. 基于第 k 子样本块的最小二乘系数估计量和误差平方和分别记为$\widehat{\beta}_{n_k,k}$ 和$SSE_{n_k,k}$.

记$\widehat{\beta}_k^{(m)}=(\beta_0^{(m)},\beta_1^{(m)},\cdots,\beta_{p_m}^{(m)})^{\mathrm{T}}$ 和$SSE_k^{(m)}$ 为截至时点 k,模型 m 的累积估计量,p_m 为模型 m 的解释变量的个数. 下面给出基于变量选择的在线更新样本估计量$\widehat{\beta}_k^{(m)}$ 和$SSE_k^{(m)}$. 首先,引入$(p+1)\times(p_m+1)$ 维度选择矩阵$p^{(m)}=(e_{m_0},e_{m_1},\cdots,e_{m_{p_m}})$,其中,$e_{m_j}$ 为 $p+1$ 维列向量,e_{m_0} 第一个元素为 1,$e_{m_j}(j\neq 0)$ 第 m_j 元素为 1,其它位置为 0. 此处 $(m_1,\cdots,m_{p_m})$ 是选择变量指标集. 定义 $X_k^{(m)}=X_k p^{(m)}$,更新$(p_m+1)\times(p_m+1)$ 维矩阵:

$$V_k^{(m)}=X_k^{(m)\mathrm{T}}X_k^{(m)}+V_{k-1}^{(m)},$$

其中,$V_0^{(m)}=0$. Wang 等给出$\beta_k^{(m)}$ 和$SSE_k^{(m)}$ 的在线更新样本估计量递推公式为:

$$\widehat{\beta}_k^{(m)}=(V_k^{(m)})^{-1}A_k^{(m)}, \quad (5)$$

$$SSE_k^{(m)}=SSE_{n_k,k}+\widehat{\beta}_{n_k,k}^{\mathrm{T}}X_k^{\mathrm{T}}X_k\widehat{\beta}_{n_k,k}+\widehat{\beta}_{k-1}^{(m)\mathrm{T}}\cdot$$

$$V_{k-1}^{(m)}\widehat{\beta}_{k-1}^{(m)}-\widehat{\beta}_k^{(m)\mathrm{T}}V_k^{(m)}\widehat{\beta}_k^{(m)}+SSE_{k-1}^{(m)}, \quad (6)$$

其中,$A_k^{(m)}=X_k^{(m)\mathrm{T}}X_k\widehat{\beta}_{n_k,k}+V_{k-1}^{(m)}\widehat{\beta}_{k-1}^{(m)}$,$\widehat{\beta}_0^{(m)}=0$,由式(5) 和式(6) 可以看出基于变量选择的在线更新估计量基于历史估计量$\widehat{\beta}_{k-1}^{(m)}$ 和$V_{k-1}^{(m)}$,不需要储存所有的原始大数据集,可以节省存储空间,提高计算效率. Wang 等进一步基于 AIC、BIC 和 DIC 信息准则来选择变量,并通过具体仿真模拟显示: 当解释变量之间不是高度相关时,BIC 准则相对于 AIC、DIC 较好.

4　研究展望

大数据表现出的高维性、海量性和实时性等特征,为大数据统计分析提出了严峻的挑战. 围绕这些挑战,分治算法、抽样法和在线更新算法已经取得重要进展,但鉴于大数据的复杂特征,仍有诸多亟待探索与解决的难题.

(1) 模型精确性和计算效率的权衡. 在大数据分析中,为了得到更精确的结果,通常不仅需要更复杂的模型,而且需要更多的计算资源,往往造成计算极其复杂,计算成本偏高,计算效率偏低. 那么如何在计算精准性和计算效率之间权衡是我们面临的一大问题. 这就需要为简单模型设计有效的算法,也即在大数据背景下,使得简单模型具有更加良好的表现. 例如,Google 公司翻译和语音识别系统,由于大数据的可获性,显著提高了传统经典模型优良精准的特征. 这个例子说明传统简单模型在大数据背景下可能具有更精确的表现. 那么,如何获得得益于大数据的简单准确模型将是一个巨大的挑战;

(2) 排序问题. 从海量大数据中筛选出最有价值的信息极其重要. 此类统计排序问题目的是将最重要或关联最强的信息排在最前面. 一般来说,最重要的信息排列在前,那么该排序算法最好. 研究者面临如何设计统计意义上合理的量度,来衡量排序的质量,以及后续的对统计推断的研究(Duchi 等[13]);

(3) 尾部特征分析. 在传统经济学中,概率 0.05 会被认为是可以忽略的稀有事件. 然而,在大数据背景下,这些所谓的稀有事件可能频繁发生,将引起特别的关注. 这就意味着尾部行为特征分析也将成为大数据问题的巨大挑战之一;

(4)大规模优化问题.众所周知,在建模中,优化起着至关重要的作用.例如最大似然估计法就是解决优化问题的常规方法.在大数据应用中,一个重要的方向便是在线优化算法.对处理实时更新的大数据流,在线优化算法应具备两个基本特征:1)在线优化算法不仅能减小样本数据的存储量,而且能存储和实时更新模型的估计参数;2)在线优化算法同时兼顾模型拟合和模型选择.比如,当新数据流来临时,模型拟合和模型选择能同时更新吗?模型选择参数能同时发生调整吗?这些是在线优化算法需要考虑的研究课题;

(5)因果推断模型.在大数据背景下,因果推断问题将面临极大的机遇与挑战.主要应用方面有:气候变化问题、医疗健康疗效比较研究以及行为经济学等问题.例如,给定100万人的电子健康记录,我们哪些药品在哪些方面有疗效?目前这些因果推断方面的研究在大数据背景下少有关注.

综合来看,尽管大数据统计分析的研究成果尚处于起步阶段,目前仅仅进行了一些初步探索.但是,围绕海量性、高速性及在线更新实时性等特征的研究文献已经为大数据统计分析提供了一个基本的研究框架,为更多有价值研究的不断涌现奠定了重要基础.

5 总结

本文在对现有大数据统计分析研究文献进行归纳和总结的基础上,具体从分治算法、抽样算法和在线更新算法三方面阐述分析,希望能够为关注大数据统计分析理论与应用的研究者与实践者提供参考.可以预见,在未来的大数据研究中,具有快捷、清晰、高效探测事物内在关系和规律的大数据统计分析领域将涌现出大量的重要研究成果.

参 考 文 献

[1] Li R,Dennis K J L,Li B. Statistical inference in massive data sets [J]. Applied Stochastic Models in Business and Industry,2013,29(1):399-409.

[2] Lin N,Xi R. Aggregated estimating equation estimation [J]. Statistics and Its Interface,2011,4:73-83.

[3] Xu C,Zhang Y,Li R,et al. On the feasibility of distributed kernel regression for big data [J]. IEEE Transactions on Knowledge and Data Engineer,2016,28(1):3041-3052.

[4] Chen X,Xie M. A split and conquer approach for analysis of extraordinarily large data [J]. Statistica Sinica,2014,24:1655-1684.

[5] Mahoney M W,Drineas P. CUR matrix decompositions for improved data analysis [J]. Proceedings of the National Academy of Sciences,2009,106(3):697-702.

[6] Drineas P,Mahoney M W,Muthukrishnans S,et al. Faster least squares approximation [J].Numerische Mathematik,2011,117:219-249.

[7] Ma P,Sun X. Leveraging for big data regression [J]. Wiley Interdisciplinary Reviews: Computational Statistics,2015,7(1):70-76.

[8] Schifano E D,Wu J,Wang C,et al. Online updating of statistical inference in the big data setting [J]. Technometrics,2016,58(3):393-403.

[9] Wang C,Chen M,Elizabeth S,et al. Statistical methods and computing for big data [J]. Stat Interface,2016,9(4):399-414.

[10] Meng C,Wang Y,Zhang X,et al. Effective statistical methods for big data analytics [C] //Saha S,Mandalis A. Handbook of Research on Applied Cybernetics and Systems Science. Hershey: IGI Global,2017: 280-299.

[11] Weisberg S. Applied linear regression [M]. Hoboken, New Jersey: John Wiley & Sons,2005:169-174.

[12] Ma P,Mahoney M W,Yu B. A statistical perspective on algorithmic leveraging [J]. Journal of Machine Learning Research,2015,16:861-911.

[13] Duchi J,Mackey L,Jordan M I. The asymptotics of ranking algorithms [J]. The Annals of Statistics,2013,41(5):2292-2323.

(责任编辑 曹 东)