DATAGURU 炼数成金

Hadoop应用开发实战案例 第12周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

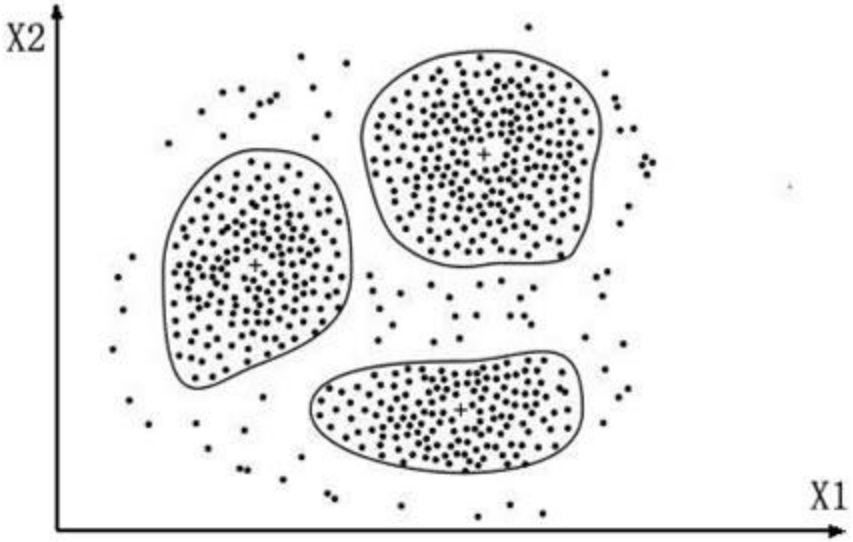
课程详情访问炼数成金培训网站

http://edu.dataguru.cn

聚类

什么是聚类？聚类和分类判别有什么区别？

聚类可以用于哪些场景？



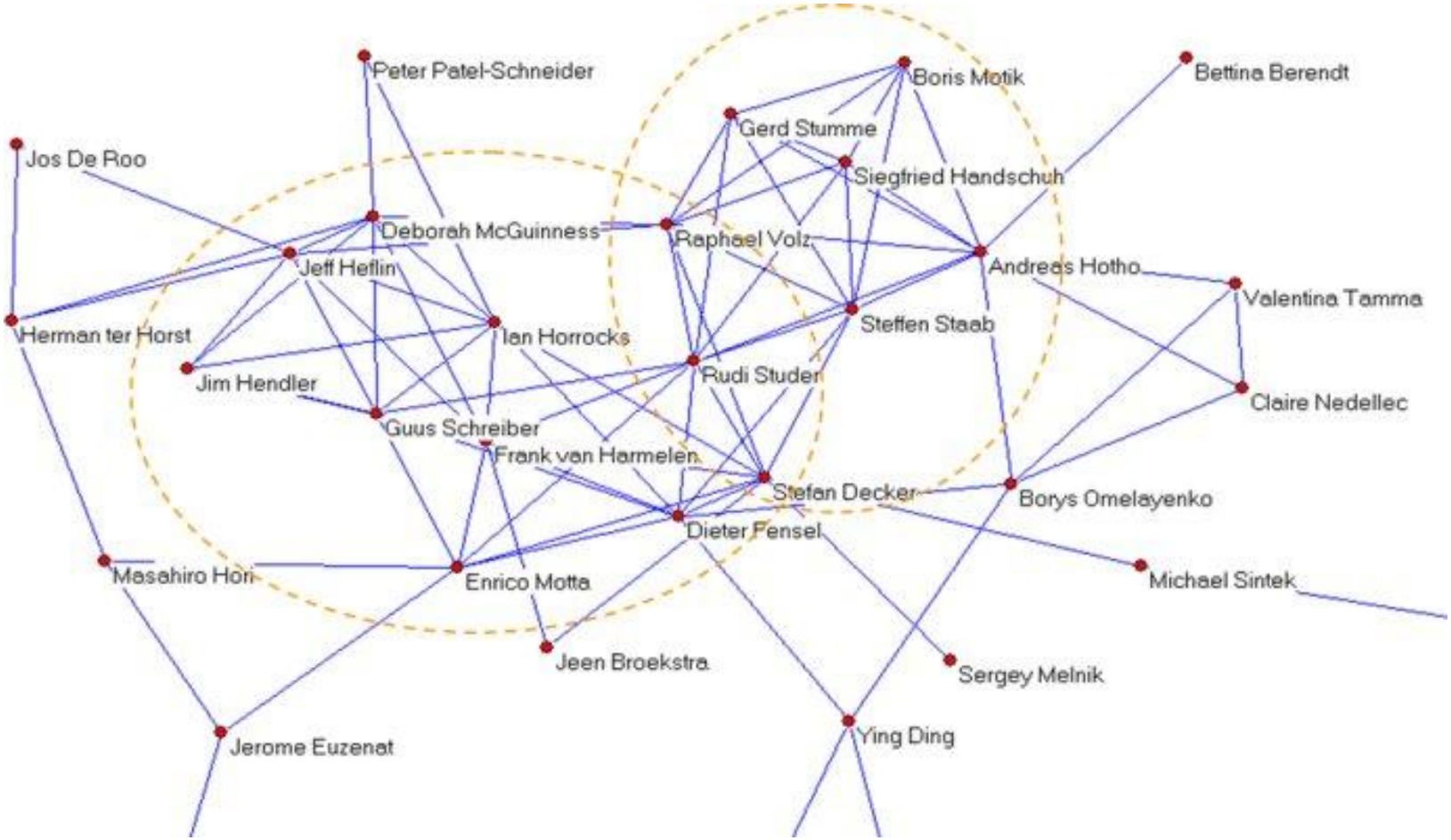
聚类应用场景：寻找优质客户

- 二八定律无处不在
- 20%的用户提供了银行80%的利润来源
- 20%的用户消费了运营商话费总额的80%
- 公司中20%的员工完成了80%的工作
- 社会中20%的人拥有80%的话语权

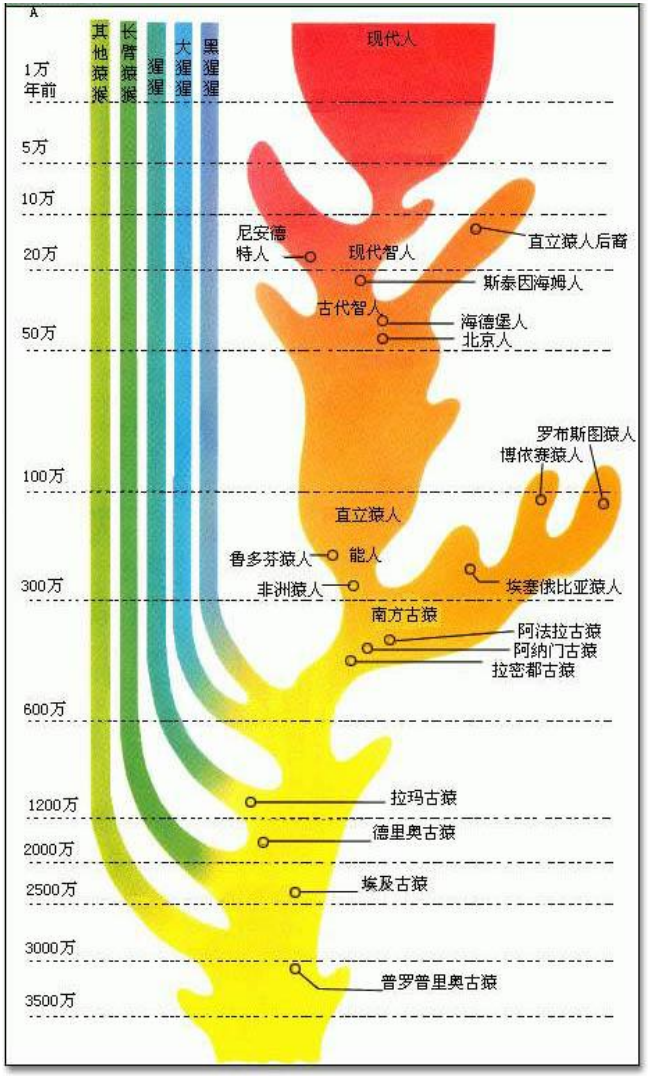
聚类应用场景：推荐系统


图书 > 计算机/网络 > 软件工程/开发项目管理 > 商品详情
看过本商品的还看了
¥137.00
Logitech/罗技 无线鼠标M545_深沉黑_激光级技术
(52条评论)
¥65.00
【当当自营】logitech罗技 M185 无线鼠标 (灰色
(931条评论)
¥40.70
大规模分布式存储系统:原理解析与架构实战（阿
(870条评论)
Recommender Systems
An Introduction
推荐系统
推荐系统(推荐系统必读经典，百度技术委员会主席廖若雪、新浪微博数据挖掘技术专家张俊林、人民搜索商务部总监常兴龙、百分点信息科技有限公司首席运营官张韶峰联袂推荐！)
当当价 ¥47.20 （8折）
定 价 ¥59.00
评 论 99.2%推荐 353条
配送至 广东广州市海珠区 有货 运费说明 本商品提供礼品包装服务
下周一(4月14日)可送达，请在17小时1分钟内下单并选择 "普通快递送货上门"
丛书名 图灵程序设计丛书
作 者 （奥地利）詹尼士 等著，蒋凡 译
出版社 人民邮电出版社
出版时间 2013-7-1
ISBN 9787115310699
所属分类 图书 > 计算机/网络 > 软件工程/开发项目管理
分享到：
送积分 472
查看大图
批量购买入口>>
我要买 1 件
加入购物车
一键购买
收藏商品
收藏人气：1
最佳拍档
¥47.20
推荐系统(推荐系统必读经典，百度技
¥39.20
推荐系统实践(《浪潮之巅》、《数学
¥48.70
机器学习实战【利用Python透析主流
¥45.60
社交网站的数据挖掘与分析（2011年J
¥739.00
【当当金牌店】WD西部数据 My Passp
¥89.00
【当当自营】logitech罗技 M215二代
新版调查
回顶部

聚类的应用场景：社区发现



聚类应用场景：生物进化树



聚类应用场景：孤立点的特殊意义

- 信用卡诈骗
- 黑客攻击

```
xmenu=1&inajax=1" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.045
06.30)"
183.3.51.76 - - [29/Nov/2013:01:27:25 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.
50727; .NET CLR 3.0.04506.30)"
183.3.51.76 - - [29/Nov/2013:01:27:26 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.
50727; .NET CLR 3.0.04506.30)"
183.3.51.76 - - [29/Nov/2013:01:27:26 +0800] "POST /member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1 H
TTP/1.1" 200 297 "http://r.dataguru.cn/member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1" "Mozilla/4.0
(compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.04506.30)"
66.249.64.1 - - [29/Nov/2013:01:30:19 +0800] "GET /home.php?mod=space&uid=50144&do=home&view=me&from=space HTTP/1.1" 200 5769
 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X) AppleWebKit/536.26 (KHTML, like Gecko) Version/6.0 Mobile/10A5376
e Safari/8536.25 (compatible; Googlebot-Mobile/2.1; +http://www.google.com/bot.html)"
66.249.64.8 - - [29/Nov/2013:01:30:44 +0800] "GET /space-uid-73446.html HTTP/1.1" 200 4782 "-" "Mozilla/5.0 (compatible; Goog
lebot/2.1; +http://www.google.com/bot.html)"
210.51.177.136 - - [29/Nov/2013:01:35:28 +0800] "GET / HTTP/1.0" 200 46531 "-" "User-Agent: Mozilla/5.0 (compatible; MSIE 6.0
;Windows XP)"
66.249.64.1 - - [29/Nov/2013:01:36:52 +0800] "GET /space-uid-73384.html HTTP/1.1" 200 4776 "-" "Mozilla/5.0 (compatible; Goog
lebot/2.1; +http://www.google.com/bot.html)"
66.249.64.1 - - [29/Nov/2013:01:38:25 +0800] "GET /space-uid-73345.html HTTP/1.1" 200 4434 "-" "Mozilla/5.0 (compatible; Goog
lebot/2.1; +http://www.google.com/bot.html)"
183.3.20.129 - - [29/Nov/2013:01:38:45 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0
.50727; .NET CLR 3.0.04506.30)"
183.3.20.129 - - [29/Nov/2013:01:38:49 +0800] "GET /member.php?mod=logging&action=login HTTP/1.1" 200 17707 "http://r.dataguru
.cn/member.php?mod=logging&action=login" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0
.50727; .NET CLR 3.0.04506.30)"
183.3.20.129 - - [29/Nov/2013:01:38:49 +0800] "POST /member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1
HTTP/1.1" 200 297 "http://r.dataguru.cn/member.php?mod=logging&action=login&loginsubmit=yes&inajax=1&ajaxmenu=1" "Mozilla/4.0
 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.04506.30)"
[root@class2room web_logs]#
```

项目背景

- 电信用户行为偏好分析

特殊人群：

- 电影？
- 网购？
- 运动？
- .....

电信数据源

- CDR-1x
  - 语音通话记录
  - 短信
  - 少部分上网数据
- CDR-do
  - 主要是上网时产生的数据
- CDR-do stream
  - 流数据

字段名	字段
IMSI	IMSI
呼叫状态	CALL_STATE
接入时间	ACCESS_TIME
主（被）叫号码	DIALED_DIGITS
呼叫标志	CALL_OR_CALLED_OR_HHO
呼叫持续时间	CALL_DURATION
最终的业务选项	FINAL_SERVICE_OPTION
BSC	BSC
释放导频_CellID	RELEASE_CELL
释放扇区_SectorID	RELEASE_SECTOR
释放经度	RELEASE_LONGITUDE
释放纬度	RELEASE_LATITUDE

优质用户

- 具有明显特征、有一定价值的用户

- 思考1：从电信的呼叫记录中可以挖掘哪些类型的优质客户？
- 思考2：此类用户具有什么比较明显的特征？

电信数据中挖掘优质用户

- 高端商务人群
  - 通话较为频繁，月平均话费高
  - 出差频率高，机场出现率高

电信数据中挖掘优质用户

- 异地情侣
  - 有固定联系人，且与该联系人通讯次数频繁
  - 通话平均时间长，通话时间段在晚上的频数多

电信数据中挖掘优质用户

- 广告用户（电话推销、垃圾短信等）
  - 呼出次数远远大于呼入次数
  - 平均通话时间短、平均呼叫间隔短
  - 与呼出对象的通信总次数很低
  - 固定联系人很少

- 快递员人群
- 乘机用户
- 网购用户
- 文艺用户
- ……

分析方法

- 多维关联筛选
  - 确定用户通信行为特征：

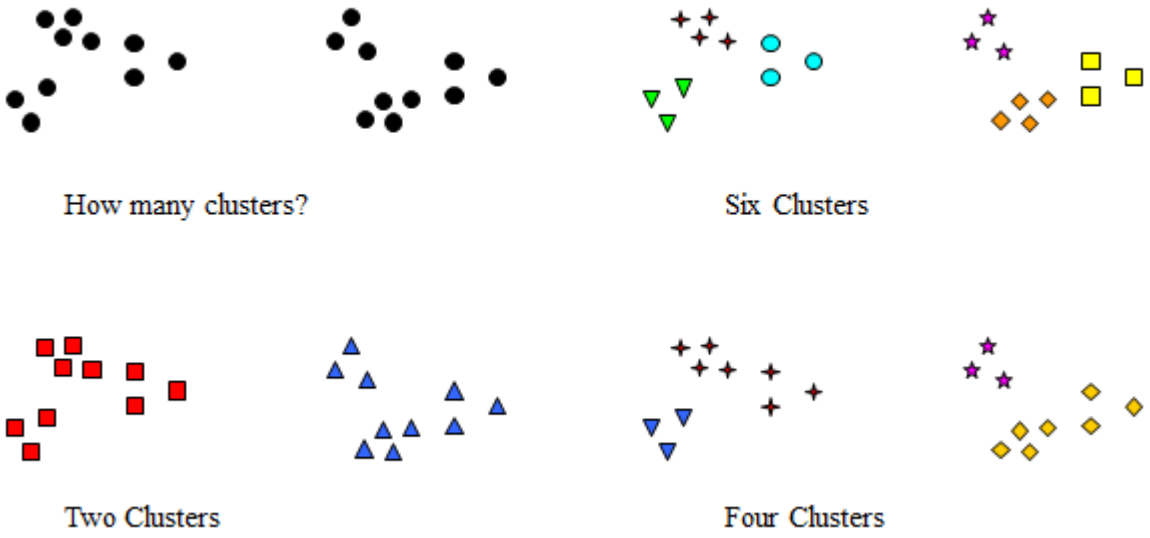
    （通信时间，通信对象，通信频次，通话时长，上网流量，通信业务）
  - 多维关联提取
- 分类器
  - 贝叶斯分类、决策树、随机森林
  - 回归分析、支持向量机、神经网络
- 聚类分析(Cluster Analysis)

聚类分析

- What is Cluster Analysis ?

  Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

- 特征：类内相近，类外相远，无监督的分类



相似度

- 欧式距离

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

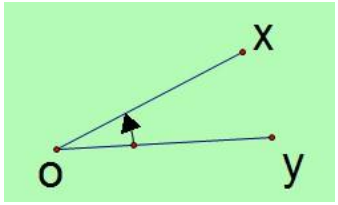
- 闵可夫斯基距离

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r}$$

r=1, k=2, |x1-y1|+|x2-y2|

- 余弦相似度

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$



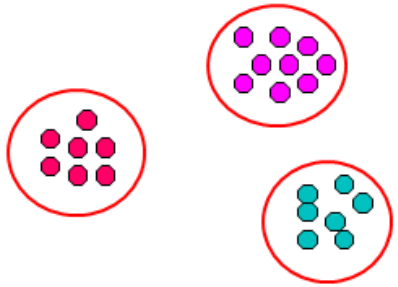
- Jaccard距离

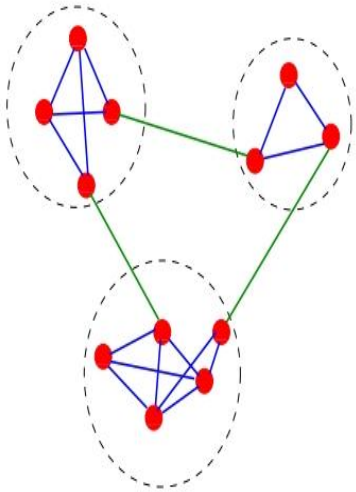
$$J(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

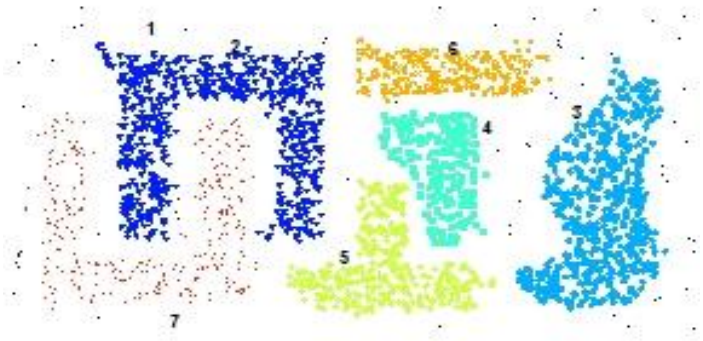
例： x=(1,0,0,1,1) y=(1,1,0,0,1) J(x,y)=2/(3+3-2)=0.5

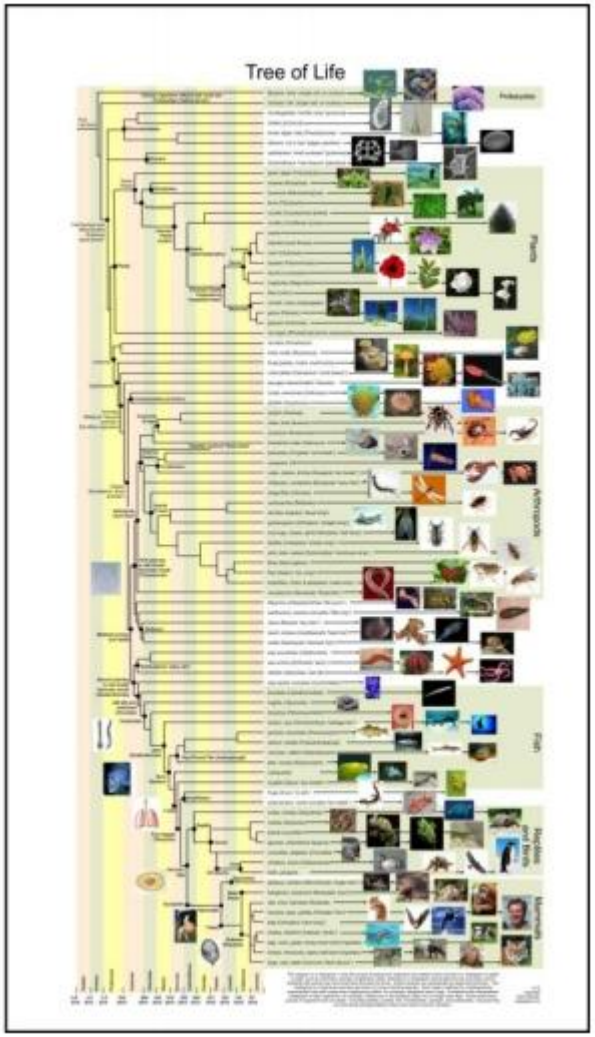
聚类算法

- 基于划分的聚类：K-means、模糊聚类
- 基于层次的聚类：层次聚类
- 基于密度的聚类：DBSCAN
- 基于图的聚类







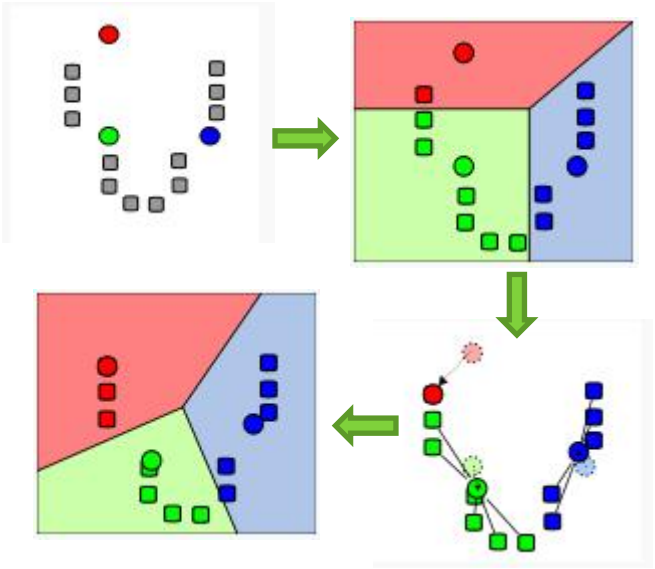


Mahout中的聚类算法

- K-means
- Canopy 聚类
- 模糊K均值（Fuzzy K-means）
- 层次聚类
- EM聚类（期望最大化聚类）
- 谱聚类
- 。。。。。。

K-means算法

- 算法思想
  - 1：随机选择K个点作为初始质心
  - 2：repeat
  - 3：　将每个点指派到最近的质心，形成K个簇
  - 4：　重新计算每个簇的质心
  - 5：until 质心不发生变化



- 算法特征：
  - 有利于发现球形或圆形簇
  - 复杂度低，为O(NKt)，其中N是对象点的个数，t 是迭代次数
  - K值不易确定，且可能产生空簇
  - 对初始质心有一定的依赖性

Canopy 聚类

- 算法思想
  - 1： 设置阈值 T1, T2, 且T1>T2
  - 2： 将相似的对象放在一个子集中，生成多个canopy集
  - 3： 对各个canopy内使用K-means聚类
- 算法特征：
  - 不需要设置 k 值，但是需要设置 T1,T2 两个阈值
  - 可并行化，计算速度快
  - 通常先用 canopy聚类来确定k值，再进行k-means聚类

模糊K均值（Fuzzy K-means）

- 数学背景
  - 模糊集合论、模糊逻辑
  - 隶属度：对象属于某个集合的概率
- 数据点集 X = {X1,X2,...,Xn}, 模糊簇集 C = {C1,C2,...,Ck};
  - 给定点Xi的所有权值之和为1：
  - 每个簇Cj 以非零权值至少包含一个点，但不以权值1包含所有点：
- 算法思想
  - 1： 选择一个初始模糊划分，即对所有点赋初始权值
  - 2： repeat
  - 3： 使用模糊划分，计算每个簇的质心
  - 4： 重新计算模糊划分，即Wij
  - 5： until 质心不发生变化

案例分析—利用电信数据挖掘快递员人群

- 快递员人群特征：

phone	call_in	call_out	in_out	voice	ms	durtime	call_sum	in_sum	voice_sum	time
A	154	445	0.34607	520	79	52461	599	0.257095	0.868114	0.655763
B	350	552	0.63406	709	193	46802	902	0.388027	0.786031	0.585025

- 真实快递员数据分析：

IMIS	DIALED_DIGITS	CALL_FLAG	FINAL_OPTION	ACCESS_TIME	DURING_TIME
A	D	0	3	2013.12.20 10:30:00	45900
A	C	0	3	2013.12.20 10:40:30	36000
A	E	1	3	2013.12.20 10:56:00	57800
B	F	0	6	2013.12.20 10:10:00	20000
……					

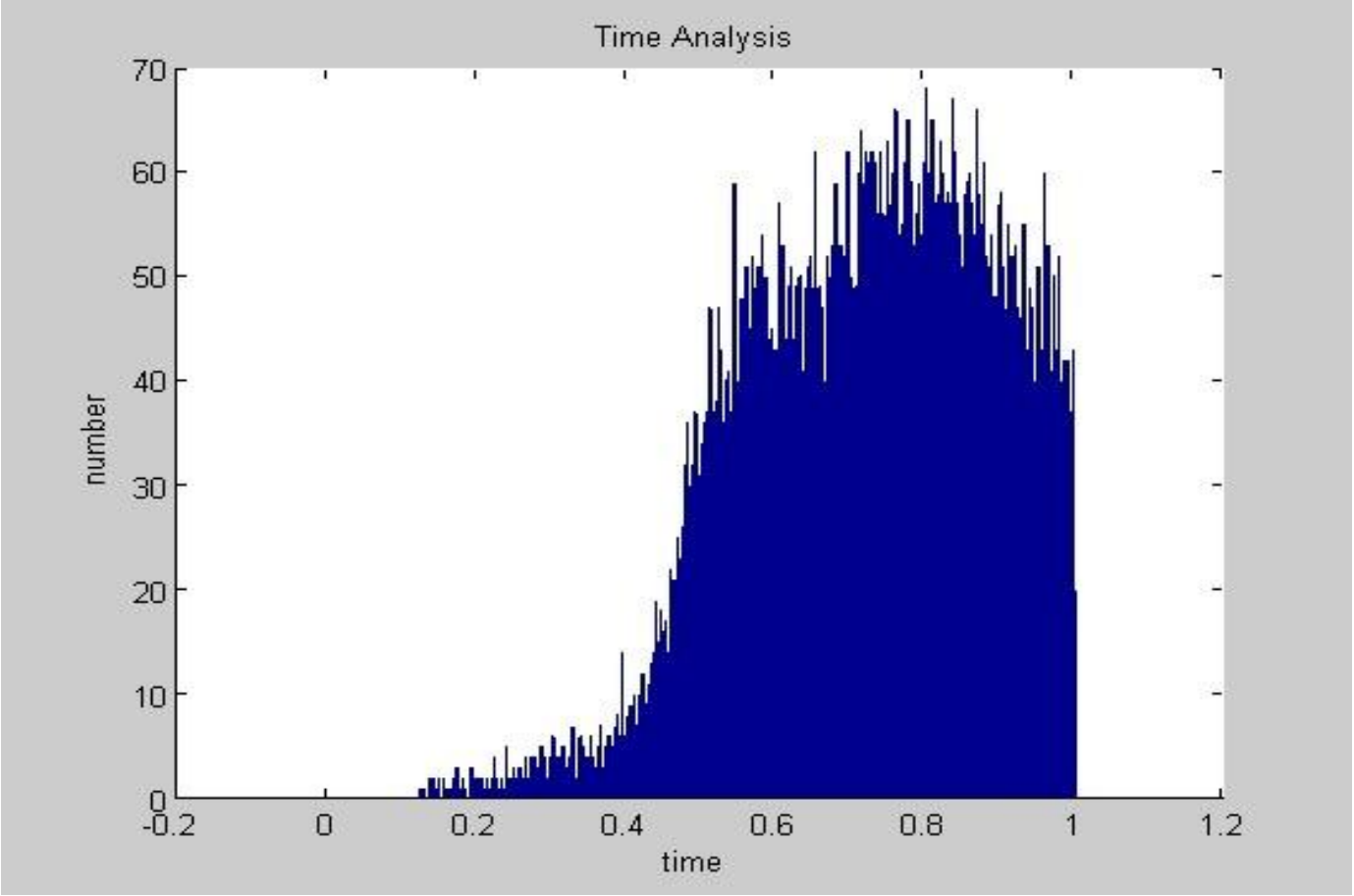
- 呼出次数>400；
- 语音次数>500；
- 平均持续时间（10s，80s）；
- 呼入/呼出 < 0.8

初步筛选

- 数据统计：
  - 每个用户的呼入呼出次数、语音短信次数、平均呼叫持续时长、呼入/呼出
- 多维度筛选：
  - 呼出次数>400，语音次数>500，平均持续时间(10s，80s)，呼入/呼出 < 0.8
- 筛选结果：
  - 400多万 ➔ 2.45万
- 统计分析：
  - 根据平均呼叫持续时长、呼出比重、语音比重分析筛选后的2.45万人

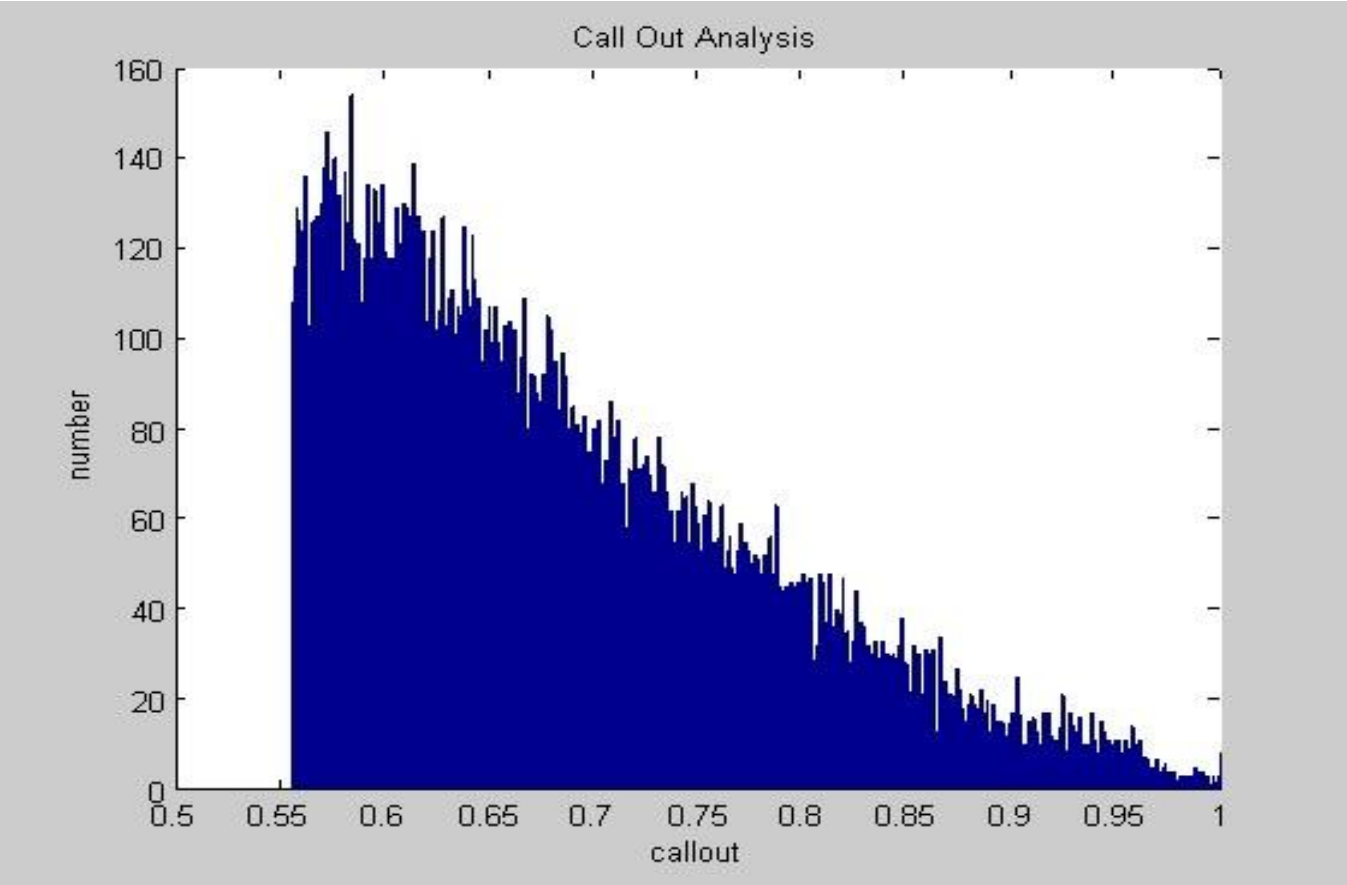
统计分析

- 该类人群的平均呼叫时长集中在40s与80s中间，（0.5, 1）。



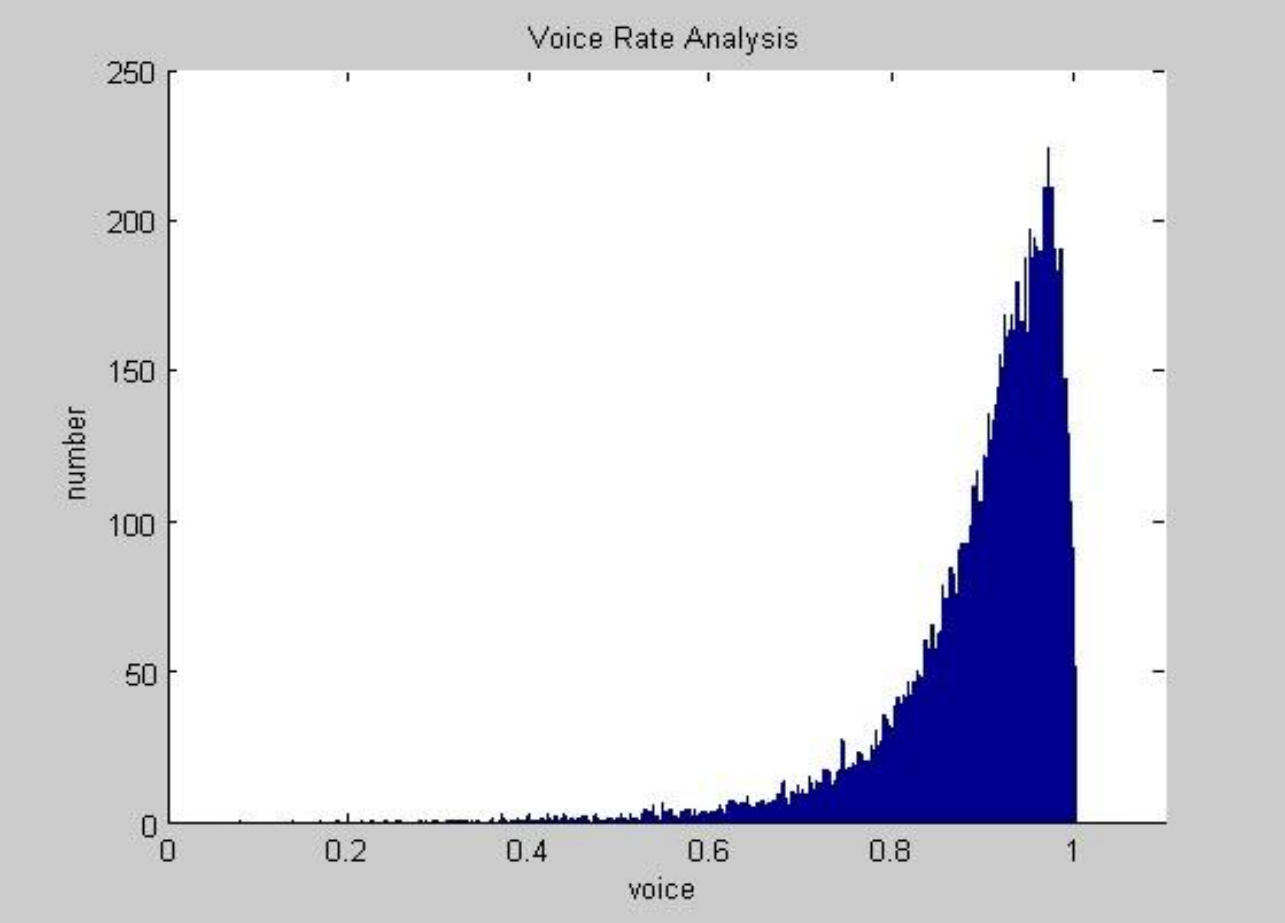
统计分析

- 该类人群的呼出比重（呼出次数/总的呼叫次数）均大于0.55，且随着比例的增大，对应人群数量近似线性递减。



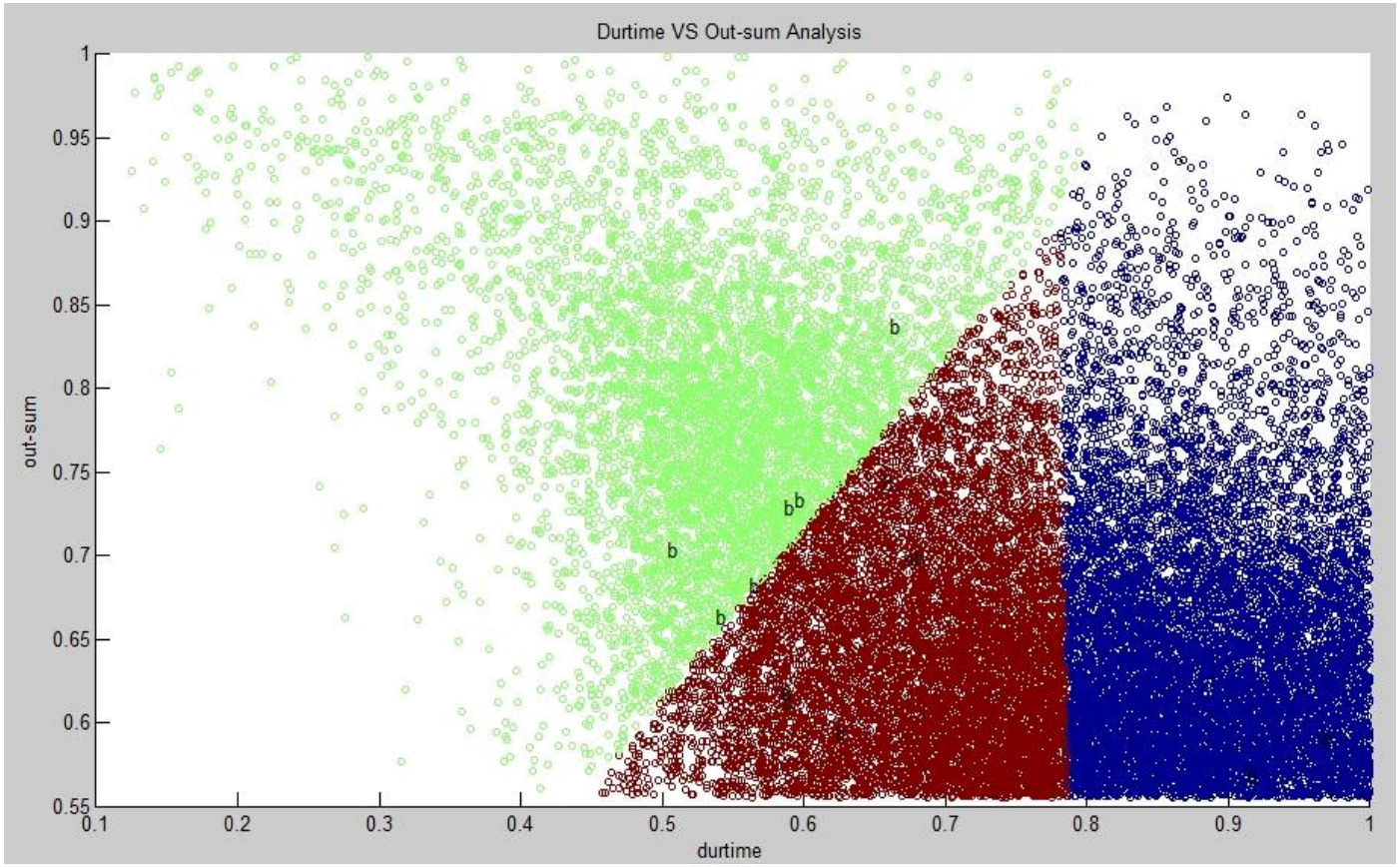
统计分析

- 语音比重（语音次数/总的呼叫次数）集中在0.8到1，相差不大。



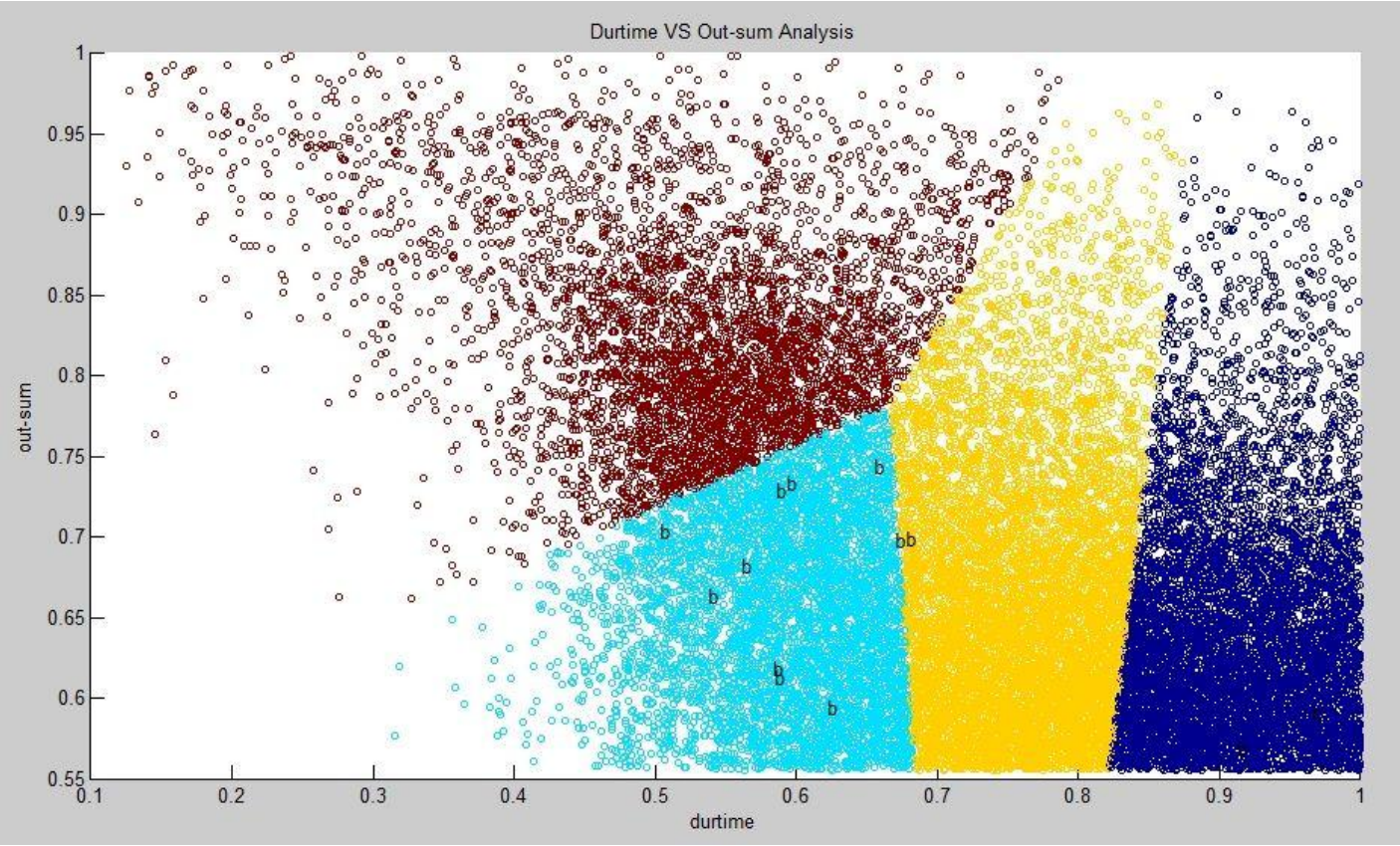
聚类分析

- 根据用户的呼叫持续时间、呼出所占比重对该人群做K-means聚类分析



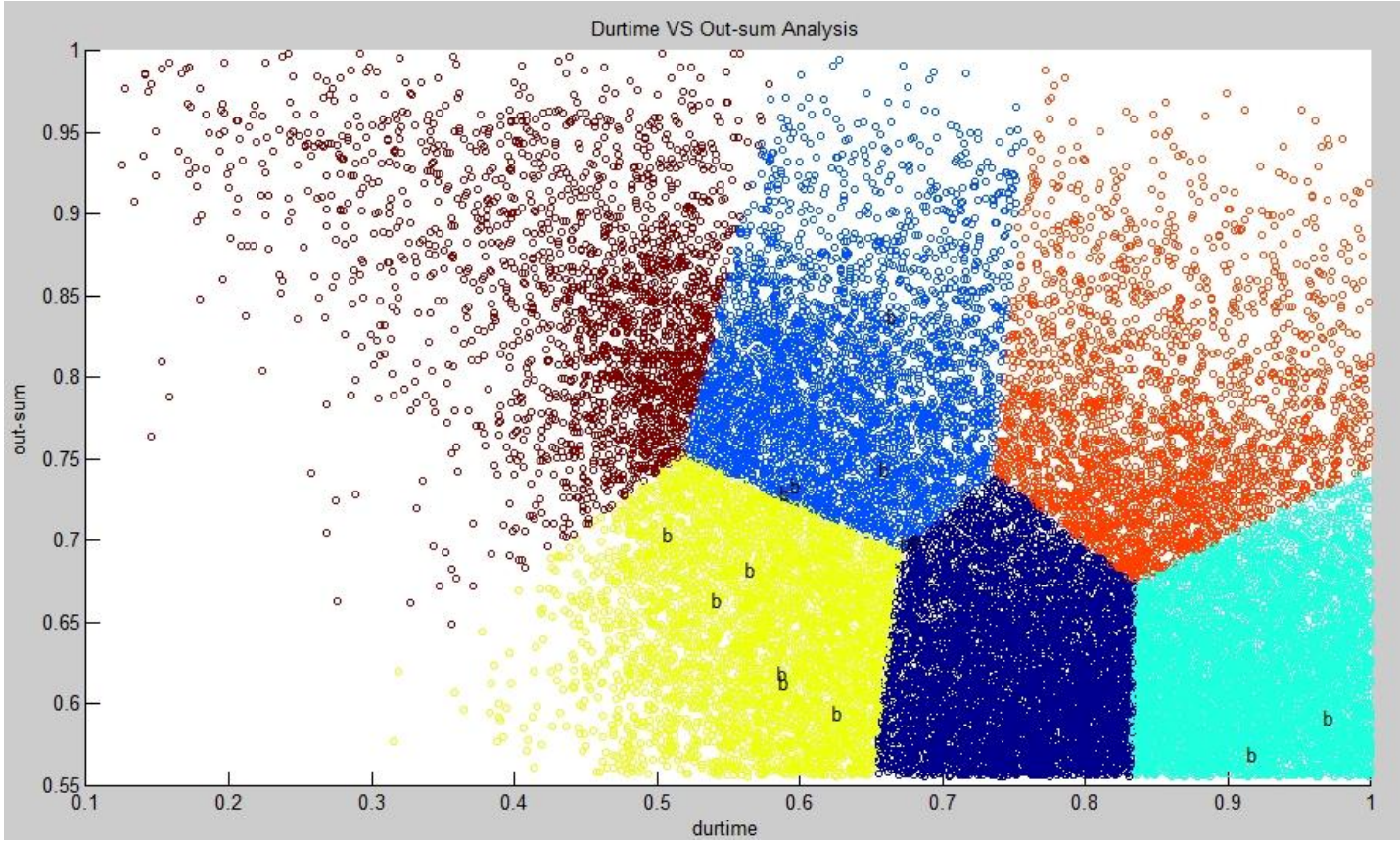
聚类分析

- 根据用户的呼叫持续时间、呼出所占比重对该人群做K-means聚类分析



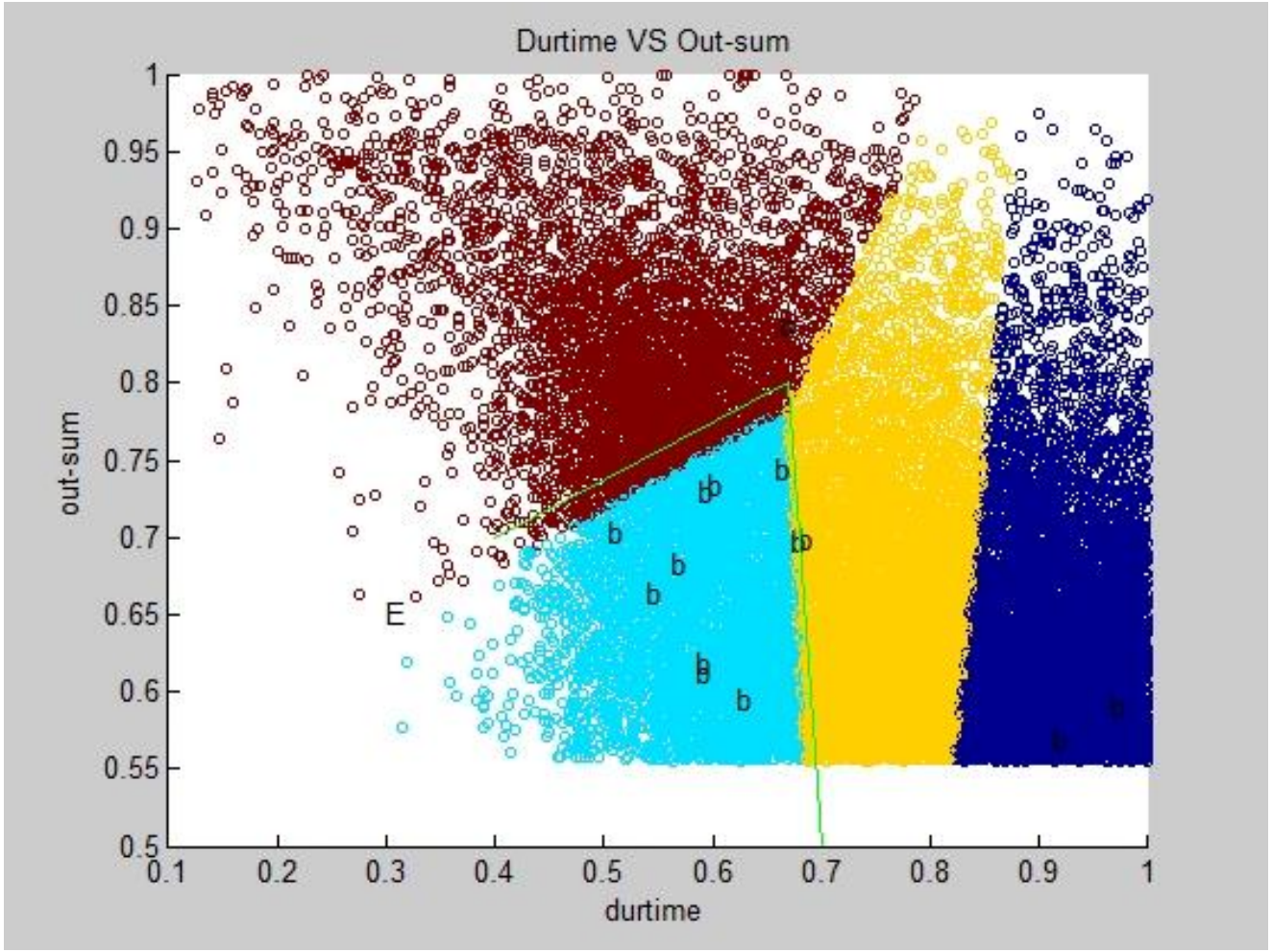
聚类分析

- 根据用户的呼叫持续时间、呼出所占比重对该人群做K-means聚类分析



聚类结果

- K-means 聚成 4 类



代码分析

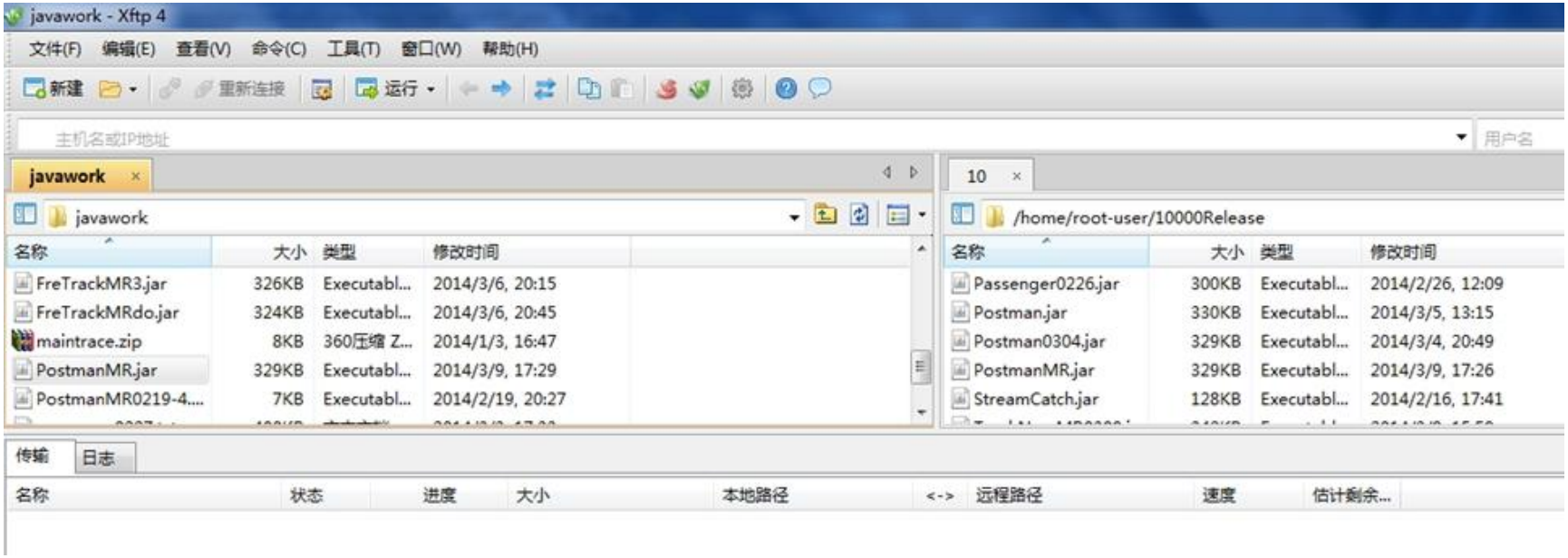
- Map函数
  - 数据预处理：提取用户工作时间（8：00~20：00）的有效数据
- Reduce函数
  - 统计用户呼出次数、语音次数、平均持续时间 、呼入/呼出

```
public static class PostmanMapper extends Mapper<LongWritable, Text, Text, Text>
{
    private final int StartTime = 28800; //开始时间（8: 0: 0）
    private final int EndTime = 72000;  //结束时间（20: 0: 0）
    private int nCurTime = 0;

    private DateFormat dateFormat = new SimpleDateFormat("yyyy-MM-dd HH:mm:ss");
    private Date dateTime; //通话时间
    private String strPhone,strParPhone,strCallDur,nCallFlag;
    private int nFinalOpt,nCallState;

    private Text outKey = new Text();
    private Text outVal = new Text();
```

（视频演示）

代码分析

- 打包上传


javawork - Xftp 4
文件(F) 编辑(E) 查看(V) 命令(C) 工具(T) 窗口(W) 帮助(H)
新建 重新连接 运行
主机名或IP地址
用户名
javawork
javawork
名称 大小 类型 修改时间
FreTrackMR3.jar 326KB Executabl... 2014/3/6, 20:15
FreTrackMRdo.jar 324KB Executabl... 2014/3/6, 20:45
maintrace.zip 8KB 360压缩 Z... 2014/1/3, 16:47
PostmanMR.jar 329KB Executabl... 2014/3/9, 17:29
PostmanMR0219-4.... 7KB Executabl... 2014/2/19, 20:27
10
/home/root-user/10000Release
名称 大小 类型 修改时间
Passenger0226.jar 300KB Executabl... 2014/2/26, 12:09
Postman.jar 330KB Executabl... 2014/3/5, 13:15
Postman0304.jar 329KB Executabl... 2014/3/4, 20:49
PostmanMR.jar 329KB Executabl... 2014/3/9, 17:26
StreamCatch.jar 128KB Executabl... 2014/2/16, 17:41
传输 日志
名称 状态 进度 大小 本地路径 <-> 远程路径 速度 估计剩余...

代码分析

- 运行程序

```
root-user@wellcell10:~$ hadoop jar ./10000Release/PostmanMR.jar
14/03/03 17:01:58 INFO input.FileInputFormat: Total input paths to process : 27
14/03/03 17:01:58 INFO util.NativeCodeLoader: Loaded the native-hadoop library
14/03/03 17:01:58 WARN snappy.LoadSnappy: Snappy native library not loaded
14/03/03 17:01:59 INFO mapred.JobClient: Running job: job_201402280005_0849
14/03/03 17:02:00 INFO mapred.JobClient:  map 0% reduce 0%
14/03/03 17:02:17 INFO mapred.JobClient:  map 1% reduce 0%
14/03/03 17:02:18 INFO mapred.JobClient:  map 2% reduce 0%
14/03/03 17:02:20 INFO mapred.JobClient:  map 3% reduce 0%
14/03/03 17:02:22 INFO mapred.JobClient:  map 4% reduce 0%
14/03/03 17:02:24 INFO mapred.JobClient:  map 5% reduce 0%
```

代码分析

- 最终结果

```
root-user@wellcell10:~$ hadoop fs -ls ./Ret/Postman
Found 22 items
-rw-r--r--   1 root-user supergroup          0 2014-03-05 13:33 /user/root-user/Ret/Postman/_SUCCESS
drwxr-xr-x   - root-user supergroup          0 2014-03-05 13:15 /user/root-user/Ret/Postman/_logs
-rw-r--r--   1 root-user supergroup      67462 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00000
-rw-r--r--   1 root-user supergroup      62239 2014-03-05 13:32 /user/root-user/Ret/Postman/part-r-00001
-rw-r--r--   1 root-user supergroup      64630 2014-03-05 13:32 /user/root-user/Ret/Postman/part-r-00002
-rw-r--r--   1 root-user supergroup      67273 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00003
-rw-r--r--   1 root-user supergroup      65665 2014-03-05 13:32 /user/root-user/Ret/Postman/part-r-00004
-rw-r--r--   1 root-user supergroup      67623 2014-03-05 13:32 /user/root-user/Ret/Postman/part-r-00005
-rw-r--r--   1 root-user supergroup      63348 2014-03-05 13:32 /user/root-user/Ret/Postman/part-r-00006
-rw-r--r--   1 root-user supergroup      67330 2014-03-05 13:32 /user/root-user/Ret/Postman/part-r-00007
-rw-r--r--   1 root-user supergroup      64500 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00008
-rw-r--r--   1 root-user supergroup      66569 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00009
-rw-r--r--   1 root-user supergroup      66684 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00010
-rw-r--r--   1 root-user supergroup      65830 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00011
-rw-r--r--   1 root-user supergroup      64351 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00012
-rw-r--r--   1 root-user supergroup      66874 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00013
-rw-r--r--   1 root-user supergroup      65628 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00014
-rw-r--r--   1 root-user supergroup      66257 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00015
-rw-r--r--   1 root-user supergroup      65842 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00016
-rw-r--r--   1 root-user supergroup      64908 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00017
-rw-r--r--   1 root-user supergroup      64656 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00018
-rw-r--r--   1 root-user supergroup      65465 2014-03-05 13:31 /user/root-user/Ret/Postman/part-r-00019
```

- 合并并保存到本地
  - hadoop fs –getmerge ./Ret/Postman ./Result/postman

概要回顾

- 项目背景
- 电信数据源
- 优质用户
- 分析方法
- 聚类分析
- 案例分析

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间