专题引子

关于聚类与差异化度量

徐华

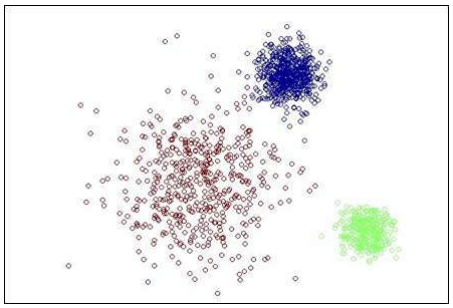
清华大学 计算机系 智能技术与系统国家重点实验室

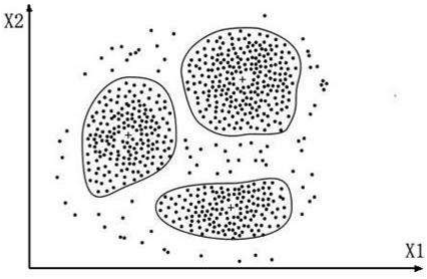
xuhua@tsinghua.edu.cn

1

关于聚类

- 现实生活中我们常常有数据，但是不知道要分成几种类型
  - 类型标签未知：数量和具体标签都不知道
- 相对于分类问题，此类问题不可能具有提前贴好类型标签的数据集（训练数据集和测试数据集）
- 有时可能知道类型标签的数量（k），而具体标签未知

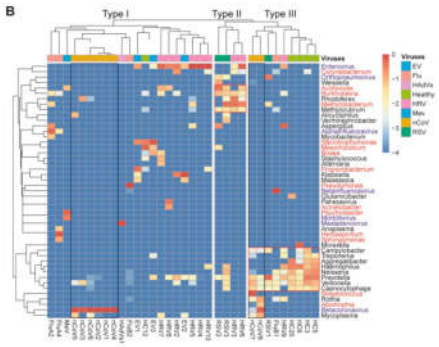




2

关于聚类的例子1

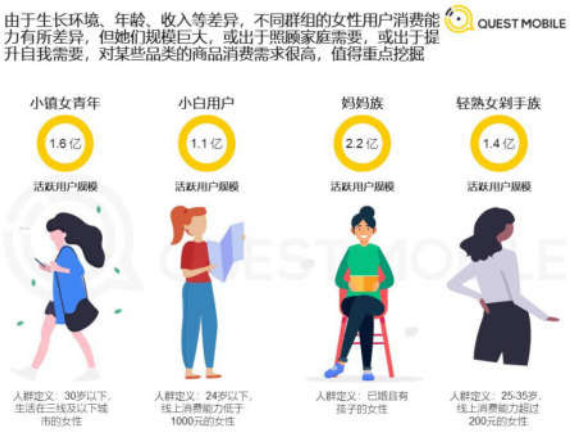
- **关于新冠病毒感染人群的聚类**
  - 新冠病毒感染者的样本聚类分析表明：某些样品的微生物群比较贫乏。研究仍然发现**新冠病毒感染者**和**社区感染性肺炎患者**均与健康对照组不同（新冠病毒感染者与健康组：R2=0.45，p =0.001；社区感染性肺炎患者与健康组：R2=0.10，p=0.002），这表明他们的肺部菌群发生了微生态失调。



3

关于预测的例子2

- **中国女性消费人群划分（1）**



4

关于预测的例子3

- **中国女性消费人群划分（2）**

作为零售平台一直以来争夺的目标群体，女性仍是这场宅经济背后的主角。除了遥控出门采购的老公，她们依据各类疫情信息，通过线上买买买，为自己和全家人置办各类商品，解决吃喝玩乐等问题，一手防疫一手托起家庭生活。

敏感型 守望家庭健康 口罩消毒液泡腾片成标配	悦己型 年轻女孩种草拔草 全方位护肤一步不能少
居家型 线上抢菜买水果 生鲜食材逆势增长	育儿型 宅家带娃“吃喝”、“玩乐”、“学习”三不误
娱乐型 带娃减肥两不误 瑜伽垫游戏机销量猛增	逆行者 医务工作者最挂念孩子 十大品类中七个与娃相关

5

关于预测的例子4

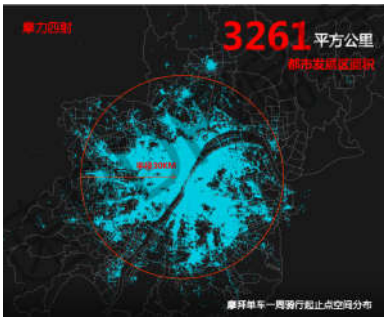
- **拍照风格的划分**

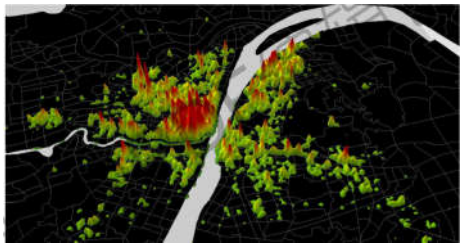
拍照风格 在拍照风格上，男女都最爱日杂风；女性更喜欢萌趣和港风，男性则偏爱文艺与潮酷

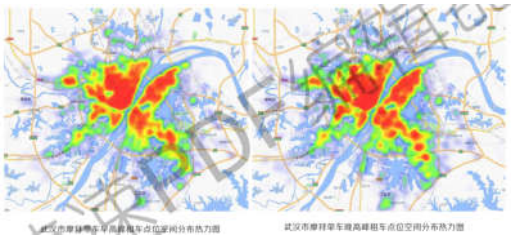
6

关于预测的例子5

- 根据共享单车使用频度对活跃区域的划分（1）





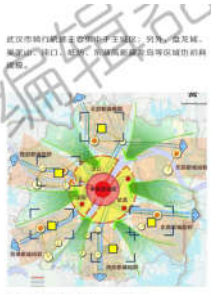


武汉市摩拜单车一周骑行起止点空间分布热力图　　武汉市摩拜单车晚高峰租车点位空间分布热力图

7

关于预测的例子6

- 根据共享单车使用频度对活跃区域的划分（2）



8

关于预测的例子7

- 同一聚集的类型：相似程度大；不同聚集的类型：差异程度大
- 如何衡量相似度和差异化程度

9

关于聚类分析

- 如何做聚类？
- 如何实现对于数据对象的相似度评价？

10

Thanks !