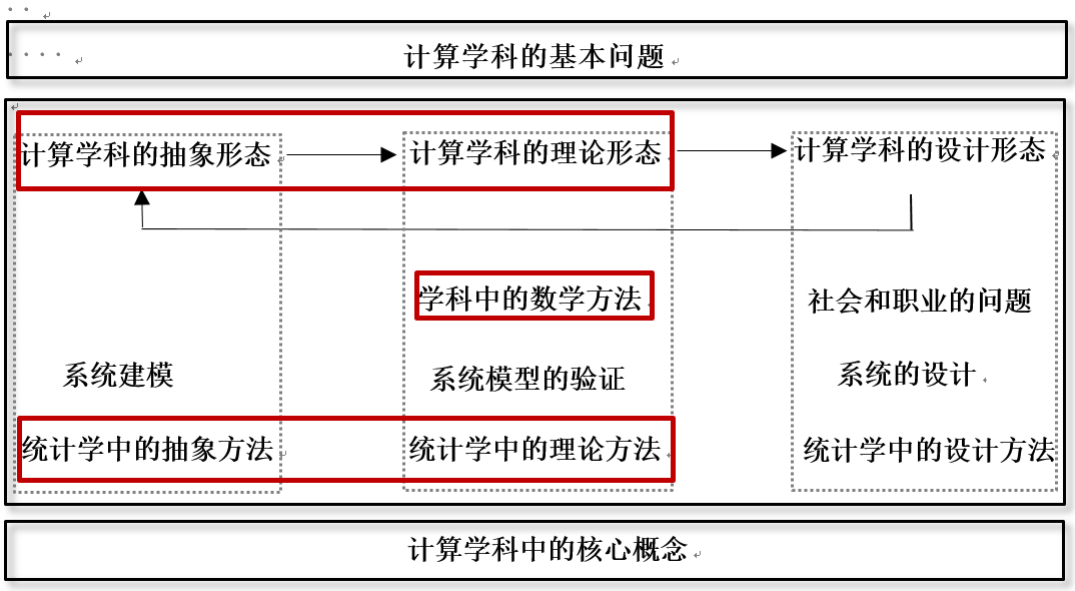
大数据计算中的计算思维——PAC 算法

李廉　　合肥工业大学

能力的评估：本案例能够置于 Bloom 分类法知识维度的“程序性知识”位置，学生学习后能够达到 Bloom 分类法认知过程维度的“分析”层次（针对大三学生，需先修概率论与数理统计）。

一、本案例课程思政的关注点

1.本案例内容在计算学科课程思维总体框架中的位置



2.科学思维可拆分为可衡量、可检验的抽象、理论和设计三个过程（学科形态，或工作范式）。本例包含抽象及理论形态的内容，其中，PAC 算法描述、误差和可靠性定义可划分到抽象形态，PAC 算法中关于误差、可靠性及随机抽取样本数之间的关系公式的证明属于理论形态。学习该案例后，学生对抽象、理论和设计三个学科形态如何区分将有进一步的认知，这种认知将为我国在三个学科形态方面的工具（含思想与方法）的创新，实现“0 到 1”的突破种下科学思维的种子。

3. 在本案例中，要求教师将 11 个品行元素中的“目标驱动、专业性、严谨”与该案例绑定在一起进行可操作性解释。

目标驱动：在大数据集中找到其中的最大元素，不能简单采用常规的搜索算法，因为会消耗过多的计算资源。如何在消耗资源较少的前提下，找到事实上可以接受的目标，是驱动的目标；

专业性：从概率学的角度理解这种近似算法的科学性、合理性；

严谨：对 PAC 算法能给出严谨的数学证明。

二、本案例的具体内容（讲解脚本）

1. 背景介绍

- 同学们好，本讲介绍一种大数据计算中的计算思维方法，即 PAC 算法。
- 我们来看这样一个大数据计算中的典型问题：设 S 是超大规模数据集合（包括流数据），如何找出其中的最大元素。
- 传统的方法（数学思维）是通过反复比较，保留当前的最大元素，直到整个数据被扫描一遍。当数据集合的元素个数为 N 时，这个算法需要 N 步。在大数据场景中，这个算法有时难以实现。
- 由于数据量巨大，存储空间和计算时间都受到了限制，无法满足精确计算的要求，这时放弃计算最优解的数学思维，利用局部数据，进行不精确计算，求得可行解，从而极大减少计算所需要的资源。
- 也就是寻找近似算法，综合考虑资源优化，使得最后得到的结果保证在一定的误差之内。
- 在这样的目标驱动下，数学家和计算机科学家们尝试使用各种方法，希望达到这样的目标：通过抽取大数据集中的一小部分数据，找出与集合中最大元素充分接近的元素D。
- 本讲要介绍的“PAC 算法”，就是一种典型的近似算法，通过引入两个概念，即求解结果的误差（近似性），以及小于这个结果的概率（可靠性），设计充分小的误差和充分大的可靠性，找出事实上可以接受的解决答案。这个方法称为概率近似正确方法（Probably Approximately Correct，简记为 PAC 方法）。
- 本讲的案例参考了我国著名数学家、计算机科学家李廉教授关于 PAC 算法的论文。

2. PAC 算法的形式化描述

- 我们把 PAC 算法所想要解决的问题再描述一下：
  设 S 是一个数据集合，具有海量的数据，甚至无穷多数据，即$|S| = N$（∞），或者是一个流动的数据集，随着时间不断有数据流入，这时如何计算最大元素。
- 在抽象形态层面，可以形式化地描述 PAC 算法如下：

  算法A称为 PAC 算法，如果对于$0 < \varepsilon, \gamma < 1$，

  $$Pr[Error(A(S), F(S)) \leq \varepsilon] > \gamma.$$

- 下面介绍这个定义里面各个部分的含义：
  - ε称为误差，γ称为可靠性概率。这两个要素也就是前面提到的评价求解结果质量的两个指标，近似性和可靠性，表示计算所得到的结果与实际结果误差小于ε的概率为γ；
  - $A(S), F(S)$分别表示对于数据集合S，PAC 算法的计算结果和问题的实际结果；
  - 设D为 PAC 算法A计算的结果，定义$Error(A(S), F(S)) = Pr\{x \geq D \mid x \in S\}$，即$S$中数据不小于$D$的比例——当 S 有限时，$Error$即为$S$中实际最大元素大于算法$A$得到的最大元素$D$的可能性；
  - $Error$还有一个等价的定义：$Error(A(S), F(S)) = |\{x \geq D \mid x \in S\}|/|S|$，其中$|S|$表示$S$中元素的个数。
- 在 PAC 算法中，当ε很小时，可以认为D就是S的最大元素，其可靠性要大于γ。也

就是通过 PAC 算法A找出了一个事实上可以接受的解决答案。

- PAC 算法将概率统计学应用到算法设计中，结合了计算思维与数学思维，并从概率学的角度给出了误差与可靠性的数学定义，很好地体现了计算科学中的算法的专业性。

3. PAC 算法中的样本量与误差的关系

- 对于前面介绍的 PAC 算法，一个直观的看法是，随着样本个数的增加，这个误差会越来越小，可靠性会越来越大，如果抽取全部数据，则结果就是精确解。
- 但是，从专业性的角度，样本个数与误差、可靠性之间的关系应该是可以量化的，这样 PAC 算法才能使人信服。
- 因此下面就介绍就是如何定量地估计样本数量与误差和可靠性之间的关系。
- 样本个数 N 与误差ε以及可靠性γ之间存在这样一个关系式：

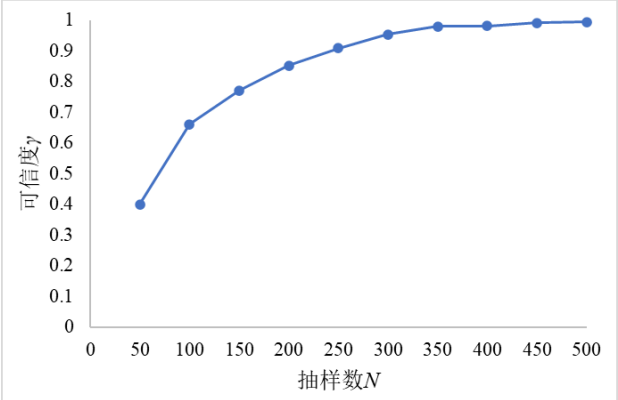
$$N + 1 \geq log(1-\gamma) \,/\, log(1-\varepsilon).$$

- 这个命题明确了样本个数与误差和可靠性之间的关系，我们可以通过这个公式，设计所需要的ε和γ，抽取需要的样本，达到可行的求解目标。
- 由于ε和γ在公式中都是以对数的方式出现，所以样本数量随着ε的减少和γ的增加不会太快，这个算法的实用性还是很好的。
- 从公式出发，我们代入几个典型的数据，观察样本数量与误差和可靠性之间的关系：
  - 当$\varepsilon = 0.05$，$\gamma = 0.95$，$N \geqslant 58$；
  - 当$\varepsilon = 0.01$，$\gamma = 0.99$，$N \geqslant 458$；
  - 当$\varepsilon = 0.01$，$\gamma = 0.99$，$N \geqslant 686$；
  - 当$\varepsilon = 0.01$，$\gamma = 0.99$，$N \geqslant 915$.
- 从下面的表格和图片中，可以更具体地看出三者之间的关系：

表 1．$\varepsilon = 0.01$时不同**抽样数N的可靠性**

抽样数N	可靠性γ
50	0.4
100	0.661
150	0.772
200	0.853
250	0.909
300	0.955
350	0.981
400	0.982
450	0.992
500	0.995

图 1. 误差 0.01 时可靠性的增加曲线



- 利用 PAC 算法，可以在样本量并不显著增加的情况下，得到具有很高可靠性的结果。

4. PAC 算法中关键结论的证明

- 上面总结的 PAC 算法步骤并不复杂，但为了确保算法的正确性，必须给出各个步骤的严谨数学证明。而严谨、细致也是设计算法的过程中的必须品行。
- 下面简单介绍先前的一个关键结论的证明：

➢ 结论：样本个数 N 与误差ε以及可靠性γ之间存在关系式：

$$N+1 \geq log(1-\gamma) \,/\, log(1-\varepsilon).$$

➢ 证明：

从S中随机抽取N个数据，其中最大者为D．设S中小于D的元素占比u满足：

$$1-\varepsilon < u \leq 1,$$

即D是最大元素的概率为u．在抽样结果Q已经发生的前提下，有：

$$Pr(0 \leq u \leq 1-\varepsilon|Q) = (1-\varepsilon)^{N+1}.$$

这一公式的具体推导如下：

$$Pr(0 \leq u \leq 1-\varepsilon|Q)$$

$$= Pr(Q|0 \leq u \leq 1-\varepsilon)\, Pr(0 \leq u \leq 1-\varepsilon) \,/ Pr\,(Q)$$

$$= \left(\frac{1}{1-\varepsilon}\int_0^{1-\varepsilon} u^N\, du\right)(1-\varepsilon) \Big/ \int_0^1 u^N\, du$$

$$= (1-\varepsilon)^{N+1}.$$

这里利用了贝叶斯公式、全概率公式，并且假定 u 是均匀分布。

令$(1-\varepsilon)^{N+1} \leq 1-\gamma$，即

$$N+1 \geq \log(1-\gamma)/\log(1-\varepsilon),$$

则

$$Pr(u > 1-\varepsilon|Q) = 1 - P\,r(0 \leq u \leq 1-\varepsilon|Q) \geq \gamma.$$

从而证明了命题。

三、教学体会

本案例通过抽象-理论-设计的划分能很好的帮助教学，将复杂知识分解，降低教学的复杂性，更好的进行教学设计。其中在抽象层面，将 PAC 算法中涉及的重要概念进行归纳，

建立概念模型，首先帮助同学们建立一个宏观印象和初步认识。在理论层面，则对算法中涉及的关键步骤和重要结论进行分析、证明，帮助同学们从不同层次上对算法进行全方位的理解。

四、激励、唤醒和鼓励同学们向上的途径

PAC 算法的关键思想，就是要针对近似的、概率的算法引入误差、可靠性这两个指标，给出抽样数量与误差、可靠性的定量关系，并对该量化关系进行严格的数学证明。通过这个案例，鼓励同学们采用严密的数学方法，对算法中涉及的关键过程进行推导及证明，并利用合适的数学工具对算法设计进行宏观指导。

五、习题

1. 设 PAC 算法的抽样结果为Q：从S中随机抽取N个数据，其中最大者为D. 记S中小于等于D的元素占比为u. 求证：当u为$[0,1]$上的均匀分布时，$Pr(Q)=\int_0^1 u^N\,du$.

参考答案：

证明：

当S中小于等于D的元素占比为某个固定值u时，抽样结果Q发生的概率为$Pr_u(Q)=u^N$.

假设u的概率密度函数为$f(u)$，则复合概率$Pr(Q)=\int_0^{+\infty} Pr_u(Q)\,f(u)\,du$.

当u为[0,1]上的均匀分布时，$f(u)=\begin{cases}1\ (0\le u\le 1)\\ 0\ (u>1\text{ 或 }u<0)\end{cases}$，可得$\int_0^{+\infty} f(u)du=1$

代入到公式中得到

$$Pr(Q)=\int_0^{+\infty} f(u)u^N du=\int_0^1 u^N\,du$$

2. 计算海量数据中的最大元素时，如果逐元素扫描则时间复杂度过大，这时可以采用概率近似正确方法（Probably Approximately Correct，PAC），通过从海量数据中抽样，在抽样数据中找出最大元素。当抽样次数达到足够多时，则可找到最大元素的可行解。该算法中可以得到 1 个命题：

$$N+1\ge log(1-\gamma)\,/\,log(1-\varepsilon)$$

其中，N为抽样次数，ε为误差，γ称为可靠性概率。

（1） 当误差为 0.05，可靠性要达到 0.95 时，需要多少抽样次数才能获得可行解。

（2） 当误差为 0.5，可靠性要达到 1-2^(-22113)时，需要多少抽样次数才能获得可行解。

参考答案：

解：

（1）

$$N+1\ge log(1-0.95)\,/\,log(1-0.05)$$
$$N\ge 57$$

需要至少抽样 57 次。

（2）

$$N+1\ge log(2^{-22113})\,/\,log(1-0.5)$$

$$N \geq 22112$$

需要至少抽样 22112 次。

参考文献

1. 杨矫云，郭思伊，李廉，基于 PAC 算法的流数据 Top-k 实时查询，华中科技大学学报（自然科学版），2018 年 46 期.