重采样方法与机器学习

毕 华 梁洪力 王 珏
(中国科学院自动化研究所复杂系统与智能科学重点实验室 北京 100190)

摘 要 Boosting 算法试图用弱学习器的线性组合逼近复杂的自然模型,以其优秀的可解释性和预测能力,得到了计算机界的高度关注.但 Boosting 只被看作是一种特定损失下的优化问题,其统计学本质未曾得到充分的关注.作者追根溯源,提出从统计学角度看待 Boosting 方法:在统计学框架下,Boosting 算法仅仅是重采样方法的一个有趣的特例.作者希望改变计算机科学家只重视算法性能忽略数据性质的现状,以期找到更适合解决"高维海量不可控数据"问题的方法.

关键词 重采样;自助法;Boosting;机器学习
中图法分类号 TP18 **DOI** 号: 10.3724/SP.J.1016.2009.00862

Resampling Methods and Machine Learning

BI Hua LIANG Hong-Li WANG Jue
(*Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract In Boosting algorithm complex natural model is approximated by the linear combination of weak learners. Due to its excellent interpretability and prediction power, Boosting has become an intensive focus in computer science field. However, it is only considered as an optimizing procedure with a specific loss function, whose nature in statistics has never obtained sufficient attention. In essence, a statistical perspective of Boosting algorithm is brought out in this paper, i.e., an interesting special case of resampling methods. The authors hope the current situation of excessive attention being paid to the performance of algorithm while the characteristic of data being ignored will be changed, such that the tasks of "high dimensional and large volume data generated in an uncontrolled manner" could be tackled more appropriately.

Keywords resampling; bootstrap; Boosting; machine learning

1 引 言

1984 年,Valiant[1] 在他的论文中提出机器学习的另类理念.他认为,学习模型无需绝对精确,只需概率近似正确(Probably Approximately Correct, PAC)即可.由此,他建立了 PAC 的理论基础.这个理论可以简单描述如下:令 $F(x)$是自然模型,$f(x)$是从样本集学习后建立的模型,$|F(x)-f(x)|\leq\varepsilon$ 以概率 $1-\delta$成立.这里的关键是"概率 $1-\delta$成立"而不是以概率 1 成立.这个理论对 Vapnik 建立有限样本统计机器学习理论有重要的意义.Kearns 和 Valiant[2-3] (1988, 1994)在 PAC 的基础上,提出弱可学习的理论.他这样描述一个概念是弱可学习:$F(x)$与 $f(x)$定义如上,$|F(x)-f(x)|\leq\varepsilon$ 成立的概率大于$(1/2)+\delta,0\leq\delta\leq1/2$.这意味着,一个概念如

收稿日期: 2008-12-26; 最终修改稿收到日期: 2009-04-03. 本课题得到国家"九七三"重点基础研究发展规划项目基金(2004CB318103)资助.毕 华,女,1979 年生,博士研究生,主要研究方向为机器学习、数据挖掘、统计学习. E-mail: hua.bi@ia.ac.cn. 梁洪力,男,1980 年生,博士研究生,主要研究方向为粗糙集理论、数据挖掘与机器学习. 王 珏,男,1948 年生,研究员,博士生导师,主要研究领域为机器学习.

果是弱可学习的,那么只要求一个弱可学习算法产生的模型的精度高于 50%,也就是比随机猜想稍好.同时他将满足 PAC 原始定义的概念可学习称为强可学习.进而,他问了如下一个问题,强可学习在什么条件下与弱可学习等价.

1990 年,Schapire[4] 回答了这个问题.他使用构造的方法证明:一个概念弱可学习的充要条件是这个概念强可学习.这是一个有些"不可思议"的结论.正是由于这个定理,开始了至今还在人们关注视野中的一类机器学习的研究,机器学习研究者将这类学习方式称为集群学习(ensemble learning)[5].从此以后,统计学家开始介入机器学习的研究.这是本文讨论的重点,我们将在本文以后部分详细说明统计学家对这个问题的描述.以后 Freund 和 Schapire 提出了 Adaboost 算法[6],由于这个算法如此简单且灵活,立即受到计算机科学技术界的推崇.特别是,人们在使用这个算法时,发现很少出现"过学习(overfitting)"现象,这个性质大大超出了人们的期望,并派生了研究这个问题的很多课题[7].

事实上,尽管 Schapire 的定理是基于 PAC 理论,但是,其使用的构造性证明方法与统计学家 Efron 比他早十余年发展的属于重采样方法的自助法(Bootstrap)没有本质区别.如果说有区别,那也只是 Schapire 的方法使用了一种特殊的采样策略.目前,大多数计算科学家称其为"富信息"策略,即,一个弱学习器不能很好学习的样本(概念),将尽可能成为下一个弱学习器着重学习的样本.这就是 Adaboost 的原理.这里,需要指出,由于机器学习强调从样本学习获得模型,因此,"概念可学习"是指一个基于样本表述的概念,可以通过学习获得一个可以在 PAC 意义下概括这个概念的模型,即,可以概括样本的模型.

为了与统计学重采样方法相比较,我们首先简单描述 Schapire 的方法:给定一个样本集,它是与自然模型同分布且独立采样获得.首先,假设样本集上每个样本对模型的贡献是相同的,即每个样本具有相同的权重,使用一个学习算法,建立一个模型;然后,根据这个模型改变样本集上样本的权重(对样本重新排序),使得在新的样本集中,不能被这个模型正确概括的样本具有较大的概率;使用这个样本集再次学习,获得另一个模型;重复这个过程,直到满足停止准则,过程结束.由于每个模型只可以正确概括部分样本,故称其为弱模型.然后,使用投票方式将它们集群,构成强模型.由于投票方式可以描述为加性模型形式,这也就是为什么 Adaboost 算法被认为是一种"集群学习"的由来.从这个过程来看,特别是基于给定样本集的采样方式,其主要贡献是解决算法设计复杂性问题,尤其是针对非线性问题的算法设计.如果我们使用的每个学习算法是线性的话,上述的学习过程就有些类似分段线性的思想.

对机器学习来说,这个方法涉及 4 个重要的要素:样本采集、采样策略、算法类型、集群方法.这 4 个要素将是本文展开讨论的线索.

与 Schapire 以及他的合作者发表他们的研究结果的同时,统计学界也开始从模型角度关注重采样方法.Breiman 很快发表了他设计的方法——Bagging[8] (Bagging Predictors).这个方法与 Adaboost 方法相比较,解决算法复杂性的意图大大降低,统计学的痕迹更为清晰.正是由于其目的与计算机科学家有区别,因此,这项研究没有像 Adaboost 那样受到计算机学界的关注.在算法类型和集群方法两个方面,Bagging 与 Adaboost 没有任何区别,它们最大区别是在于"采样策略".具体地说,Bagging 沿袭了经典重采样方法——随机采样策略,而 Adaboost 则使用"富信息"策略.这个差别导致了"样本采集"步骤不同,后者暗示,"富信息"策略一定是基于满足独立同分布的当前给定样本集,否则"重采样"过程就没有"富信息"一说了,而前者则没有这个限制,它暗示的样本集既包含已经观测到的样本,也包含以后可能被观测到样本.显然,对关注从样本集通过算法设计,建立模型的计算机学界,后者更具有吸引力.由于在原理上,这些方法没有本质的区别,因此,目前在重采样意义下,大家仍沿用 Schapire 对其的称谓,将这类方法统称为 Boosting 方法.

从自然模型或样本集多次采样建模的角度来看,重采样方法已经有很长的历史(见本文第 2 节).然而,将重采样方法引入传统统计学的研究应该是 Quenouille[9] 于 1949 年提出的"刀切法"(Jackknife).但是,真正包含样本采集、采样策略、算法类型、集群方法 4 个要素,而目前最具影响力的重采样方法则是 Efron[10] 在 1979 年提出的"自助法".

高维海量不可控数据的涌现,对统计学是一个挑战,算法复杂性也已成为统计学家不得不面对的严肃问题.但是,通过"样本采集"获得的样本集,并应用统计方法获得的结论,对自然模型的真实性拟

合仍然是统计学的本质. 目前, 高维海量不可控数据的涌现对统计学提出了挑战性的问题, 为了解释这些问题, 我们需要了解统计学对数据统计分析的发展历程.

2 高维数据的两个基本问题

统计学始于被观测的数据, Wegman[11] 把统计描述为一种将原数据转化为信息的方法, 以区别于传统统计学的描述——传统统计学是关于收集和分析带随机性误差的数据的科学和艺术[12]. 从统计学的发展可以看出数据的采集方式经历了大样本到小样本, 再到大样本的过程. 在 1908 年以前统计学的主要用武之地是社会统计(尤其是人口统计)问题, 后来加入生物学统计问题. 这些问题涉及到的数据一般都是大量的, 自然采集的. 而所采用的方法, 以拉普拉斯中心极限定理为依据, 总是归结到正态. 到 20 世纪, 受人工控制的实验条件下所得数据的统计分析, 日渐引人注意. 由于实验数据量一般不大, 直接导致了依赖于近似正态分布的传统方法的失效, 在这种小样本的研究方向上, Gosset et al. 发展了确定正态样本统计量的精确分布的理论[13]. 无论大样本理论还是小样本理论, 它们的共同特点是数据维数一般不大, 最多几维, 即自然模型涉及的变量数量很少. 然而, 现在我们面临自然涌现的数据除了观测的数据数量剧增之外, 最大的不同是, 数据维数少则几十维, 多则上万维. 如果再考虑数据性质的复杂性和数据表述的多样性, 这不仅对计算机科学是一个挑战性的问题, 对统计学同样是一个挑战性的问题. 例如, 银行的巨额交易数据, 电话呼叫记录, 文本数据等, 数据量达到 GB 甚至 TB 级. 适合分析和处理以精心设计实验获得独立同分布、同方差和低维数据的传统统计学理论已不能适应, 需要新的思考[14].

在统计建模中高维数据会遇到两个困难:

(1) Bellman 的维数灾难(Curse of Dimensionality)现象[15]. 维数灾难现象表明, 在给定模型精度下估计模型, 需要的样本数量将随着维数的增加指数增长(图 1). 而与此相关的问题是空空间现象(empty space phenomenon)[16], 即高维空间的本质上是稀疏空间. 一个典型的例子是高斯分布中的 3σ 准则: 当样本集在二至三维空间时, 采用高斯函数, 可以证明, 90%以上的样本集分布在 3σ 范围以内. 然而, 当维数显著增加时, 样本集的分布更多地集中在高斯函数的边界(3σ 以外)而不是中间. 这表明在高维样本集中, 数据可能大多数分布在超球的外壳, 而不是在球的中心. 由此产生的困难是, 在多元数据分析中缺乏一般性的方法来直接分析高维空间的密度估计和几何性质, 因为相对低密度的区域包含了样本集的大部分, 反而高密度区域可能完全没有数据存在[17].

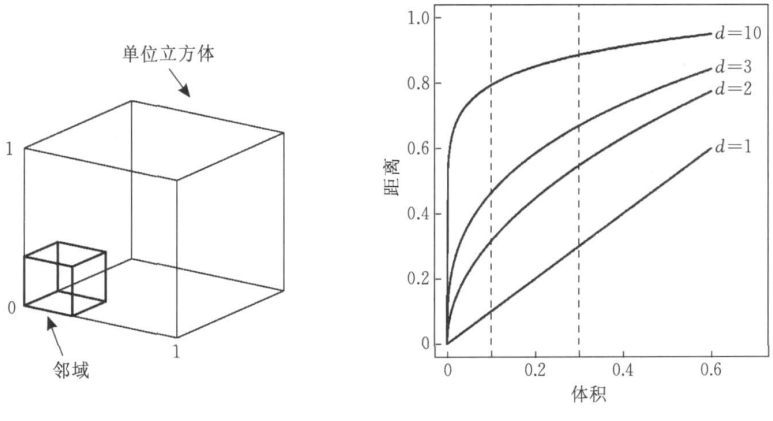

图 1 高维空间的维数灾难[18]

(2) 不适定问题. 我们对自然模型, 几乎一无所知, 如果使用传统统计学的理论、方法和理念, 估计概率密度函数, 这一定是一个不适定问题[19]. 20 世纪初 Hadamard 在某些情况下求解线性算子方程 $Af=F, f\in \boldsymbol{F}$ 的问题(寻找满足这一等式的函数 f)是不适定的. 即使方程存在唯一解, 如果方程右边有一个微小变动(如用 F_δ 取代 F, 其中 $\|F-F_\delta\|<\delta$ 任意小), 也会导致解有很大的变化(即可能导致 $\|f_\delta-f\|$ 很大). 20 世纪后半叶, 人们发现根据数据估计密度函数这个统计学中的主要问题是不适定

的: 使泛函 $R(f)=\|Af-F_\delta\|^2$ 最小化的函数 f_δ 并不能保证在 $\delta\to 0$ 时是方程真实解的一个好的近似.

高维空间的数据在拟合模型时的稀疏性, 使得所获得的样本集不足以表现自然模型; 传统统计学的不适定问题使得我们无法在高维复杂数据的情形下精确估计自然模型. 这两个有关高维的、复杂的、自然涌现数据的问题, 是重采样方法出现与成长的温床. 特别是, 当前一些重要的领域, 例如, 银行交易数据、文本数据、Web 数据都是自然涌现的, 不但数据量庞大, 而且维数很高, 并且可能不能简单以一个固定的样本空间进行描述, 即数据不能使用相同维数的向量表述. 而重采样方法恰恰为处理这类数据提供了工具, 并在理论上给出了统计解释.

3　重采样方法的思想来源

重采样方法的思想来源大致有两个方面: (1) 试验设计; (2) 抽样调查. 其出现的本原是为了更准确地获得"有代表性的样本"①. 由于当前海量涌现的高维数据具有天生的稀疏性, 因此, "获得有代表性样本"对需要满足同分布条件的各种机器学习研究具有重要的现实意义.

重采样方法最早可以追溯到 20 世纪 30 年代 Fisher(1935)提出的配对随机化检验(Randomization Test)和 Pitman(1937)提出的两个独立样本②的随机化检验[20]. 对两样本情形, 试验者从可能不同的自然模型中得到两个样本, 希望用统计假设检验来判断两个自然模型是否相同, 以决定"两个自然模型相同"这个零假设是否被拒绝③. 一个直观的方法是将两个样本组合成一个有序的样本, 不管每个值是来自哪个自然模型, 从小到大给样本赋"秩", 而检验统计量就可能是来自其中一个自然模型观测值的"秩和". 如果这个秩和太小(或者太大), 就意味着来自这个自然模型的值趋向于比来自另一自然模型的值小(或者大, 视具体情况而定). 由此可知, 如果与一个样本相关的秩趋向于比另一个样本相关的秩大, 则"两个自然模型相同"这个零假设可能被拒绝.

Fisher(1935)用数据本身作为秩来判断配对数据是否来自同一自然模型, 描述如下:

两个独立的随机样本集 $\{z, y\}$ 分别来自两个自然模型 F, G

$$F\to z=(z_1, z_2, \cdots, z_n);$$

$$G\to y=(y_1, y_2, \cdots, y_m).$$

希望根据样本检验"两个自然模型相同"这个零假设, $H_0: F=G$, 如果 H_0 为真, 两样本来自同一自然模型. 检验统计量是观测值之和: $T=\sum_{i=1}^{n} z_i$, 将 $\{z, y\}$ 混合成一个样本集, 从中抽取出 n 个样本, 在 H_0 假设条件下, 每一种 n 个样本的组合方式都是等概率的, 考虑所有组合的可能性, 可得零分布④. 之后采用标准的假设检验原理构造概率 p 值, 做出接受或者拒绝的结论.

在随机化检验中采用数据组合的方式构造假设检验的过程, 体现出了早期的重采样思想, 当两个样本集融合在一起时, 除非来自不同自然模型, 否则重新采样后的统计指标和原样本统计指标应没有差别. 此方法先于计算机发展, 所以一般限制在小样本数据上, 在样本容量较大的情形下, 需要的计算工作量很大, 在当时, 其应用必然受到限制.

几乎是同一时期, 抽样调查方面也开始冒出了重采样的萌芽, 这种早期的重采样思想是从有限自然模型中无重复采样. 在 Pearson 的统计框架中, 针对一个自然模型, 其对应着一个庞大的却有限的样本的集合. 在理想情况下, 科学家会搜集所有的这些样本, 并确定其分布参数. 如果无法搜集到全部样本, 那么就搜集一个很大的并且具有代表性的数据子集. 通过大量的且具有代表性的子集计算模型的参数, 如果数据具有足够的代表性, 被计算出的参数将与自然模型的参数相同. 然而 Pearson 学派的方法存在一个根本性的缺陷, 如果所获得的数据被称为"便利样本"(opportunity sample), 即属于那些最容易得到的数据, 这些数据并不能真正代表自然模型.

20 世纪 30 年代的早期, 印度发现了一个便利抽样的典型案例, 为了估计孟买码头上大批黄麻的价值, 需要从每包中抽取一些样品, 黄麻的质量由这些样品来确定. 抽样是将一把中空的圆形刀片插入包中, 再拔出来, 刀片中央的空处带出了少量的黄

① 其本质就是同分布条件.

② 这里的"样本"就是"样本集", 在统计学中, 由于一般不考虑单独样本的问题, 因此, 在计算机科学中使用的"样本集", 在统计学中就称为样本. 由于本文涉及统计学的内容, 对它的概念和事实, 使用"样本集"很别扭且可笑, 因此, 在不致引起混乱时, 我们有可能不加指出地使用"样本"一词, 但是, 其含义就是计算机科学中的"样本集".

③ 零假设: 关于自然模型未知分布的信息所做的统计假设. 根据零假设可以构造出一个性质优美的统计量, 通过这个统计量来做统计推断.

④ 承认零假设成立的基础上, 构造一个统计量, 这个统计量的分布即为零分布.

麻,但是由于天气和包装运输的原因,外层黄麻会变质,而由于在中间的黄麻被压紧,并结成一块,导致空心刀片难以插入,这样,所取的样本多是外层已经变质的黄麻,这种"便利样本"就会产生偏差,由此,导致评价整包黄麻质量偏低,实际上整包黄麻的质量要高得多[21].

这个例子说明了收集具有代表性样本对估计模型准确性的重要性.为了收集能够准确估计自然模型的具有代表性的子样本,当时出现了"判断样本"(judgment sample)的方法[21].这个方法是将自然模型划分为几个子模型,每个子模型都由某些样本来"代表",这些"代表"的样本组成的集合作为判断样本.但是只有对自然模型有充分了解之后,才能将自然模型划分为一些能用个体样本来代表的子模型,这样判断样本才具有代表性,如果我们对自然模型已经了解得那么清楚,就无需进行抽样[21].

Mahalanobis[21]建议采用随机样本(random sample)来推断有限自然模型.这种采样得到的样本优于便利样本和判断样本.最初 Mahalanobis(1946)在研究作物产量上使用交叉抽样方法(Half Sample),之后 McCarthy(1966)将其扩展到抽样调查领域.在抽样调查领域,仅从自然模型随机抽取,得到所有可能样本中的一次采样,并依此来推断自然模型,其推断结果是否准确可靠,无法衡量,通过重复采样法进行抽样得到 K 个子样本集,由于各个子样本集都独立且采样方式相同,若各子样本集的估计结果一致或者比较接近时,推断结果的真实性比较容易让人信服.此时的采样方法是基于从自然模型上重复采样的原则.1969 年 Hartigan 提出了 Random Sub-Sampling 方法,并首次将此方法用在统计量估计中[22].

传统统计学需要研究的问题是:如何利用样本 $X=\{X_1, X_2, \cdots, X_n\}$ 中的信息,对自然模型分布做出判断.将样本中的信息加工处理,用样本上的函数来构造统计量 $T=T(X)$(如样本均值、样本方差、回归曲面、分类函数等),用统计量来体现自然模型的信息.统计量只依赖于样本,而与参数无关.

无偏性是衡量统计量的一个基本准则,其实际意义是无系统误差,即统计量的数学期望等于自然模型的参数 $\theta(ET=\theta)$.对实际问题,无偏估计一般是不可能的,我们只是希望能够找到偏差较小的统计量,或者采用某种方法降低统计量的偏差.重采样方法刀切法[9,20]就是其中一种.

4 刀切法

1949 年,Quenouille 提出了刀切法,这是近代重采样方法的标志,以后,由 Quenouille(1949,1956)和 Tukey(1958)不断完善,重采样方法成为统计学的重要方法之一.

刀切法的原始动机是降低估计的偏差.常用做法是:每次从样本集中删除一个或者几个样本,剩余的样本成为"刀切"样本,由一系列这样的刀切样本计算统计量的估计值.从这一批估计值,不但可以得到算法的稳定性衡量(方差),还可以减少算法的偏差.这个方法暗示,刀切法的样本集需要事先给定,即,它的重采样过程是在给定样本集上的采样过程.

最简单的一阶刀切法描述如下[9]:

假设独立同分布的样本 $X=\{X_1, X_2, \cdots, X_n\}$ 来自一个未知概率模型 $F\theta$, $\theta=\theta(F)$是未知参数,$\hat{\theta}=T(X)$是估计统计量,则 θ的刀切法估计为

$$\tilde{\theta}=n\hat{\theta}-(n-1)\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i},$$

其中 $\hat{\theta}_{-i}=T(X_{(i)})$是刀切样本集 $X_{(i)}$上的统计量,$X_{(i)}=\{X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n\}$,是把原样本集中第 i 个样本剔除后剩余的 $n-1$ 个样本组成的集合.

刀切法的最重要的性质是:刀切估计可以将偏差从 $O(1/n)$减少到 $O(1/n^2)$,并可以修正估计为无偏估计,但是并不能保证减少方差.这个性质描述如下[23]:

设 $X=\{X_1, X_2, \cdots, X_n\}$为独立同分布样本集,$X_i\sim F(x,\theta)$,其中 $\theta\in\Theta$为未知参数,统计量 $T(X)$为 θ的估计,若其偏差为

$$Bias(\theta)=E(T(X)-\theta)=\sum_{k=1}^{\infty}\frac{b_k(\theta)}{n^k}=O\left(\frac{1}{n}\right),$$

则 θ的刀切法估计的偏差为 $Bias_J(\theta)=O\left(\frac{1}{n^2}\right)$.

虽然刀切法可以降低估计偏差,但当参数不光滑(smooth)时,刀切法会失效.此处光滑是指样本集上的微小变化,只会引起统计量的微小变化.最简单的不光滑的统计量是中位数(median),中位数是刻画随机变量分布"中心"的统计量.满足 $P(X\leq m(X))\geq 1/2$ 且 $P(X\geq m(X))\geq 1/2$ 的实数 $m(X)$称为中位数,在样本集上,样本中位数定义为 $m(X)=1/2[X_{[n/2]}+X_{[n/2]+1}]$.通俗地说,将一维样本排序,处在最中间位置的那个数据(或最中间两个数据的

平均数)即为这组数据的中位数. Efron 指出[24] 刀切法在估计中位数时会失效, 而自助法可以有效地给出中位数的估计. 用老鼠数据[24] 的例子来说明, 9 个排好序的样本分别为

10, 27, 31, 40, 46, 50, 52, 104, 146,

这个样本集的中位数是 46(样本个数是奇数, 中位数为最中间位置的样本). 如果改变第 4 个样本 $x=40$, 当 x 增加至并且超过 46, 中位数才会改变, 之前中位数不改变. 当样本从 46 继续增加直至 50, 中位数和此样本值相同, 超过 50 之后, 中位数变为 50. 使用一阶刀切法估计中位数, 先去掉第一个样本 $x=10$, 剩余 8 个样本的中位数是 48(46 与 50 的算术平均值), 依次去掉相应的第 i 个样本, 得到如下中位数估计结果:

48, 48, 48, 48, 45, 43, 43, 43, 43.

刀切法只得到 3 个不同的中位数估计, 方差较大. 而自助法的采样方法使得样本集变化较大, 会得到比较敏感的中位数变化. 并且, 在大样本性质上, 中位数的刀切法估计的标准差是不相合的(不能收敛到真实的标准差). 而自助估计是相合的.

5　自助世界

1979 年 Standford 大学统计系的 Bradley Efron 在统计学刊物《The Annals of Statistics》上发表了开创性论文——"自助法: 从另一个角度看刀切法(Bootstrap Methods: Another Look at the Jackknife)"[10]. 发表过程具有戏剧性, 最初, 杂志编辑毫不客气地拒绝了这篇文章, 理由是"太简单", 目前, 这个方法的影响可从有影响的重要杂志发表有关文章上得到证实: 从 1982 年开始几乎在每个数理统计期刊上都刊登一篇或者多篇与自助法相关的文章. 并且关于自助法主题的论文也不断出现在计算机学科的杂志上.

Efron 在 1979 年的这篇文章指出了自助法与刀切法的关系. (1) 自助法通过经验分布函数构建了自助法世界, 将不适定的估计概率分布的问题转化为从给定样本集中重采样. (2) 自助法可以解决不光滑参数的问题. 遇到不光滑(Smooth)参数估计时, 刀切法会失效, 而自助法可以有效地给出中位数的估计. (3) 将自助法估计用泰勒公式展开, 可以得到刀切法是自助法方法的一阶近似. (4) 对于线性统计量的估计方差这个问题, 刀切法或者自助法会得到同样的结果. 但在非线性统计量的方差估计问题上, 刀切法严重依赖于统计量线性的拟合程度, 所以远不如自助法有效.

Efron 的这篇文章是对刀切法的一种新的统计学解释. Efron 将刀切法纳入了自助法的体系中, 并构建了从真实世界(自然模型)到自助世界的采样过程. 这里, 自助世界是基于经验分布函数从给定样本集重采样获得.

样本集 $X=\{X_1, X_2, \cdots, X_n\}$ 来自一个未知概率模型 F, $\theta=\theta(F)$ 是我们关注的未知参数, $\hat{\theta}=T(X)$ 是估计参数的统计量, 它们可以通过传统统计方法(极大似然, MAP 等)获得, 定义 $R(X,F)=T(X)-\theta(F)$. 然而我们不仅关注估计值本身, 同时也关注统计量的准确程度, 是无偏估计吗? 距离真实值的偏差是多少? 稳定吗? 方差是多少? 但是这样的问题往往无法回答, 因为我们不了解自然模型本身, 我们面对的只有从自然模型中的采样结果——样本集.

我们可以在给定样本 X 的条件下, 构造 F 的估计 F_n, 然后从分布 F_n 中重新生成一批随机样本 $X^*=\{X_1^*, X_2^*, \cdots, X_m^*\}$. 如果 F_n 是 F 的一个足够好的估计, 那么 X 与 F 的关系会从 X^* 和 F_n 的关系中体现出来(图 2).

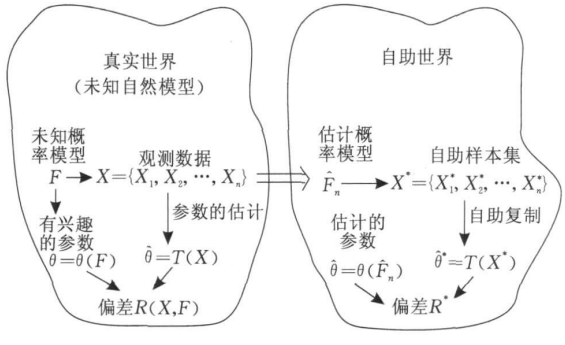

图 2　自助采样示意图[24]

自助法定义如下[10]: 样本集 $\{X_1, X_2, \cdots, X_n\}$ 来自一个未知概率模型 F, 关注统计量 $T(X_1, X_2, \cdots, X_n; F)$, 定义 F_n: $F_n(X)=\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$(其中 $I(\cdot)$ 为示性函数)是样本集 $\{X_1, X_2, \cdots, X_n\}$ 上的经验分布函数(Empirical Distribution Function), 其中每个样本的概率均为 $1/n$. 从 F_n 上 m 次随机采样得到自助样本集为 $\{X_1^*, X_2^*, \cdots, X_m^*\}$, 目的是用自助样本集上的统计量 $T(X_1^*, X_2^*, \cdots, X_m^*; F_n)$ 的分布去逼近原样本集上统计量 $T(X_1, X_2, \cdots, X_n; F)$ 的分布. 其中 m 表示自助样本集中样本的个数, n

表示原始样本集中样本的个数.产生过程如下:

$$F \xrightarrow{\text{iid}} \{X_1, X_2, \cdots, X_n\} \rightarrow T(X_1, X_2, \cdots, X_n; F)$$

$$F_n \xrightarrow{\text{iid}} \{X_1^*, X_2^*, \cdots, X_n^*\} \rightarrow T(X_1^*, \cdots, X_n^*; F_n)$$

从自然模型采样得到样本集,基于此样本集进行学习.如果样本集是对自然模型的独立同分布的采样,那么,在统计上,这样的样本集对自然模型是理想的,它可以很好地拟合自然模型.传统统计学的样本是定义在事先给定的空间上,即空间维数确定,通常可以理解为欧式空间中的点.对自然模型进行估计,并基于这个估计使用自助法得到自助样本集,可以不受样本空间维数固定的制约,并且可以追加新样本.学习的模型在统计意义下可对自然模型进行解释.重采样的次数是有限的,需要我们设计采样方法使得重采样样本构建的算法具有代表性,虽然自助法本身没有对算法类型做任何限制,但是弱可学习这个条件对于算法建模来说,容易满足,并且能够适用在自助样本集上.从自助法的采样过程来看,弱可学习建立的模型只依赖于部分样本,为了得到自然模型的拟合,需要考虑某种集成方法,将这些自助样本集上的学习算法集群起来.

6 自助法的统计性质

我们斟酌再三,还是决定在本文的正文中包含这一节.在这一节中,我们描述了自助法的统计性质,并简单说明了这些性质的证明.正是这些性质解释了自助法之所以被广泛应用的原因.我们之所以描述这些性质,主要原因是:这篇文章的读者应该是计算机科学家和工程师,他们是从 Adaboost 之类的算法了解自助法的,而讨论这些算法的文章,往往不介绍这些统计学的结论.当我们试图将这个方法用于解决更为广泛复杂的问题,而不是仅仅局限在算法设计复杂性时,例如,经验模型问题、结构化数据问题等,我们就需要了解这些性质以及它们成立的条件.我们将有些重要定理的证明罗列在正文之中,就是显现地引起读者注意这个问题,以便激发大家的联想.

6.1 自助样本集上均值的相合性

定义原样本集上估计偏差为 $R(X,F)=T(X)-\theta(F)$,自助样本集上估计偏差为 $R^*=R(X^*,F_n)$.注意,这里衡量的是自助样本集与原样本集的差别,并试图用 R^* 近似 R.

最早统计学方面关于自助法收敛的定理,特指自助样本集上样本均值的收敛,结论是 1981 年由 Bickel 与 Freedman[25] 和 Singh[26] 分别独立给出的.他们分别指出方差或三阶矩有限的独立同分布随机变量的样本均值在自助样本集上的渐进相合性.考虑最简单的参数均值 μ,常用的统计量样本均值 $\mu_n=\frac{1}{n}\sum_{i=1}^{n}X_i$,样本方差 $s_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu_n)^2$.

由中心极限定理,$R(X,F)=\sqrt{n}(\mu_n-\mu)/s_n \xrightarrow{P} N(0,1)$,即样本均值分布收敛到标准正态分布.

自助样本集上的均值估计定义为 $\mu_m^*=\frac{1}{m}\sum_{i=1}^{m}X_i^*$,自助样本集上的方差 $s_n^*=\frac{1}{m-1}\sum_{i=1}^{m}(X_i^*-\mu_m^*)^2$.在重采样过程中,$\{X_1^*, X_2^*, \cdots, X_m^*\}$ 被看成从经验分布函数 F_n 上独立同分布的采样,$R^*=\sqrt{m}(\mu_m^*-\mu_n)/s_m^*$,如下定理说明自助样本集上的 R^* 的弱收敛性.

定理 **1**[25]. 一维情况.

假设样本 $\{X_1, X_2, \cdots, X_n, \cdots\}$ 满足独立同分布,并具有有限方差 σ^2.沿着几乎所有的样本序列 $X_1, X_2, \cdots, X_n, \cdots$,给定 $\{X_1, X_2, \cdots, X_n\}$,当 $n, m\rightarrow\infty$ 有:

(1) $\sqrt{m}(\mu_m^*-\mu_n)$ 条件概率弱收敛于 $N(0,\sigma^2)$;

(2) $s_m^*\rightarrow\sigma^2$ 条件概率收敛:即 $\forall\varepsilon>0$, $P\{|s_m^*-\sigma|>\varepsilon|X_1, X_2, \cdots, X_n\}\rightarrow 0$, $a\ s$.($a\ s$ 即几乎处处收敛或概率 1 收敛).

定理 **2**[25]. 高维情况.

假设样本 $X_1, X_2, \cdots, X_n, \cdots$ 满足 R^k 空间上的独立同分布,并具有有限方差 $E\{\|X_1\|^2\}<\infty$,F_n 是给定 $\{X_1, X_2, \cdots, X_n\}$ 上的经验分布函数,$\{X_1^*, X_2^*, \cdots, X_n^*\}$ 是从 F_n 上条件独立同分布采样.沿着几乎所有的样本序列 $X_1, X_2, \cdots, X_n, \cdots$,当 $n, m\rightarrow\infty$,有:

(1) $\sqrt{m}(\mu_m^*-\mu_n)$ 的条件分布依概率收敛到正态分布,其均值为 0,协方差为 $cov(X)$.

(2) $\{X_1^*, X_2^*, \cdots, X_m^*\}$ 的经验协方差条件收敛到 $cov(X)$.

1981 年 Singh 进一步证明了在三阶矩有限的条件下自助样本集上统计量的一致收敛性,同时收敛速度比传统正态方法的收敛速度快.

定理 **3**[26]. 问题描述同前,定义 $\|\ \|_\infty$ 即 $\sup_{x\in R}|\ |$:

(1) 若 $EX^3<\infty$,F 非格分布①,那么

① 格分布是一种离散分布,其取值表示为 $a+bn$ 的形式,其中 $a, b\neq 0$,n 是整数

$n^{1/2} \| P\{n^{1/2}(\mu_n - \mu)/\sigma \leq x\} - P^*\{n^{1/2}(\mu_n^* - \mu_n)/s_n \leq x\} \|_\infty \to 0$, a s ;

(2) 若 $EX^3 < \infty$, F 格支撑为 h[①],

$\limsup_{n\to\infty} n^{1/2} \| P^*\{n^{1/2}(\mu_n^* - \mu_n)/s_n \leq x\} - P\{n^{1/2}(\mu_n - \mu)/\sigma \leq x\} \|_\infty = h/\sqrt{2\pi\sigma^2}$ a s .

由上面的两个定理可知, 自助法的理念认为: 如果独立同分布样本采自自然模型, 同时满足方差有限或者三阶矩有限这个条件, 并且样本个数无限, 通过自助法采样构造出的自助世界可以估计出真实世界的有效信息, 并且估计的速度也优于传统的统计方法. 样本均值在传统统计学中是一个重要统计量, 它衡量了样本的平均程度的刻画, 并且是正态分布的重要参数, 自助方法能够在大样本性质下(样本个数趋于无穷, 同时自助样本集中的样本个数也随着趋于无穷)具有均值的相合性, 只需满足独立同分布和矩有限两个条件.

6.2　均值自助估计的相合性

1990 年 Drago[27] 指出, 之前那些研究自助法收敛的文章主要集中在样本容量和自助样本容量趋向无穷时的收敛性质. 他定义了诸如均值、方差、均值标准差此类统计量的自助估计, 并给出在样本容量有限的条件下, 这些统计量的自助估计随着 $B\to\infty$ 的收敛性.

先给出统计量的自助估计的定义: 独立地进行自助法采样 B 次, 得到 B 个自助样本集 $X^{*b} = \{X_1^{*b}, X_2^{*b}, \cdots, X_m^{*b}\}$, $b = 1, 2, \cdots, B$ (每个自助样本集都包含 m 个自助样本). 可以得到第 b 个自助样本集上的统计量: $T(X^{*b}; F_n) \equiv T^b$, $b = 1, 2, \cdots, B$.

定义统计量的自助估计为

$$T^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B} T^b.$$

偏差的自助估计为

$$bias_B = \frac{1}{B}\sum_{b=1}^{B} T^b - T = T^{*(\cdot)} - T.$$

方差的自助估计为

$$Var_B = \frac{1}{B}\sum_{b=1}^{B}(T^b - T^{*(\cdot)})^2.$$

定理 **4**[27]. 样本集 $\{X_1, X_2, \cdots, X_n\}$ 独立同分布, 且方差有限, 考虑最简单的统计量样本均值时, $B\to\infty$, 均值方差的自助估计就是均值方差的传统估计, 即

$$Var_B = \frac{1}{B}\sum_{b=1}^{B}(\mu^{*b} - \mu^{*(\cdot)})^2 \to \widetilde{Var}(\mu_n).$$

需要注意的是统计量的自助估计偏差衡量的是与原样本集上统计量的差别, 而非与真实世界的参数的差别.

证明.

(1) 先给出样本均值方差的传统形式:

样本均值 $\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, 样本方差 $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n)^2$. 样本均值的方差为

$$Var(\mu_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}Var(X_i) = \frac{1}{n}\sigma^2.$$

用样本方差代替方差可得样本均值的方差估计:

$$\widetilde{Var}(\mu_n) = \frac{1}{n}s_n^2.$$

(2) 样本均值的自助方差估计

均值的自助估计为

$$\mu^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\mu^{*b}, \quad \mu^{*b} = \frac{1}{n}\sum_{i=1}^{n} X_i^{*b}.$$

均值的方差的自助估计为

$$Var_B = \frac{1}{B}\sum_{b=1}^{B}(\mu^{*b} - \mu^{*(\cdot)})^2.$$

根据 Tchebycheff 不等式:

$$P\{|Var_B - E(Var_B)| < \varepsilon\} \geq 1 - \frac{1}{\varepsilon^2}D(Var_B).$$

由(Cramer(1946))定理可以得到

$$\lim_{B\to\infty} D(Var_B) = 0, \quad E(Var_B) = \frac{1}{n}s_n^2.$$

所以由弱大数收敛定理可得:

$$\lim_{B\to\infty} Var_B = E(Var_B) = \frac{1}{n}s_n^2,$$

所以得到结论. 证毕.

算法建模中样本容量都是有限的, 研究在有限样本的情况下样本均值的性质是重要的, 从定理的证明可以看出, $\lim_{B\to\infty} D(Var_B) = 0$, 样本均值的方差在自助法的意义下, 是稳定的. 重采样的这个过程当 $B\to\infty$ 时可以得到很稳定的估计量.

7　自助法的理论基石

自助法之所以成功, 是因为自助法不需要对未知自然模型做任何假设, 也无需事先推导出估计量的精确解析式, 只需重采样并计算估计值. 这样看来, 本质上是一种非参数方法[28]. 在实现过程中, 计算机的地位不容忽视[29]. 因为其中用到了大量的计

① 格支撑: 如果格分布表示为 $a + bn$ 的形式, 其中 a, $b \neq 0$, n 是整数, 那么 b 就是其支撑.

算,如果没有计算机,自助法理论只能是纸上谈兵.同时由于对于自然模型参数的一些估计无法得到明确的解析式,自助法的出现使得我们绕过这种繁琐的理论推导.自助法有两个重要的理论基础,其暗示了这种方法需要满足的条件.

Glivenko-Cantelli 引理

1916 年 Cantelli 发现了统计学基本原理即 Glivenko-Cantelli 引理,指出尽管存在一些数据,对其概率分布一无所知,但数据本身可以用来构造一个非参数分布.虽然其数学函数不优美,还有很多断点,但是还是可以通过增大观测值的数量,来使不那么美的经验分布函数(empirical distribution function)越来越接近真实的分布函数.经验分布的构造只需一系列的简单数学[30].

早期的Glivenko-Cantelli 引理针对一维变量,1933 年, Kolmogorov-Smirnov 发现经验分布函数收敛到真实分布函数的渐近准确估计,并指出这种收敛不依赖于真实分布函数的具体形式.这为自助法在高维分布上的使用提供了定理保证.

非参数检验

非参数检验有两个很重要的问题:若数据具有一个已知的参数分布,比如正态分布,这种情况下我们采用非参数分析方法会产生什么问题?若数据不太适合采用参数模型(parametric model),那么数据必须偏离参数模型多远时,使用非参数方法才会更优.1948 年 Pitman 成功地解决了两大疑问.他指出:当知道参数模型和本应使用特定的参数检验时,即使采用非参数检验,效果也根本不差;如果数据稍稍偏离参数模型,则非参数检验将远远地胜于参数检验[30].这就为自助法在非参数检验方面扫清了障碍.传统的参数检验依赖于正态分布的假设,但是如果假设错误,会造成很大的偏差,自助法本质上是非参数的不依赖于自然模型的具体形式.

8 自助法的几何解释

1993 年, Efron[24] 给出了自助法的几何解释.在这一节中,不再把参数 θ 看成样本 X 的函数,而是固定样本 X 把概率分配给每个样本.设概率向量 $\boldsymbol{P}^*=(P_1^*, P_2^*, \cdots, P_n^*)$ 满足 $0\leq P_i^*\leq 1$ 和 $\sum_{i=1}^{n} P_i^*=1$,使得 $F^*=F(P^*)$ 是分布函数.定义 θ^* 作为 P^* 的函数,记为 $T(P^*)$. $\theta^*=T(P^*)\equiv t(F^*(P^*))$ 这样的定义使得 F^* 上的函数 t 转换成了 P^* 上的函数 T.向量 P^* 是一个 n 维单纯型(simplex). $n=3$ 时单纯型是一个等边三角形(图 3).

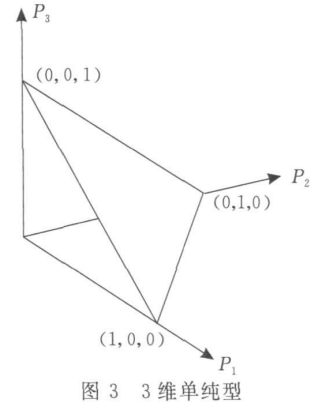

图 3 3 维单纯型

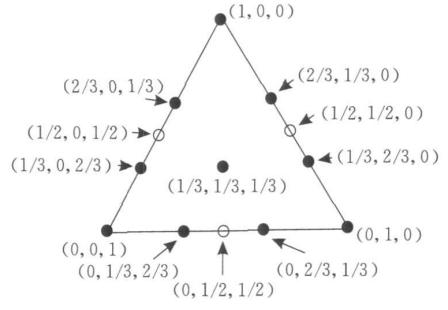

图 4 单纯型的平面图

我们定义 $\boldsymbol{P}^0=\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)^{\mathrm{T}}$,此时 $T(\boldsymbol{P}^0)$ 表示 3 维单纯型的中心(图 4).刀切法的统计量为 $\hat{\theta}_{(i)}=T(\boldsymbol{P}_{(i)})$,其中 $\boldsymbol{P}_{(i)}=\left(\frac{1}{n-1}, \cdots, 0, \frac{1}{n-1}, \cdots, \frac{1}{n-1}\right)^{\mathrm{T}}$ (见图 4 的空心点).

而自助法相等于一个多项分布抽取 nP^*:

$$P^* \sim \frac{1}{n} Mult(n, P^0).$$

此分布的均值向量和协方差矩阵如下:

$$P^* \sim \left(P^0, \left[\frac{I}{n^2}-\frac{P^0 P^{0\mathrm{T}}}{n}\right]\right).$$

$T(P^*)$ 构成了单纯型上一个曲面,如图 5 所示.

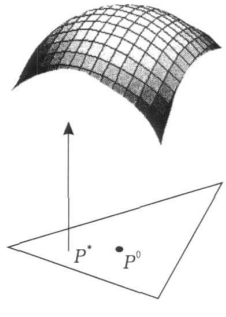

图 5 $T(P^*)$ 构成的曲面

自助法的几何解释使得我们可以从样本集上采样策略来理解自助法, 所关心的统计量 $T(X)$可以看成 P^*的函数, P^*被认为是自助样本集上的采样策略.

9　自助法研究现状

Efron 于 1979 年提出了自助法, 他指出通过经验分布逼近真实分布, 可以在原样本集上"有放回抽样", 从而得到自助样本集, 重复 B 次得到统计量的偏差和方差, 并证明了自助法和刀切法在估计中位数上的显著不同. 1991 年 Efron 在 Science 上发表论文 (Statistical Data Analysis in the Computer Age[31]), 确认自助法是计算机时代运算密集型 (Computer-intensive) 统计方法之一. 1996 年 Breiman 研究了"重采样技术"在估计和分类器设计中的应用, 提出了 Bagging[8] 和 Arcing[32]. 2003 年 Statistical Science 出版特刊, 以纪念自助法发表 25 周年. 30 年的系统研究使得有关自助法的书籍和论文层出不穷, 并广泛应用在统计学的各个领域. 这里, 我们仅仅简要介绍自助法在丰富统计学研究上的几种典型应用.

在 21 世纪之前, 自助法的研究主要集中在统计量上, 换句话说, 就是研究这种方法在统计学上的性质. 在 20 世纪后期, 自助法的研究开始与统计建模的研究联系在一起. 之后, 它的研究与机器学习的研究越来越难以区分, 尽管在文章的风格和关注的问题上, 可以看出作者属于哪个领域, 但是, 从解决的问题来看, 不外乎就是分类与回归.

如果是从给定有限样本集进行重采样, 这时, 独立同分布是自助法必须满足的条件, 目前所获得的所有有关的性质均需要满足这个条件. Efron 最早在 1979 年的论文中使用的例子就是独立同分布数据的自助法, 它是重采样理论中发展最早的结果.

对于独立同分布样本$\{X_1, X_2, \cdots, X_n\}$, 如前所述, 选取经验分布 F_n 作为真实分布 F 的估计, 当确定 F 的经验分布后, 从自助世界采样这一过程就马上简化为"有放回的简单随机抽样" (因为经验分布的本质也就是每个样本出现的概率都相等), 即从$\{X_1, X_2, \cdots, X_n\}$中有放回地随机抽取$\{X_1^*, X_2^*, \cdots, X_m^*\}$. 比如我们想要估计的统计量是均值, 即 $T_n=\sqrt{n}(X_n-\mu)/\sigma$, 其自助法估计是$T_n^*=\sqrt{m}(X_m^*-\mu^*)/\hat{\sigma}$.

如何衡量自助法结果的逼近程度呢?上节给出 Bickel 和 Freedman 以及 Singh 证明自助法相合性的定理, 是渐进相合的. 而 Singh 进一步证明了在方差有限的情况下, 自助法近似比传统正态近似的收敛速度要快. 自助法的优良性质都是在矩存在且有限的情况下成立的, 如果此条件不再符合, 通常的自助法就会失效. 如随机变量的均值有限但方差无限, 具体如自然模型服从稳定分布($\alpha\in(0, 2)$), 此时方差是无限的, 这种情况下 Athreya[33] 证明: 如果仍然对统计量 T_n进行常规的自助法逼近, 那么这个近似将收敛到一个随机极限. 解决的办法是重采样的样本量适当取小, 让 $m=o(n)$, $n\to\infty$.

1987 年 Athreya 指出在方差有限的情况下, 统计量的自助法估计才具有优良性质, 如果条件不满足, 通常的自助法就会失效. 并证明了如下定理.

定理 **4**[33]. 样本集$\{X_1, X_2, \cdots, X_n\}$独立同分布, $EX_1^2=\infty$, 任给实数集$\{x_1, x_2, \cdots, x_n\}$, 随机向量序列 $H_n(x_i, \omega)$, $i=1, 2, \cdots, k$收敛到随机向量 $\boldsymbol{H}(x_i, \omega)$, $i=1, 2, \cdots, k$, 其中 $H_n(x_i, \omega)=P(T_n\le x|X_n)$, $T_n=n(\mu_n^*-\mu_n)/\max(X_1, X_2, \cdots, X_n)$.

证明见文献[33].

从这个定理可以得知, 在方差不存在的条件下, 自助法会失效, 均值的自助估计的分布收敛到一个随机向量, 而非原样本集上的统计量的分布.

Efron 在他 1979 年的论文中考虑了多元回归的情形, 主要是用自助法来近似回归系数的分布.

我们知道回归模型的常用方法是最小二乘法 (LSE), 但最小二乘法的使用有 3 个假设条件: 误差项独立且等方差; 自变量不受噪音污染; 误差项方差有限. 由于最小二乘法对模型假设很敏感, 使得有重尾①的误差分布或者有异常点(outlier)时最小二乘法就会失效. 在面对异方差的误差②、残差有相关结构, 非线性模型, 非高斯误差分布或者更复杂的数据结构诸如此类的回归问题时我们可以采用自助法[34].

有两种自助法处理回归模型. 第一种是将自助法参数化, 对回归残差进行重采样. 首先用所有样本建立回归模型, 得到相应残差之后将残差中心化, 然后从中心化的残差 ε_i^* 中重采样, 并根据之前估计的回归方程重新计算因变量的值, 即

① 重尾即高阶矩发散, 从某一阶开始的高阶统计量变成无穷大, 直接限制了系统的预测性. 传统统计学中常用的正态分布和指数分布族不是重尾分布.

② 异方差指回归模型的误差方差不相等.

$$Y_i^* = x_i\beta + \varepsilon_i^*, i = 1, 2, \cdots, n,$$

得到 y 值后重新计算回归系数,重复这样的步骤,最后就可以用重采样计算的回归系数来近似其分布.

第二种回归自助法是自助法采样的直接应用.对样本集进行有重复采样,基于重采样的自助样本集建立回归模型.Efron 指出第二种回归自助法在模型假设偏差的情形下不敏感.

自助法最初是为了评估统计量准确性或评估预测精度而生的一种方法.如何用它来改进估计或改进预测呢?机器学习中的算法多是不光滑非线性的,基于此原因,自助法在改进算法性能方面的应用效果受到了 Breiman 的关注[35-36].1996 年 Breiman 将"重采样技术"用在估计和分类器设计中,提出了 Bagging,即 Bootstrap aggregation(自助聚集),同时 Schapire 基于 PAC 框架提出了 Boosting.此后 Boosting 广泛应用在计算机各个领域,并且保持很好的效果.

这和机器学习的特点有很大关系.机器学习算法试图在假设空间中找到最优的函数(最接近真实函数的假设函数).设计假设空间时,需要注意两个问题:假设空间的大小和假设空间是否包含真实函数.如果假设空间很大,则需要更多的样本来限制搜索[37].Dietterich[37] 在统计、计算和表示等 3 个方面总结了 Boosting 集群各个弱分类器的效果优于单个分类器的原因.

(1) 统计上的原因

对于一般的学习任务,往往要搜索的假设空间十分巨大,但是能够用于训练分类器的训练集中样本个数不足以用来精确地学习到目标假设.这时,学习的结果便可能是一系列满足训练集的假设,而学习算法能够选择这些假设之一,作为学习到的分类器的输出.然而,通过机器学习的过拟合问题的研究[38],我们看到,能够满足训练要求的多个假设,却不一定在实际应用中具有同样好的表现.这样,当学习算法要选择哪个假设作为输出时,就会面临一定的风险.如果把多个假设集群起来,则能够降低这种风险(这可以理解为通过集群使得各个假设和目标假设之间的误差得到一定程度的抵消).

(2) 计算上的原因

如果学习算法不够好,就无法解决上述搜索问题.已经证明,试图使用人工神经网络和决策树学习,得到最好的模型是一个 NP-hard 问题[5,37].其它的分类器设计算法也面临着类似的计算复杂度的问题.这使得我们只能使用某些启发式的方法来降低寻找目标假设的复杂度,但这样的结果是,找到的假设不一定是最优的.通过把多个假设集群起来能够使得最终的结果更加接近实际的目标函数值.

(3) 表示上的原因

由于假设空间是人为规定的,在大多数机器学习的应用场合中,实际目标的假设可能并不在假设空间之中.如果假设空间在某种运算下不封闭,那么,我们通过把假设空间中的一系列假设集群起来,就有可能表示出不在假设空间中的目标假设.

在机器学习中分类器的优劣受以下几方面影响:模型的不正确假设;分类器复杂程度太低以至于不足以拟合模型;分类器变量的错误选择;分类器的不稳定性(样本集的微小变化会引起分类器的较大变化).采用弱分类器集群可以避免这些条件限制,从而通过集成方法得到强学习器.

10 自助法与机器学习

1991 年,Efron 和 Tibshirani 在 Science 上发表文章,将统计建模划分为 3 个步骤,并以 3 个问题的方式说明这 3 个步骤[31],这 3 个问题是:

(a) 收集怎样的数据来解答我们面临的问题?

(b) 从这些数据中能得到什么结论?

(c) 这些结论的可信程度有多高?

应该说,统计机器学习与 Efron 讨论的统计建模(Statistical Modeling)是有差别的,但根据 Breiman 在 2001 年《统计科学》(《Statistical Science》)杂志上的文章[39],统计机器学习可以看作统计建模文化中的一种,因此,Efron 对统计建模需要考虑的 3 个问题,也就成为统计机器学习考虑的问题了.

统计建模基于给定样本集,并认为这个样本集是对自然模型的一次采样,因此独立同分布是统计建模的重要条件,只有保证此条件,才能有信心从样本集中学习到适用的统计模型.很多研究,严格假设数据来源的分布,甚至更进一步假设为正态分布,但如果此假设有误,建模就会出现严重偏差.统计机器学习避免对数据的来源作出假设,如何获得对自然模型的一个近似数学模型就是机器学习的主要问题了.

对机器学习来说,根据算法设计的特点,基于重采样技术的机器学习需要解决的 4 个问题是样本采

集、采样策略、算法类型和集群方法.

(1) 样本采集. 一般来说, 此步骤与机器学习无关. 但采样的数据是否足够体现自然模型的信息, 对机器学习来说却至关重要. 如果数据不足或偏离太大, 机器学习无论如何努力, 也无济于事. 与机器学习一样, 统计学对这个问题也十分重视. 一般地说, 无论机器学习还是统计学, 解决这个问题的一个自然的考虑是追加样本. 但是, 由于追加样本付出的代价太大(时间代价或者花费代价)或者实验条件无法重现使得这样的方式不可行. 考虑经验模型也许是一个可行的方法. 经验模型可以看做是领域专家对自然世界的一种解释性建模, 而它在重采样方法中, 可以理解为专家从自然模型中采样, 并由此获得的模型. 因此, 它们可以纳入重采样理论框架之中.

(2) 采样策略. 这个问题与模型集群有关. 假设一个问题被多个子模型[①]描述, 这些子模型是基于重采样样本建立的, 如果模型个数可以趋于无穷大, 此问题可以完美解答. 但由于计算的限制, 模型个数是有限的, 所以具有不同性质的模型个数必须平衡, 否则在数量上占有优势的模型将使得解答偏向这些模型给出的解答, 这就失去多个模型集群的本意了. 因此, 在建立模型时, 采用"富信息"策略来保证模型的差异性, 以保证被集群的模型不具有数据量上的绝对优势. 对机器学习来说, 采样策略是必须的, 其原因是, 我们无法处理无穷次重采样. 传统统计学并不涉及这个问题.

(3)算法类型. 重采样没有显现指出使用何种方法建立模型, 也没有指出所建模型的表述形式, 甚至没有说明模型应该满足的条件(例如, Schapire 的弱分类器需要满足比随机猜想稍好), 仅仅要求通过对特定问题的观察数据来选择模型. 究其原因, 主要是重采样方法的理论建立在采样次数趋于无穷大的基础上. 但是, 对计算而言, "采样次数趋于无穷大"是永远不能满足的, 因此, 在机器学习中使用这个方法, 还是需要满足弱可学习条件为好. 重采样对算法和模型这种相对宽松的条件, 使得我们可以对此有更多的期望, 以解决目前很多难以解决的问题, 我们将在最后一节继续讨论这个问题.

(4)集群方法. 如何将已建立的子模型组合在一起, 使其成为一个统一的模型, 这就是集群的任务. 模型集群也有学者称为"元学习(meta learning)", 其含义就是将所有子模型理解为一个空间的基集合, 集群就是在这个由子模型张成的空间上的学习. 目前, 这个学习的基础是加性函数, 或直观地说, 就是考虑子模型的线性组合, 学习就是为每个子模型加上不同的权值. 事实上, 加性模型暗示的就是一种投票机制, 并以少数服从多数作为问题求解的解答. 如果采用"一人一票制", 就没有在子模型空间上的学习问题了. 但是, 被集群的子模型求解能力是不同的. 这就像对一家公司的投资人, 不同的投资数量对公司具有不同的决定权一样, 需要对这些子模型加权. 可是如何加权呢? 这就需要评价子模型的能力, 可以用重采样的方式解决这个模型评价问题.

尽管机器学习需要考虑的问题与统计学不尽相同, 但是, Efron 提出的 3 个问题, 应该是机器学习不能逃避的问题. 当然, 对具体问题的具体考虑则不同. 其中采样次数的有限性和样本采集是在给定的有限样本集还是自然模型可能是最大的区别.

对机器学习来说, 样本集和采样策略是重要的. 后者由于涉及"富信息"问题, 受到机器学习研究者较大的关注. 但是, 前者似乎更具重要意义, 就是如何采样以及样本形式的要求. 重采样方法对这些要求比传统机器学习的要求低得多. 传统机器学习, 所有样本需要定义在事先确定的空间上, 这是一个很大的限制. 通过重采样方法建立的集群学习算法尽管对建立同一子模型需要考虑这个限制, 但是, 对不同子模型却没有要求它们必须考虑使用相同形式的样本, 除非重采样仅仅被使用在给定的样本集上, 因此, 大大扩展了应用的范围. 另外, 对经验模型, 即仅仅表现为简单规则的由领域专家总结的知识, 同样可以理解为由专家经过采样后进行分析建立的模型, 这是一个非常有趣的事实.

研究者知道, 经验知识是人工智能研究的基石之一. 但是, 人们一般不将这类知识的获得与从样本集经过数学方法建立的模型联系在一起, 而且这两类研究甚至是排他的. 50 年前, 人工智能就是针对统计建模的缺陷而提出的一种处理复杂信息的方法, 而 20 余年前, 人工神经网络(统计机器学习的前身)又是在人工智能面临困境时发展的产物. 难道两者只能存在之一吗?尽管统计机器学习的泛化理论对理论研究和应用具有足够的吸引力, 但是, 要求采样的样本集与自然模型"同分布"的条件, 很多情况下难以满足. 这样, 很多从这类样本集发展的算法和方法可能就难以向实际问题推广. 尽管经验模型同

① 本文在基于 PAC 的讨论中, 根据 Valiant 的说法, 使用"弱模型", 这里称它们为"子模型", 但是, 它们的特点是相同的.

样具有这个问题,但是,将这类知识作为一种补充(在某种情况下甚至是决定性的)不失为一种有效的方案.当然,应该说明我们假设"专家之所以称为专家,因为他们的确对某个领域有独到的见解".当我们将专家的经验知识理解为他们从观测数据总结的模型时,除了建立模型的方法我们未知之外,观测过程就可以理解为一次采样,而重采样方法对如何从数据建立模型,即建模方法,没有确定的限制,这样,将经验模型集群就是完全可行的方案了.

在基于重采样方法的机器学习中,对建立子模型的方法没有限制,暗示其背后起作用的样本可以来自给定数据集,也可以来自自然模型.

集群方法无需多言,只要注意如果将其理解为一类学习的话,它的建模是在子模型张成的空间上,而不是样本集或自然模型所依赖的空间.另外,有些子模型可能有特殊性质,特别是如果考虑经验模型的话,这就需要仔细研究损失函数的设计.同时,对一个特定问题,并不是所有子模型都是有意义的,有的甚至是干扰,这就需要在子模型空间上对其变量进行选择,这是目前另一个热门话题了.

11 总结与讨论

像大多数计算机科学家一样,我们首先关注重采样方法不是从统计学角度,而是弱可学习定理,并由此长时间研究集群学习,即,Boosting.我们欣赏这个方法的原因是,使用如此简单的方法,来解决算法设计复杂性.当时人们的研究主要集中在这类集群学习如何与 Margin 联系在一起,以便研究这类学习范式的泛化能力.

近几年,一些统计学家加入机器学习研究,并将机器学习研究纳入他们的研究范围,由于他们的研究结果大多数发表在统计学杂志上,我们不得不开始寻找这些我们不熟悉的杂志,并浏览这些杂志上的文章.

事实上,我们第一次阅读 Duda 和 Hart 合著的经典著作《模式分类》(《Pattern classification》)[40]第二版时,发现作者将这个版本发展成为一个模式识别(也可以称为统计机器学习或基于统计的学习)研究的大辞典,因此,作者将重采样方法作为一个小节放在"独立于算法的机器学习"一章中.这个小节的讨论不是从 PAC 开始,而是基于统计学,因此,并没有包含弱可学习定理的显现说明,但是,对 Schapire 构造型的证明原理,是以算法的形式表述.这意味着,作者可能并不欣赏有限样本的统计理论.鉴于这本书并不是一本关于统计学的著作,因此,这一小节中并没有包含重采样方法的理论结果.这样,就使得我们无法了解重采样方法的基础假设和条件.

恰恰此时,我们开始关注变量选择的研究进展,重采样中最重要的方法——自助法的创始人 Efron 在 2004 年给出了一个变量选择的算法 LARS[41],我们注意到,他研究变量选择的目的之一就是考虑自助法中的子模型选择.尽管作者没有对这个论题展开讨论,但是,却引起我们的好奇,并就困惑我们已久的几个问题发生共鸣.

(1)"富信息"的采样策略依赖给定样本集,如果这个样本集不能满足同分布条件,这个"富信息"就成为无米之炊.

(2)目前在机器学习研究中已有大量算法,评价的方法主要是 Benchmark 样本集的测试,相差几个百分点的算法就能够评价其优劣吗?这些算法是否应该保留,并让它们在问题求解时起不同作用呢?

(3)经验模型一直被排斥在统计机器学习之外,这合理吗?它们可以成为一类子模型吗?

为此,我们感到需要详细了解重采样方法的统计学解释,我们发现,由于隔行的原因,重采样方法已经发展成为统计学习的基本方法之一,为了说明这个现实,我们将 1979 年自助法发表之后三十年间出版有关此问题论述的重要专著罗列如下:

(1)Efron 关于重采样方法的理论基础的著作[42].

(2)Hall 关于自助法和 Edgeworth 展开的著作[43].

(3)Mammen 关于自助法收敛的著作[44].

(4)Efron 和 Tibshirani 关于自助法在标准误的估计、各种复杂数据结构、回归分析、偏差估计、与刀切法的对比、区间估计、置换检验、交叉验证等方面详细理论和应用的著作[24].

(5)Shao 和 Tu 关于刀切法和自助法理论和应用的著作[45].

(6)Davison 和 Hinkley 关于自助法方法的应用的著作[46].

(7)Good 关于自助法用于各种假设检验的著作[47].

通过对重采样方法的统计学研究,使得我们有必要认清在机器学习中存在的问题及待研究的方

向. 我们从以下几个方面说明这个问题.

(1) 自助法世界的分布

在自助法世界中, 随机变异(random variation)有两个来源: (1) 从自然模型中随机采样得到的原样本集; (2) 从原样本集重采样得到自助样本集. 传统统计学主要集中在第一个变异来源, 解决方式有两种: 无限样本的大样本理论或者小样本的精确分布. 基于目前统计学的研究现状, 这两种解决方式都会失效. 无法获得无限样本, 即使是无限样本, 由于高维数据的稀疏性, 使得样本集不足以拟合自然模型; 而精确分布又是一个不适定的问题, 人们试图采用自助法作为逃脱此困境的办法. 自助法可以不必额外增加样本, 也避免估计分布, 但是, 从原样本集重采样这个过程会增加新的随机变异. 第 6 节介绍的 1981 年 Bickel 和 Freedman[25]、Singh[26] 的两个定理都是针对简单统计量样本均值的, 指出在大样本的意义下, 如果矩有限, 自助样本集可以消除第二种随机变异.

(2) 自助法与数据描述的无关性

在自助样本的构建过程中, 与数据的描述无关, 给我们在高维空间的维数约简提供了可能. 虽然这种约简还不能达到人能够理解的程度, 但是相信至少提供了某种解释性.

(3) 自助法与 Boosting 的关系

从自助法的几何性质可以得知, 所关心的统计量 $T(X)$可以看成 P^*的函数, 而自助法相当于一个多项分布抽取 nP^*, $P^* \sim \frac{1}{n} Mult(n, P^0)$. 而 Boosting 可以看成"富信息"策略的重采样, 其 $P^* = (\omega_1, \omega_2, \cdots, \omega_n)$, 基于此点, 可以试图将自助法和 Boosting 统一纳入重采样的框架中.

重采样方法的一系列的理论结果促使计算机科学家开始考虑使用这个方法, 只要我们注意这个方法需要满足的条件, 即, 独立同分布与矩有限. 对有限观察来说, 这个理论并没有比其他统计方法为我们建模带来根本性的改善, 但是, 这个方法暗示我们, 可以多次采样, 多种方式采样, 不受建模方法限制, 这为我们带来了巨大的想象与发展空间. 特别是, 使得将专家经验模型作为一次采样后的模型成为可能. 这使得人工智能与统计机器学习相互排他的研究一致达成理解成为可能.

整体而言, 对计算机科学来说, 与其说重采样方法是一个理论不如说是一个"方法论"更为恰当. 事实上, 重采样告诉我们了一种工作的方法, 它绝不是仅仅简单地解决算法设计复杂性而已. 这正是我们写这篇综述的本意.

今年恰逢"自助法"发表 30 周年, 作为计算机科学领域的研究者, 将这个在统计学发展并被广泛应用的理论用于解决机器学习研究中的问题, 并希望它在机器学习研究中扮演更为重要的角色, 应该是一件有趣的事情.

参 考 文 献

[1] Valiant L G. A theory of learnable. Communications of the ACM, 1984, 27(11): 1134-1142

[2] Kearns M, Valiant L G. Learning Boolean formulae or finite automata is as hard as factoring. Cambridge, MA: Harvard University Aiken Computation Laboratory. Technical Report TR-14-88, 1988

[3] Kearns M, Valiant L G. Cryptographic limitations on learning Boolean formulae and finite automata. Journal of the ACM, 1994, 41(1): 67-95

[4] Schapire R E. The strength of weak learnability. Machine Learning, 1990, 5(2): 197-227

[5] Dietterich T G. Ensemble methods in machine learning// Proceedings of the Multiple Classifier Systems. Cagliari, Italy, 2000: 1-5

[6] Freund Y, Schapire R E. Experiments with a new Boosting algorithm// Proceedings of the Thirteenth International Conference on Machine Learning (ICML). Bari, Italy, 1996: 148-156

[7] Breiman L. Prediction games and arcing classifiers. Neural Computation, 1999, 11(7): 1493-1517

[8] Breiman L. Bagging predictors. Machine Learning, 1996, 24(2): 123-140

[9] Miller R G. The jackknife-a review. Biometrika, 1974, 61(1): 1-15

[10] Efron B. Bootstrap methods: Another look at the Jackknife. The Annals of Statistics, 1979, 7(1): 1-26

[11] Wegman E J. Computational statistics: A new agenda for statistical theory and practice. Journal of the Washington Academy of Science, 1988, 78(1): 310-322

[12] Chen Xi-Ru. An Opportunity Mathematic. Beijing: Tsinghua University Press, 2000(in Chinese)
(陈希孺. 机会的数学. 北京: 清华大学出版社, 2000)

[13] Gosset W S. The probable error of a mean. Biometrika, 1908, 6(1): 1-25

[14] Wegman E J. On the eve of the 21st century: Statistical science at a crossroads. Computational Statistics and Data Analysis, 2000, 32(3): 239-243

[15] Bellman R. Adaptive Control Processes: A Guided Tour. Princeton, New Jersey: Princeton University Press, 1961

[16] Scott D W, Thompson J R. Probability density estimation in higher dimensions//Gentle J E. Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface. Amsterdam: North Holland-Elsevier Science Publishers, 1983: 173-179

[17] Li Zi-Qing, Zhang Jun-Ping. Subspace Statistical Learning in Face Recognition//Wang Jue et al. Machine Learning and Its Applications. Beijing: Tsinghua University Press, 2006 (in Chinese)
(李子清, 张军平. 人脸识别中子空间的统计学习//王珏等. 机器学习及其应用. 北京: 清华大学出版社, 2006)

[18] Hastie T, Tibshirani R, Friedman J H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag, 2001

[19] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

[20] Simon J L. Resampling: The New Statistics. Arlington, VA: Resampling Stats, Inc. 1997

[21] Mahalanobis P C. Recent experiments in statistical sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, 1946, 109: 325-378

[22] Hartigan J A. Using subsample valuesas typical values. Journal of the American Statistical Association, 1969, 64 (328): 1303-1317

[23] Wasserman L. All of Nonparametric Statistics. New York: Springer, 2006

[24] Efron B, Tibshirani R. An Introduction to the Bootstrap. New York: Chapman & Hall Ltd, 1993

[25] Bickel P J, Freedman D A. Some asymptotic theory for the bootstrap. Annals of Statistics, 1981, 9(6): 1196-1217

[26] Singh K. On the asymptotic accuracy of Efron's Bootstrap. Annals of Statistics, 1981, 9(6): 1187-1195

[27] Drago C, Zoran R. Some asymptotic behavior of the bootstrap estimates on a finite sample on a finite sample. Statistical Papers, 1990, 31(1): 41-46

[28] Mooney C Z, Duval R D. Bootstrapping: A Nonparametric Approach to Statistical Inference. London: Sage Publications, 1993

[29] Diaconis P, Efron B. Computer-intensive methods in statistics. Scientific American, 1983, (5): 116-130

[30] Salsburg D. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. New York: Henry Holt & Company, 2002

[31] Efron B, Tibshirani R. Statistical data analysis in the computer age. Science, 1991, 253(5018): 390-395

[32] Breiman L. Arcing classifiers. The Annals of Statistics, 1998, 26(3): 801-824

[33] Athreya K B. Bootstrap of the mean in the infinite variance case. The Annals of Statistics, 1987, 15(2): 724-731

[34] Chernick M R. Bootstrap Methods: A Guide for Practitioners and Researchers. 2nd Edition. New York: Wiley, 2007

[35] Breiman L. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. JASA, 1992, 87(419): 738-754

[36] Breiman L. Heuristics of instability and stabilization in model selection. The Annals of Statistics, 1996, 24(6): 2350-2383

[37] Dietterich T G. Machine learning research: Four current directions. AI Magazine, 1997, 18(4): 97-136

[38] Mitchell T. Machine Learning. McGraw-Hill, 1997

[39] Breiman L. Statistical modeling: the two cultures. Statistical Science, 2001, 16(3): 199-231

[40] Duda R O, Hart P E, Stork D G. Pattern Recognition. 2nd Edition. New York: John Wiley & Sons, 2001

[41] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Annals of Statistics, 2004, 32(2): 407-499

[42] Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1982

[43] Hall P. The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag, 1992

[44] Mammen E. When Does Bootstrap Work? Asymptotic Results and Results and Simulations. New York: Springer-Verlag, 1992

[45] Shao J, Tu D. The Jackknife and Bootstrap. New York: Springer-Verlag, 1995

[46] Davison A C, Hinkley D V. Bootstrap Methods and Their Application. Cambridge: Cambridge University Press, 1997

[47] Good P. Permutation, Parametric and Bootstrap Tests of Hypotheses. 3rd Edition. New York: Springer-Verlag, 2005

BI Hua, born in 1979, Ph. D. candidate. Her research interests include machine learning, data mining, statistical learning.

LIANG Hong-Li, born in 1980, Ph. D. candidate. His research interests include rough set theory, data mining and machine learning.

WANG Jue, born in 1948, professor, Ph. D. supervisor. His research interests focus on machine learning.

Background

Adaboost algorithm proposed by Schapire is a universal method for improving accuracy of any given learning algorithm while maintaining excellent interpretability. It has been intensive focus among computer science field since appearance. The research focus is on the design of algorithm, namely, optimizing procedures with various specific loss functions. In essence, a statistical perspective of boosting algorithm is brought out in this paper, i. e., an interesting special case of resampling methods. Recently, the modeling of high dimensional and large volume data generated in an uncontrolled manner have become urgent tasks, and resampling method is one of the key technologies for this type of data. In other words, the algorithm design as the first step in machine learning should be followed by the statistical analysis in which the difference between the mathematical model and natural model will be examined carefully.

In machine learning, there are four essential elements—data acquisition, sampling strategy, specific learning algorithm and ensemble method. The authors hope the current situation of excessive attention being paid to the performance of algorithm while the characteristic of data being ignored will be changed. Especially, the characteristic of the data is paid more attention to in this paper. Accuracy is not the exclusive target of the algorithm. Meanwhile, interpretability of the model is very critical in practice. Therefore, the accuracy and the interpretability of a model are of equal importance during the statistical modeling. Resampling methods extends the types of data which can be handled by learning algorithms while improves the interpretability of the model.

This work is supported by the National Basic Research Program of China (973 Program) under grant No. 2004CB318103. The main research of this project is to solving and developing the theory and applications of data modeling in large-scale with complex characteristics. The researches involve in machine learning theory and applications.