

文章编号:2095-6134(2015)01-0136-04

简 报

基于核方法和高斯混合模型的手机上网时长 统计分析及应用*

常楠, 张三国†

(中国科学院大学数学科学学院, 北京 100049)
(2014 年 3 月 3 日收稿; 2014 年 3 月 28 日收修改稿)

Chang N, Zhang S G. Mobile network access time statistical analysis and its application based on kernel method and Gaussian mixture model[J]. Journal of University of Chinese Academy of Sciences, 2015,32(1):136-139.

摘 要 从不同手机型号角度,分析上网时长的分布,并用核方法和高斯混合模型对手机上网时长数据进行建模和分析。实际数据分析表明,手机上网时长分布具有双峰现象,本文对此现象给出了合理解释。

关键词 移动设备; 手机; 上网时长; 核密度估计; 高斯混合模型

中图分类号:O213.9 文献标志码:A doi:10.7523/j.issn.2095-6134.2015.01.022

Mobile network access time statistical analysis and its application based on kernel method and Gaussian mixture model

CHANG Nan, ZHANG Sanguo

(School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract Based on kernel method and Gaussian mixture model, we analyze the distribution of the mobile network access time from a new view, the mobile model view. The analysis of real data suggests that the distribution of the mobile network has two peaks. We give some reasonable interpretations for such a phenomenon.

Key words cellular device; mobile phone; access time; kernel density estimator; Gaussian mixture model

随着 3G 和 4G 网络的兴起,各式各样的移动设备开始普及,智能手机也渐渐成人们生活的必需品。Gokul et al.^[1]曾经指出,每个人用的移动设备比如智能手机一定程度上反映出这个人的行为特征,所以用同一类型智能手机的人在一定程度

上有着相同的特性。比如喜欢用 iphone 的人一般都比较年轻时尚,喜欢用功能机的人普遍比较守旧踏实,在经济上也要比用 iphone 的人要差一些。更深层次地说,用不同手机型号的人可能在性格、年龄、职业和经济情况上都有着不同的特性,

* 教育部回国留学人员科研启动基金及中国科学院大学校长基金(Y05101AY00)资助

† 通信作者, E-mail: sgzhang@ucas.ac.cn

因此有必要从手机型号着手来探究用户习惯. 移动上网方便是智能手机最重要的特性之一, 智能手机与移动上网已经紧紧地联系在了一起, 使用不同型号手机的人在移动上网方面也有不同的特征, 从不同型号手机上网数据来研究用户习惯, 也就成了一个新颖而有效的手段.

手机是大众使用频率最高的移动设备. 据我们了解, 迄今为止, 手机上网数据由于涉及到隐私和法律方面的因素, 一般实际数据难以获得. 相关方面公开的研究工作也不是很多. Falaki et al. [2] 曾经研究手机流量特征和用户行为的多样性, 但是他们只用了 255 个用户的数据. Huang et al. [3] 研究影响手机网络应用用户认知的因素, 他们的工作主要集中于个别的手机型号, 包括苹果, 三星和 palm 手机. Shafiq et al. [4] 测量上网流量的时间和空间的动态, 结果表明, 不同的设备型号有着不同的上网模式. 从以上看, 尚未有学者就众多手机型号的角度研究用户习惯. 本文的数据采集于 300 万个手机, 这 300 万个手机来自 769 个手机型号, 为了让分析更有意义, 我们采集到每个手机连续 7 天的上网时长总和, 然后把每个型号手机的平均上网时长当做这个手机型号的上网时长代表. 那么每个手机型号的上网时长代表用这个手机型号的这一类人的上网数据. 通过统计分析我们发现一个有趣的现象, 这 769 个手机型号的上网时长在 70 min 和 350 min 的时长点上密度比较高, 即代表着用手机在一周内上网时长为 70 min 和 350 min 的人要多一点.

1 方法

通过对数据的整理发现使用 iphone 的人群一周内平均上网 1 745. 431 min, 平均每天 249. 347 min, 而所有手机品牌一周内的平均上网时长为 619. 887 4 min, 中位数为 660. 485 9 min, 说明使用 iphone 手机的人群平均上网时长要多于大多数其他手机品牌. 但单纯地分析一个手机品牌上网时长意义是不大的, 有必要探讨这 769 个手机型号在一周内平均上网时长有着怎样的分布. 图 1 是各种型号手机平均上网时长的分布直方图, 从中可以看出, 在 50 ~ 100 和 350 ~ 400 min 的密度最高. 单凭直方图无法给出有效的量化峰值点在哪里, 所以有必要用光滑的核方法来处理.

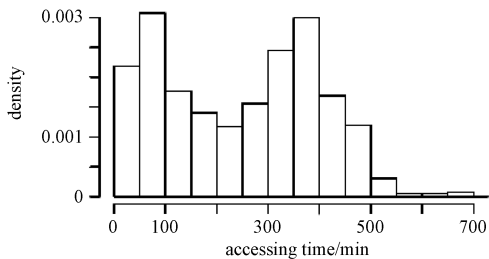


图 1 各种型号手机平均上网时长分布直方图

Fig. 1 Histogram of average mobile network access time on models

1.1 核方法

对来自于密度为 $f(x)$ 的样本集 $S = \{x_i\}_{i=1, \dots, n}$, 任意 x 点的核密度估计 (KDE) 为

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x_i - x),$$

其中, K 是核函数, 满足 $K(t) \geq 0$ 和 $\int K(t) dt = 1$, $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$, 这里 h 为窗宽. 常用的核函数有以下几种:

1) 高斯核函数

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty;$$

2) Epanechnikov 核函数

$$K(x) = \frac{3}{4} (1 - x^2) I(|x| \leq 1);$$

3) Biweight 核函数

$$K(x) = \frac{15}{16} (1 - x^2)^2 I(|x| \leq 1).$$

一般来说, 核函数对于估计结果的影响并不大, 但是窗宽的选择却对估计结果有很大影响. 常见选择窗宽的方法有拇指法则、最小二乘交叉验证方法、似然交叉验证法. 本文采用最小二乘交叉验证方法, 该方法的想法是选取 h 使得估计的平均最小方差最小. \hat{f} 和真实函数 f 的积分均方差为

$$\begin{aligned} I_n &= \int [\hat{f}(x) - f(x)]^2 dx \\ &= \int [\hat{f}(x)]^2 dx - 2 \int \hat{f}(x) f(x) dx + \int f(x)^2 dx. \end{aligned}$$

我们注意到上式的最后一部分和窗宽 h 没有关系, 第二部分可以这样来估计:

$$\begin{aligned} \int \hat{f}(x) f(x) dx &= \frac{1}{n} \sum_{i=1}^n n \hat{f}_{-i}(X_i) = \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i}^n k_h(x_i - x_j), \end{aligned}$$

这里的

$$\hat{f}_{-i}(x_i) = \frac{1}{n-1} \sum_{j=1, j \neq i}^n k_h(x_i - x_j)$$

是 $f(X_i)$ 的减一核估计. 对于第 1 部分, 我们知道

$$\begin{aligned} \int [\hat{f}(x)]^2 dx &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \int k_h(X_i - x) k_h(X_j - x) dx \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{k}(X_i - X_j), \end{aligned}$$

这里的 $\bar{k}(v) = \int k(u)k(v-u) du$. 这里选择 h , 使得

$$CV_f(h) = I_{(1n)} - 2\hat{I}_{2n}$$

最小, 这里的

$$I_{(1n)} = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{k}(X_i - X_j),$$

$$2\hat{I}_{2n} = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k_h(X_i - X_j)$$

记为 \hat{h} . 设 \hat{h} 是上述交叉验证问题的解. 具体可参考文献[5].

图 2 是生成的直方图和核密度估计的叠加图, 从图 2 可以看出, 分布图是一个双峰图, 这样就促使我们想到一个特殊的参数模型——高斯混合模型(GMM). 对于参数方法, 如果模型假设正确, 那么它的估计效率优于非参数方法.

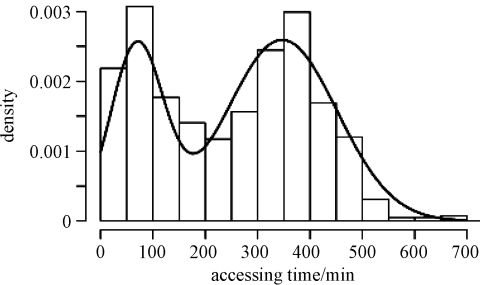


图 2 各种型号手机平均上网时长分布核方法估计图

Fig. 2 Plot of average mobile network access time, estimated using the kernel method, on models

1.2 高斯混合模型

一般地, g 个高斯密度混合模型, 其密度如下:

$$f(y; \Theta) = \sum_{i=1}^g \pi_i \phi(y; \mu_i, \Sigma_i),$$

其中, $\Theta = \{\pi_i, \mu_i, \Sigma_i\}, i = 1, \dots, g, \phi(y; \mu_i, \Sigma_i)$ 是一个均值为 μ_i 协方差阵为 Σ_i 的 p 维正态分布, π_i 满足 $\pi_i \geq 0, \sum_{i=1}^g \pi_i = 1, 0 < i \leq g$, 设有 n 个独立观察样本 $y_j, 1 \leq j \leq n$, 那么 Θ 的对数似然函数为

$$\log L(\Theta) = \sum_{j=1}^n \log \left\{ \sum_{i=1}^g \pi_i \phi(y_j; \mu_i, \Sigma_i) \right\},$$

其模型参数的极大似然估计一般采用 EM 算法^[6-7]来计算.

设 z_1, \dots, z_n 是没有观察到的示性向量, 其中, z_j 的第 i 个元素 z_{ij} 取值为 0 或者 1, 如果 y_j 来自于第 i 个高斯模型, 那么 $z_{ij} = 1$, 否则为 0. 令 $x_i = (y_i^T, z_i^T)^T, 1 \leq i \leq n$, 那么 $y_i = (1 \leq i \leq n)$ 和 $x_i (1 \leq i \leq n)$ 分别称为不完全数据和完全数据. 对于 Θ 的完全数据的对数似然为

$$\log L_c(\Theta) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \{ \log \pi_i + \log \phi(y_j; \mu_i, \Sigma_i) \}$$

在 EM 算法中上式中的 z_{ij} 被当做缺失数据, 在第 $t+1$ 步迭代中, 记

$$\tau_i(y_j; \Theta^t) = \pi_i^t \phi(y_j; \mu_i^t, \Sigma_i^t) / \sum_{h=1}^g \pi_h^t \phi(y_j; \mu_h^t, \Sigma_h^t)$$

是 y_j 属于第 $i (i = 1, \dots, g; j = 1, \dots, n)$ 个高斯模型的后验概率, 那么 E 步计算 Q 函数 $Q(\Theta; \Theta^t)$ 为

$$\begin{aligned} Q(\Theta, \Theta^t) &= \sum_{i=1}^g \sum_{j=1}^n \tau_{ij}(y_j; \Theta^t) \{ \log \pi_i + \\ &\quad \log \phi(y_j; \mu_i, \Sigma_i) \}, \end{aligned}$$

在此基础上, 分别定义:

$$T_{i1}^t = \sum_{j=1}^n \tau_i(y_j; \Theta^t),$$

$$T_{i2}^t = \sum_{j=1}^n \tau_i(y_j; \Theta^t) y_j,$$

$$T_{i3}^t = \sum_{j=1}^n \tau_i(y_j; \Theta^t) y_j y_j^T,$$

则 M 步的参数更新为

$$\pi_i^{t+1} = T_{i1}^t / n,$$

$$\mu_i^{t+1} = T_{i2}^t / T_{i1}^t,$$

$$\Sigma_i^{t+1} = \{ T_{i3}^t - T_{i1}^{t-1} T_{i2}^t T_{i2}^{t-1T} \} / T_{i1}^t,$$

不断重复上述 2 步, 直到收敛.

2 实际数据分析和结果

接下来将以上方法应用于实际例子. 首先对各种不同手机型号上网时长分布进行核密度估计, 鉴于高斯核的广泛应用性, 这里采用高斯核, 关于窗宽的选择通过最小二乘交叉验证方法, 计算得窗宽 $h = 36.81$, 而 2 个峰值分别在 68 min 和 354 min. 对于参数方法的高斯混合模型, 由于实际数据是一个双峰问题, 所以采用 2 个高斯密度混合模型进行拟合, 并且每一个高斯模型都为

维, 方差记为 σ_j^2 , 得到上述参数 θ 估计为

$$\theta = \begin{cases} \pi_1 = 0.316; \mu_1 = 69.3, \sigma_1^2 = 2\,557.3; \\ \pi_2 = 0.684; \mu_2 = 346.9, \sigma_2^2 = 11\,074.3. \end{cases}$$

为了方便比较, 我们把原分布的直方图、非参数核密度估计, 参数高斯混合模型的结果画在图 3 中。

图 3 展示了各种型号手机平均上网时长分布的高斯混合模型图。图中包含直方图、GMM 拟合曲线和 KDE 估计曲线。横轴为访问时长 (min)，范围从 0 到 700；纵轴为密度 (density)，范围从 0 到 0.003。GMM 拟合曲线 (实线) 和 KDE 估计曲线 (虚线) 均显示了双峰分布特征，峰值分别位于约 70 min 和 350 min 处。

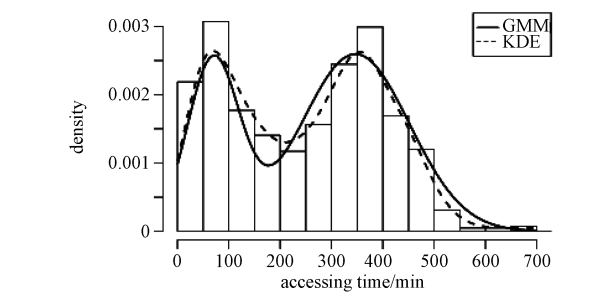


图 3 各种型号手机平均上网时长分布高斯混合模型图

Fig. 3 Plot of average mobile network access time, estimated using Gaussian mixture model, on models

用高斯混合模型得到的 2 个峰值分别在 70 min 和 350 min 时, 此结果和核方法估计的结果相差不多。有趣的是此结果和我们先前的认知是不一样的: 在分析数据之前, 我们原以为手机上网时长分布应该是单峰的, 上网时长多的和上网时长少的所占比例较小。前面已经指出, 每一个手机型号代表着某一类型的人, 也就是说上述现象可以解释为用手机上网的人在上网时长较少和上网时长较多的 2 个时长点上比较多, 并不是我们想象的用手机上网时长适中的占最高比例。进一步查看每周上网 70 min (平均每天 10 min) 的高峰点的手机型号, 多是一些评价良好的功能机, 比如 samsung-s7350c 和 nokia5000 等, 这些功能机由于屏幕和运行速度限制等, 不能运行过于复杂的网络应用, 因此上网时间不长。运营商在为使用这些品牌手机的客户制定套餐时, 应该多推荐一些上网流量少的。另一个高峰期在 350 min, 也就意味用这部分手机型号的客户每天会有将近 1 h 花在手机上网, 这些手机型号主要包括 nokia700 和 mt917 等中端智能机, 这些手机价格合理, 上网方便受到很大一部分人群的青睐。运营商可以适当地为这部分手机品牌的客户推荐一些流量比较多

的套餐。我们还注意到, iphone 等高端智能机等虽然上网时间很长, 但是由于价格因素等的限制, 使用人群并没有中端智能机多。

3 展望

本文从手机型号角度对手机上网时长进行统计分析, 用核方法和高斯混合模型拟合数据, 表明其分布具有双峰现象, 并给出双峰原因的合理解释。我们下一步的工作还需要结合以下 2 个方面, 进行更深一步的探讨: 第 1, 双峰的出现与用户的套餐资费以及所上的网站有无较大关系, 如果有, 关系如何体现; 第 2, 文中虽然对模型进行了拟合, 但是并没有对拟合效果进行评估, 应该通过一些拟合优度检验来评价。

References

[1] Gokul C, Jan B, Daniel G P. Who's who with big-five: analyzing and classifying personality traits with smart-phones [C] // Proceedings of the 15th Annual International Symposium on Wearable Computers. 2011.

[2] Falaki H, Mahajan R, Kandula S, et al. Diversity in smartphone usage [C] // Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services. 2010.

[3] Huang J, Xu Q, Tiwana B, et al. Anatomizing application performance differences on smartphones [C] // Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services. 2010.

[4] Shafiq M Z, Ji L, Liu A X, et al. Characterizing and modeling internet traffic dynamics of cellular devices [C] // Proceedings of the ACM Sigmetrics Joint International Conference on Measurement and Modeling of Computer Systems. 2011.

[5] Li Q, Jeffrey S R. Nonparametric econometrics: theory and practice [M]. Princeton University Press, 2011.

[6] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society. Series B (Methodological), 1977, 39(1): 1-38.

[7] Wu C F J. On the convergence properties of the EM algorithm [J]. Ann Statist, 1983, 11(1): 95-103.