统计、计算和未来

李 航
华为诺亚方舟实验室

关键词：统计加计算 机器学习

前言

科学研究发现，许多高等动物拥有空间意识、数量意识[1]，甚至社会关系意识，可是没有证据表明这些动物拥有时间意识——这也许是人与动物的主要区别之一。人可以面对未来进行想象、展望和规划，而动物却不具备这一能力[2]。

渴望了解未来、热衷预测未来是人的本性。哲学家克尔凯郭尔曾说："人生的过去可以去解读，但未来只能去经历 (Life can only be understood backwards; but it must be lived forwards)"。这也许是大家关心未来的根本原因。

从人类文明的摇篮期开始，预测未来就有着重要意义。3500 年前的商朝，人们笃信通过占卜可以预测未来，出征、收成、婚嫁、生育，甚至晴雨、出行、狩猎都要进行占卜。流传下来的甲骨文实际上多是刻在龟甲兽骨上的占卜记录。

但人类真正拥有科学的预测未来的手段是从 17 世纪开始的——这个工具就是**统计加计算**，发展至今成为**统计机器学习**。随着科学技术的发展，统计加计算变得愈来愈强大，不断改变人类的生活方式，给人类带来巨大的影响。

帕斯卡的故事

统计加计算的第一位代表人物应是 17 世纪法国科学家帕斯卡 (Blaise Pascal)。帕斯卡幼年丧母，体弱多病，由父亲带大成人。但他聪明伶俐，勤奋好学，十几岁时就展现出非凡的才华，可惜的是他只活了 39 岁。他短暂的人生里在数学、物理、哲学等多个领域取得了辉煌的成就。

图1 帕斯卡(1623~1662)

看到作为税务官的父亲整天忙于繁杂的数字计算，帕斯卡产生了开发数字计算机器的想法。他 19 岁时（1642 年）发明了人类历史上第一个机械计算

图2 帕斯卡发明的计算器

[1] 是指判断数量多少的能力，而不是指识别数字的能力。

[2] 动物不拥有时间意识的观点来自物理学家加来道雄(Michio Kaku)。

器，可以进行加减乘除运算，其基本原理是通过齿轮进行进位。帕斯卡先后研制了 50 多个样机，终于在 3 年后发布了实用品（图 2）。

帕斯卡在数学上也有很深的造诣。有人向帕斯卡求教，希望帮助解决当时的难题“点数问题 (problem of points)”，帕斯卡通过与另一位法国著名数学家费尔玛往来 6 封书信，一起找到了问题的答案。

其解法如下：假设两个人赌博，通过多局博弈争抢一笔奖金。两人每局获胜的概率均是 1/2，多局博弈后，最先赢得指定局数的人将获得所有奖金。假设由于某个原因，赌博必须中途终止（两人都未赢得指定的局数），应该如何分配奖金才算合理？这个问题看上去简单，实则不然。因为两人平分奖金，或者按照赢的局数比分配都有一定的不合理性。费尔玛和帕斯卡提出的解决办法是：假设一个人需要再赢 r 局才能获得奖金，另一个人需要再赢 s 局才能获得奖金，那么两人应该按以下比例分配奖金：

$$\sum_{k=0}^{s-1}\binom{r+s-1}{k} \Big/ \sum_{k=s}^{r+s-1}\binom{r+s-1}{k}$$

其思路是，计算两个人在剩下的比赛中所有获胜的可能性，并按比例分配。

在讨论过程中，帕斯卡提出了相当于“数学期望”的概念，被认为是人类历史上第一个有关现代概率论的论述[3]。概率的本质是计算未来的可能性，而统计的本质是通过对过去的观察计算未来的可能性。从帕斯卡的时代（1654 年）开始，概率统计这一预测未来的工具逐步发展起来，成为科学技术的重要支柱。

上述解法主要依赖于组合数计算。为了能够很快地进行计算，帕斯卡又发明了著名的“帕斯卡三角形”[4]，是由组合数（即二项式系数）组成的三角形阵列（如图 3）。从三角形顶点的组合数 $\binom{0}{0}=1$ 开始，自上而下一层层地计算三角形底边各个点的组合数 $\binom{n}{k}$，每个点的组合数等于它的左上方与右上方两个组合数之和。它的原理是：

$$\binom{n}{k}=\binom{n-1}{k-1}+\binom{n-1}{k}$$

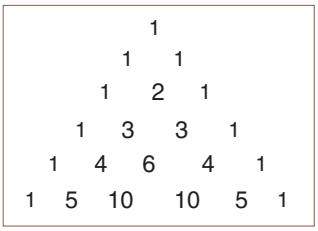


图3　帕斯卡三角形

帕斯卡的工作涉及统计加计算的所有重要概念：系统、算法及概率。

霍列瑞斯的故事

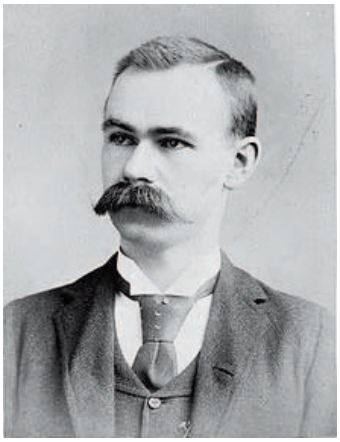


图4　霍列瑞斯(1860~1929)

另一位与统计加计算密切相关的人物是 19 世纪美国的发明家霍列瑞斯 (Herman Hollerith)。美国每 10 年进行一次人口普查，之前的人口普查都是手工统计，直到 1890 年遇到了问题。因为根据上一次的经验估计，完成这次普查需要 13 年的时间，这是无法接受的。霍列瑞斯发明的穿孔制表机 (punched card tabulator) 解决了这一问题，只用 6 年的时间就完成了人口普查的统计工作，如图 5。霍列瑞斯后来创建了自己的公司，专门生产穿孔制表机，这家公司后来发展成为著名

[3] “概率”这个词是在18世纪初贝努利的著作中第一次出现的。

[4] 在中国被称为 “贾宪三角形”，或“杨辉三角形”，分别比帕斯卡三角形早600年和400年。

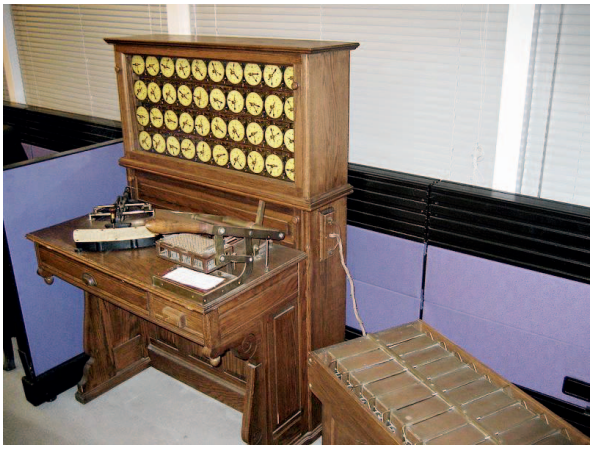


图5 霍列瑞斯发明的穿孔制表机

的 IBM 公司。

穿孔制表机被认为是当代计算机的前身，可以用于大规模数据的统计。穿孔制表机的操作原理是，首先在卡片上记录样本的属性，一个样本用一张卡片。如果某个样本的某个属性为 1，就在对应卡片的对应位置上打一个孔，记录这一信息。然后用制表机对大量的穿孔卡进行统计处理，一次处理一张卡片。有多根探针，一根探针对应一个属性；如果卡片的某个位置有孔，则该位置的探针就能从孔穿过，与下面的金属相连，形成电路，对应的属性就被记录一次。这样，就可以对大量样本的属性进行快速的统计。

从霍列瑞斯的时代开始，电子计算机逐渐发展起来，成为人类有史以来最强大的工具。

统计加计算以及机器学习

统计加计算的核心是计算未来各种可能性的大小。假设我们把骰子投掷到桌上，如果我们能够准确测量出骰子的初始速度和角度、空气的阻力、桌面的弹力与摩擦力，基于物理原理和数学分析方法，我们就能精确地计算出骰子落到桌面时哪个面朝上。如果我们只关心点数为 1 的那一面朝上的可能性，那么这种计算就没必要，因为按照概率统计原理，其可能性是 1/6。概率统计的特点就是忽略过程和细节，而只关注结果。客观世界是极其复杂的，通过解析的方法很难计算出结果，而概率统计可以帮助我们计算出各种结果出现的可能性（当然在很多情况下，计算各种可能性也需要非常复杂的计算）。这一功能使统计加计算变得非常实用，这就是它作为预测未来工具的强大之所在。

人的判断不完全是客观和理性的，会受到情感、经验、兴趣、习惯等主观因素的影响，也会受到信息的不完全与不准确的制约，而且人的计算与存储能力远远不及计算机。统计加计算会帮助人做出合理的判断，并提供强大的工具。

从 20 世纪 50 年代开始，统计加计算逐步演进成为统计机器学习，也就是通常人们说的机器学习。机器学习的核心是利用统计学原理，基于数据进行建模，通过模型对客观事物进行分析、预测和判断，已经成为一个拥有严密与完整理论的领域。

机器学习技术可以由概率、系统、算法三个维度来描述（见图 6）。概率是统计的基础，系统和算法是计算的基础。迄今为止，统计机器学习已有大量的算法，比如决策树、提升方法、核方法、神经网络、贝叶斯方法、图模型、非参数化方法、稀疏方法，诸多的系统（比如 Hadoop、Spark）被开发出来，并应用在各个领域。

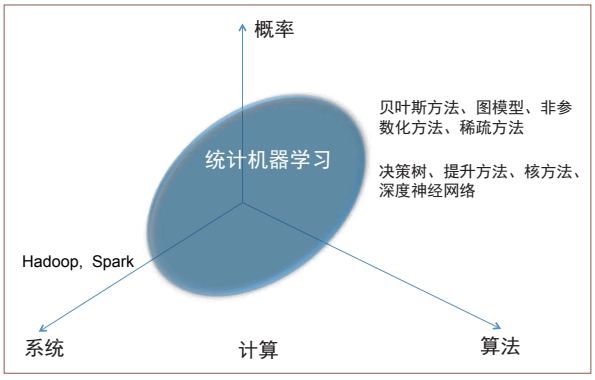


图6 统计机器学习技术概观

现在是大数据时代，计算机的处理能力呈指数级增长，积累的数据量也呈指数级增长，这给机器学习带来巨大的机遇和挑战，需要不断开发更多更快更准的平台、算法和工具。通过机器学习手段预测未来的技术在不断演进，取得了丰硕的成果。

通信网络中的预测

大数据给通信带来了巨大的挑战，其规模大、速度快、种类多的特点在通信中也有充分体现。例如，在移动骨干网，流量带宽为40G~100Gbps；在大型数据中心，流量带宽可达到1Tbps规模。如何对大数据传输的通信网络进行有效的预测、控制和管理成为一个极具挑战的问题。机器学习有望成为实现这一目标的有力工具。

在通信网络中，规模庞大的数据流被称为大象流(elephant flow)，规模微小的数据流被称为老鼠流(mouse flow)。大象流出现的比例虽然不高，但会占据网络带宽，给网络带来堵塞。所以，及时检测、预测大象流，采取必要的措施是网络管理的重要课题。

华为诺亚方舟实验室开发的技术可以很好地对大象流进行检测与预测，前者已应用到华为的产品中。

大象流检测

大象流检测是指当数据流源源不断通过的时候，统计其规模大小，判断其是否是大象流（有大量的数据包）。我们开发的LD-Sketch算法可以在线高速地检测通过的大象流。

完成这一任务的一个简单方法是，用哈希函数将数据流索引，当数据流到来时，利用哈希索引对数据流的规模进行快速统计，规模超过一定阈值的数据流是大象流；在这个过程中，将规模小的数据流丢弃，以防止数据溢出。但这个方法存在两个问题。一个是哈希函数会有碰撞(collision)，不同的数据流被统计在一起，检测时产生伪正(false positive)；另一个是实际的大象流（在还没有达到阈值时）被错误丢弃，检测时产生伪负(false negative)。

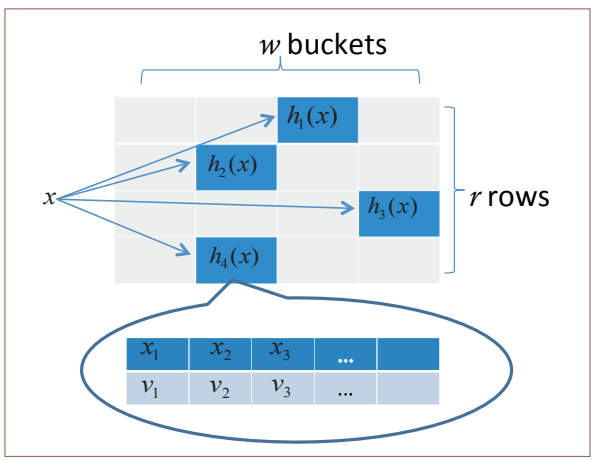


图7 LD-Sketch算法的数据结构：哈希函数阵列及相关的计数阵列

LD-Sketch算法能够很好地解决以上问题。图7显示了LD-Sketch算法的数据结构，包括一个哈希阵列，每一个哈希阵列的元素有一个相关的计数阵列。LD-Sketch算法通过使用哈希阵列减少伪正，通过使用相关的计数阵列保证没有伪负。

首先，用多个不同的哈希函数对数据流进行索引，将结果保存在哈希阵列的不同行中，每一行的每一个元素对应哈希函数的一个具体取值、一个具体的桶(bucket)。针对某个哈希函数，取值相同的数据流会被放到同一个桶中，即产生哈希碰撞。为避免哈希碰撞，可以增加哈希函数的个数，即阵列的行数，这样有希望将不同的数据流均匀地分到不同的桶中。

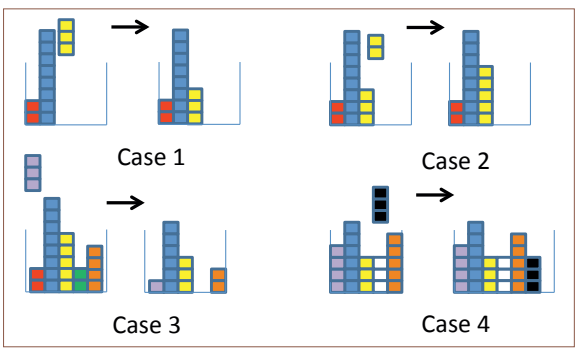


图8 LD-Sketch算法：计数阵列更新的四种情况

分配到哈希阵列的每个元素，即每个桶内的数据流的规模在相关的计数阵列中被统计。规模超过阈值的数据流被认为是大象流。图8以模拟“俄罗斯方块”游戏的方式直观地显示了在计数阵列中统计数据流规模的四种不同情况。第一种情况，计数阵列未满，新的数据流加入，这时在阵列中加入新的数据流的统计；第二种情况，计数阵列未满，已有的数据流加入，这时在阵列中更新已有数据流的

统计;第三种情况，计数阵列已满，新的数据流加入，假设数据流的整体规模小于某个阈值，这时从所有的数据流中统计减去最小的数据流统计，将计数矩阵部分清空（类似俄罗斯方块），加入新的数据流的统计;第四种情况，计数阵列已满，新的数据流加入，假设数据流的整体规模大于某个阈值，这时扩大计数矩阵，在阵列中加入新的数据流的统计。丢失数据流信息只有在第三种情况时出现，但因为减去的是最小的数据流统计并有记录，理论上不会错过大象流，即不会产生伪负。

LD-Sketch 算法的空间复杂度和时间复杂度都达到了理论上的最优水平，其性能在实际测试中超过了已有的所有算法。

大象流预测

大象流预测是指数据流刚到的时候，系统只看它的包头信息 (packet header information)，就能判断其是否是大象流。这好比是邮递员通过邮件的单据来判断邮件的大小，如果有一定的规律，这个预测也能做到很准。大象流预测在局域网管理（比如数据中心的网络管理）中非常有用。

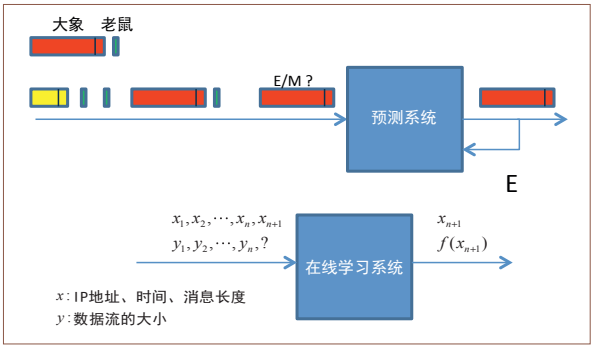


图9 大象流预测问题

大象流预测可以定义为在线回归学习问题 (online regression learning)，一个数据流是一个样本，由特征向量和流的规模组成，如图 9 所示。特征向量表示数据流的地址、时间等信息。学习系统不断地从数据中学习回归模型，再对于新给定的样本基于特征对其规模进行预测。我们的方法假设预测模型属于高斯过程回归 (Gaussian Process Regression)，从流数据中在线学习该模型。

高斯过程回归是非参数化的、多元回归问题的学习方法。假设要学习的函数遵循高斯过程，通过已给训练样本和要预测样本的联合高斯分布，给出要预测样本的期望和方差，也就是做出预测，如图 10 所示。

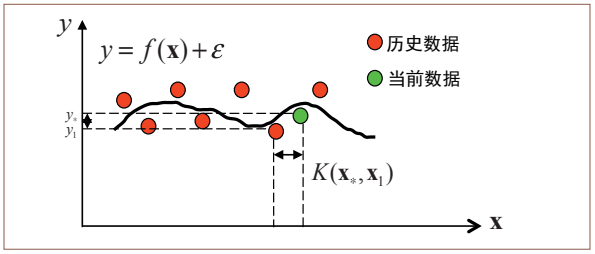


图10 高斯过程回归

具体地说，假设已给训练样本$(\mathbf{X}, \mathbf{y})$，要预测样本$(\mathbf{x}_*, y_*)$，训练样本与预测样本遵循多元高斯分布：

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim N(\mathbf{0}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma_n^2 \mathbf{I} & K(\mathbf{X}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{X}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix})$$

其中$N(\mathbf{0}, \mathbf{\Sigma})$表示均值为$\mathbf{0}$，协方差矩阵为$\mathbf{\Sigma}$的多元高斯分布，$K(\cdot,\cdot)$表示核函数，$n$是训练样本个数，$\sigma_n^2$是噪音的方差，$\mathbf{I}$是单位矩阵。预测样本的期望和方差如下：

$$\bar{y}_* = E[y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*] = K(\mathbf{x}_*, \mathbf{X})[K(\mathbf{X},\mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1} \mathbf{y}$$

$$\mathrm{cov}(y_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{X})[K(\mathbf{X},\mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1} K(\mathbf{X}, \mathbf{x}_*)$$

如果用所有n个样本进行学习，计算复杂度是$O(n^3)$，现实中不可以接受。我们的方法是在线学习算法，先用m个样本学习模型，然后增量式地学习，每次处理一个样本，部分更新模型，重复这个过程，这样算法复杂度变成$O(m^2 n)$。在标准数据集上大象流预测的准确率达到 97.9%，处理一个数据流的时间仅是 10 微秒。

未来的预测

我们正在经历人类历史上前所未有的变化与进步，物理世界中的人、物、事，精神世界中的思想、情感，都在信息世界中被数字化，并永久地保留下来，而且这个过程的规模和速度在不断地扩大，这

也是大数据现象发生的根本原因。大数据、统计加计算将会不断地、大幅度地提高我们预测未来的能力……

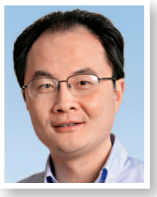
畅想未来，我们工作、生活的方方面面也许都可以用基于统计加计算的“预测机器”来辅助。这个预测机器会在人生的每一章节、每一个片段对未来做出精准的预测。比如，预测机器能帮助你设计出行计划：如果你想到某一个地方，告诉机器目的地，它就会给你路径、交通方式、时间、成本、舒适度等多个选项。预测机器还能帮助你管理健康，可告诉你两天以后得感冒的概率非常高，还会给你饮食、睡眠、运动方面的建议。预测机器也能帮助你择偶。如果你考虑跟某人结婚，它能帮助你预测婚姻成功的指数是多少，未来十年二十年离婚的概率有多大。预测机器还能帮助你选择就业。如果你考虑到某家公司工作，它能够估算你的合适程度以及预测未来在这家公司成功的概率，未来十年二十年可能的收入是多少。所有这些之所以能成为可能，是因为在信息世界里每个人的大量信息都被记录下来了，使得预测机器有强大的统计加计算的预测能力。

结语

可以预见，随着更庞大的数据的积累，更强大的计算机的开发，更准确的机器学习方法的发明，我们将能够更好地预测未来。从帕斯卡的时代至今仅有四个世纪，人类预测未来的能力已发展到了令人惊叹的水平，展望今后预测未来技术的发展，会让我们感到无比的兴奋与向往。商朝时人们凡事都要占卜，将来人们也许凡事都要预测，这一点是我们可以想象得到的。■

致谢：

华为诺亚方舟实验室黄群、陈志堂、何诚、耿彦辉详细介绍了他们的大象流工作，加州大学伯克利分校郁彬教授分享了霍列瑞斯的故事，在此对他们表示感谢。



李 航

CCF专业会员。华为技术有限公司诺亚方舟实验室主任。主要研究方向为信息检索、自然语言处理等。

HangLi.HL@huawei.com

参考文献

[1] Blaise Pascal, Problem of Points, Herman Hollerith, Tabulating Machine, Wikipedia.

[2] Huang, Lee. Ld-sketch: A Distributed Sketching Design for Accurate and Scalable Anomaly Detection in Network Data Streams. InforCom, 2014.

[3] Poupart, Chen, Fung, and Geng. Proactive Network Routing Control System with Flow Size Prediction. 2016, to appear.

CCF教育工作委员会在京召开会议

2016年4月27日，新一任CCF教育工作委员会（简称“教育工委”）在北京召开首次工作会议。CCF秘书长**杜子德**，教育工委主任**杜小勇**、副主任**唐卫清**、**张铭**及20余名委员出席会议。杜小勇主持会议，并介绍了教育工委的定位和职责。CCF常务理事、中科院计算所研究员唐卫清介绍了计算机工程教育认证的相关情况，北京大学教授张铭介绍了计算机知识体系的相关情况。

会议讨论明确了2016年度教育工委工作计划和分工。针对高校计算机教育培养目标分类标准制定的问题，教育工委拟从对毕业要求的分层（基本、高级）和对知识体系的分类（系统、软件、应用）两个角度开展工作。