重尾分布与轻尾分布

潘镇涛　21721273

1　承前启后

在启发觅食算法的生物行为基础中，如果捕食者像无头苍蝇一样漫无目的的做布朗运动，真的能很有效的地寻到到猎物吗? 在数学上可以证明，布朗运动和分子自由扩散一样，单位速度的分子在时间t内平均只有$\sqrt{t}$的位移量。若捕食者采用这种策略，需要很久才能够成功。因而基于布朗运动的觅食算法的效率也十分低下，很难适用于大多数现实应用场景。

如何设计更高效的搜索方法呢? 一组巴西物理学家于1999年提出了一个设想，认为“莱维飞行”比布朗运动有更高的搜索效率，因此自然会偏向与采用“莱维飞行”捕食的生物，基于“莱维飞行”设计更为高效的觅食算法。为了理解“莱维飞行”的概念，需要对概率分布的轻尾性和重尾性有深入的认识。

2　概率分布

2.1　指数分布

指数分布是描述泊松过程中的事件之间的时间的概率分布，即事件以恒定平均速率连续且独立地发生地过程。它是几何分布的连续模拟，它具有无记忆的关键性质。例如，如果T是ij一元件的寿命，已知元件使用了t小时，它总共使用至少$s+t$小时的条件概率与一开始使用时算起它使用至少s小时的概率相等。在$x \to \infty$的时候，指数分布是以指数的速度趋近于0的。因此，以指数分布为分界线，来区分轻尾分布和重尾分布。

指数分布的概率密度函数：

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

2.2 轻尾分布

比指数分布尾部更薄的概率分布是轻尾分布。在$x \to \infty$的时候，它们下降到0的速度比指数分布的速度快得多，因此在尾部质量更小。也就是说，在远离峰值的尾部区域，时间发生的概率更低一些。Gumbel 分布是轻尾分布的一个例子，以及正态分布也属于轻尾分布。轻尾分布往往并不能很好地反映“真实世界”的数据。所以正态分布用来对那些主流事件发生较多，非主流事件发生较少的情况进行建模更为合适。

2.3 重尾分布

当$x \to \infty$时,重尾分布下降到0的速度慢于指数分布。因而尾部质量较大。在PDF的曲线下将会有更多的体积。重尾分布往往会有很多异常值和非常高的值。重尾分布的尾巴越重，样本中出现一个或多个不成比例的值的概率就越大。其数学定义为:

$$\lim_{x\to\infty} e^{\lambda x} F(x) = \infty$$

其中，$\lambda > 0$, $F(x)$是尾分布函数。

如果从重尾分布中随机抽取样本，很有可能会由于有一个或多个非常大的值会导致统计信息出现巨变。例如，如果从国人的收入中抽取样本，那ô抽取的大部分样本数据将相对较少。但是，容易抽到一个或两个值可能会很大（即异常值）的样本。比如，先抽出99 个人的收入样本，大约人均5万元，然后突然抽到一个千万富豪。这些大数值倾向于歪曲您的样本统计数据：平均数会非常误导（对于上述示例，数额会达到数百万美元），样本方差可能会非常大，样本均值通常会低估总体平均数。另外，重尾分布不受中心极限定理限制。

重尾分布更适用于对那些离峰值较远的稀有事件也会有相当的概率发生的情况。重尾分布作为一个大的类别，还包含三个重要的子类别，分别是肥尾分布（Fat-tailed distribution），长尾分布（Long-tailed distribution）和次指数分布（Subexponential distribution）。

2.4 长尾分布

长尾分布，或者说长尾理论是一个与互联网发展分不开的概念。说到这里就不得不先提一下传统商业中的帕累托法则（Pareto principle），又称为二八定律。比如80%的财富集中在20%的人手里，图书馆里20%的书可以满足80%的顾客。于是大家往往只关注在PDF 图中最左面的20%的顾客，以期满足80%，来实现效益的最大化。

但在一些网上零售业中，如Amazon和Netflix，数据表明右端黄色的尾巴虽然平均需求小但是由于数量巨大，导致其总的营销收益甚至超过主流的商品。这一发现似乎对商业界的触动极大，也说明了正确建模的重要性。如果用指数分布进行建模，这些远端的需求也许就会被忽视；而用长尾分布进行建模就可以发现这些新的需求从而带来效益的提高。长尾分布的数学定义为：

$$\lim_{x\to\infty} Pr\left(X > x + t | X > t\right) = 1$$

就是说，当x很大的时候，很有可能x实际上更大。

2.5 肥尾分布

从建模的角度来看，肥尾分布就是针对那些罕见事件虽然发生的概率低，但也必须要考虑到的情况。比如一个保险公司考虑灾害的发生和保险的定价，那ò像自然灾害这种情况，如果不考虑的话就可能面临真的发生时要赔很多的情况。因为正如肥尾分布的名字所体现的，即使在远离峰值的远端，那些罕见事件还是有相当的概率会发生的。虽然我们常常用正态分布对很多时间进行建模，但当一个事件的本质是肥尾分布而我们误用了正态分布或指数分布时，就存在着对“小概率事件真的发生”这种危险的低估。据说美国股市历史上的黑色星期五，千禧年的互联网泡沫破灭，以及2008年前后的金融危机都是这种错误的真实案例。

肥尾分布的数学定义为:

$$\lim_{x\to\infty} Pr\left[X > x\right] \sim x^{-\alpha}, \alpha > 0$$

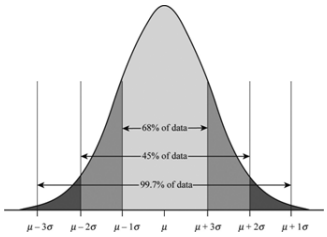
也就是说，在x 较大的地方，肥尾分布趋于0的速度是明显慢于指数分布和正态分布的。柯西分布（Cauchy distribution）就是一类典型的肥尾分布。

3 觅食算法中的分布特性

3.1 觅食算法中的轻尾性

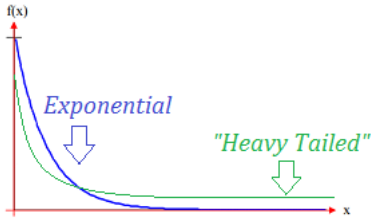
前面说过，布朗运动的步长服从正态分布。正态分布是一种典型的轻尾分布, 也就是说，不太可能取得极端值的分布。

表达式$y = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$。其中$\mu$是均值，$\sigma$是标准差。当$x$增大时，函数下滑趋势非常快，因此数学家又将正态分布函数纳入速降函数空间范畴。速降函数是专门为傅里叶分析而量身定做的。



3.2 觅食算法中的重尾性

而在现实生活中，大多数的事情不能用轻尾分布来描述，例如保险领域的保险金-事故可以看作是稀有事件，但一旦发生并通过审核，保险公司就必须大量金额，这样一来保险金的分布就很可能取到很大的值。为了描述这样容易取得极值的随机变量，我们需要引入重尾分布。



典型特点: 大头短+小尾长。”莱维飞行”的平均位移是t^{γ},$\gamma > \frac{1}{2}$.

莱维过程$\{X(t), t \geq 0\}$是一种随机过程，它满足的条件比布朗运动宽松:

- $X0$ 几乎处处为0;
- 独立增量性;
- 稳定增量性;
- 样本轨道右连续.

连续的布朗运动和离散的泊松过程都是莱维过程的特例。因此可以大胆猜测，莱维过程就是带“跳跃”的布朗运动。正是这些不连续性的“跳跃”给予莱维过程“重尾”的特性。已经有相当多的研究表明很多动物的移动模式可以用莱维过程来描述。而近些年通过对人类的移动数据（通话次数、出租车等）的挖掘,也惊奇地发现人类的移动模式也和莱维过程高度吻合。

4 总结

本文从基于布朗运动设计的觅食算法切入，介绍了基于布朗运动的觅食算法的缺陷，引出莱维飞行在觅食算法中的应用。进一步的，本文较深入的介绍概率分布中的重尾分布和轻尾分布，进而介绍觅食算法的重尾性和轻尾性，最后给出了莱维飞行在觅食算法成功被应用的原因。