分布式统计计算

8. 并行的工程应用

李舰

华东师范大学

2019-01-07

目 录

目 录

并行效率的描述

- 加速比
  - 最简单的评价指标是加速比（Speed Up），即串行程序执行时间与并行程序执行时间的比值：
  - $SpeedUp = \frac{T_{serial}}{T_{parallel}}$
  - 当我们增加并行程序数量时，并行程序执行时间 $T_{parallel}$ 将减少，加速比将提高。
- 阿姆达尔定律
  - 阿姆达尔定律（Amdahl's law）引入了处理器的数目 N，重写了加速比的公式：
  - $SpeedUp(N) = \frac{1}{(1-p)+\frac{P}{N}}$
  - 其中算法的可并行比例为 $P(0 <= P <= 1)$。
  - 如果 P 不为 0，加速比的上限是固定的，比如如果 $P = 0.9$，那么加速比最高也只是 10 倍。

并行的选择

- 查普尔定律
  - 查普尔定律（Chapple's law）指出，如果要执行新的并行代码，需要努力实现并行代码足够次数来抵消花费在开发它的时间，这样才能体现价值：
  - $(T_{parallel} \times N) + T_{parallel_algorithm_development} << T_{serial} \times N$
  - 在实际的应用中，还要考虑经济成本。
- 并行方式的选择
  - 选择并行范式，例如基于 R 的 Snow、MPI 等。
  - 确定硬件加速的可能性，比如 GPU 计算、多核计算等。
  - 选择并行框架，比如 Hadoop、Spark 等。
  - 选择算法库，比如 BLAS、机器学习算法库等。
- 预估计算效率
  - 估算算法复杂度。
  - 估算加速比。

目 录

分块均值方法

- 简介
  - 分块均值（Chunk Averaging）简称 CA，从 1999 年以来，不同作者对分块均值的特定形式进行了研究。
  - 假设 $\hat{\theta}$ 通过在我们的数据上使用函数 g() 生成，那么可以将数据按行分为 r 块，对每个分块使用 g()，对 r 个结果取均值，从而得到我们对 θ 的 CA 估计量 $\widetilde{\theta}$。
- 渐进等价
  - CA 方法的可靠性来自于 CA 估计量 $\widetilde{\theta}$ 在统计上和完全估计量 $\hat{\theta}$ 是等价的。
  - 可以证明，如果数据是独立同分布的，而且完全估计量是渐进多元正态分布，那么 CA 方法也会生成一个渐进多元正态的估计量。而且最重要的是，分块估计量和原始估计量有相同的渐进协方差矩阵。
- 并行实现
  - 对于一些复杂的方法，如果数据量巨大，可以无需从算法层面上并行，只需要将数据拆分，然后使用 CA 方法实现“易并行”的处理即可。

BLB 和变量子集

- BLB 方法
  - Bag of Little Bootstraps 方法简称 BLB。在 BLB 中，和 CA 方法一样，人们同样关注数据分块，但分块是随机的。
  - 挑选 s 个大小为 b 的分块，对于每个分块，我们使用标准的 bootstrap，有放回地抽取 r 个大小为 b 的样本。之后对所有分块取均值，和 CA 类似。
  - 可以证明，对于独立同分布数据，BLB 产生的估计量和 $\hat{\theta}$ 是渐进等价的。
- 变量子集
  - 考虑一个回归或者分类问题，可以对预测变量取子集，而不是对观测值取子集，这种形式也被称为 Boosting。
  - 假设有 50 个预测变量，用逻辑斯蒂回归来预测一个二元变量 Y，可以随机地选择 k 对预测变量，对每对变量使用模型，而不是一次调用 50 个变量建模。这样我们可以生成 k 个预测值，然后使用多数原则来预测新的 Y 值。
  - 需要注意，和 CA 以及 BLB 方法不同，Boosting 方法无法产生一个统计上等价的估计量。

目 录

目 录

环境对象

- 什么是环境
  - 所有的数据对象都存在于环境，类似于平行世界，不同环境之间互不干扰。
  - 通常我们只操作一个全局环境 `.GlobalEnv`，在很多和环境操作的相关函数中都是默认值，但是可能可以通过参数 `envir` 来指定特定的环境。
  - 可以任意新建各种环境。
- 环境相关的函数
  - 新建环境：`new.env`。
  - 判断是否为环境对象：`is.environment`。
  - 获取上层环境：`parent.env`。
- 数据的存取
  - 列出环境内的对象：`ls`。
  - 在环境中赋值：`assign`。
  - 删除环境中的对象：`rm`。

命名空间和搜寻路径

- 命名空间
  - 每个包都有独立的命名空间，空间内的对象名互不干扰。
  - R 中的命名空间是特殊的环境。
  - R 3.0 之后强制要求 NAMESPACE 文件。
  - 可以通过正则表达式设置函数是否被外部调用。
- 关于内部函数
  - 约定俗成的做法是将内部函数以“.”开头，统一存放在 R 包里的 util.R 脚本中。
  - 不需要暴露给用户的函数作为内部函数。
  - 内部函数可以不写文档。
- 搜寻路径
  - search 函数返回所有的环境。
  - 其顺序是有意义的，R 会按照顺序依次查找。
  - 位于第一个环境中的变量实际上就是全局变量。

option 选项与全局设置

- 全局设置
  - 在很多操作中，我们希望数据（或者某些设置）不会随着函数的结束而消失，甚至不会因为 R 的重启而消失，需要进行全局的设置。
  - 一种常用的方式是使用“<<-”来赋值全局变量。
- options 相关函数
  - options() 返回所有选项。
  - getOption 可以获取某个选项信息。
  - options 函数可以用来修改某些选项。
- options 的应用
  - 除了修改内置的选项以外，还可以新建选项。
  - options 可以代替全局变量来存放全局的信息，在函数式编程的最佳实践中，不提倡使用全局变量，对于一些需要全局进行设置的变量，在 R 中通常存放在 options 中。
  - 重启 R 之后 options 会被重置，如果希望重启后也能保存，可以在 etc 文件夹的 Rprofile.site 文件中进行修改。

系统信息

- 查看本机信息
  - Sys.info 返回系统信息。
  - Sys.getlocale 可以查看字符环境。
  - .Platform 显示当前系统平台信息。
  - R.home 查看当前 R 的安装路径。
  - R.Version 查看当前 R 版本的详细信息。
  - .libPaths 查看当前包的路径。
- 查看环境变量
  - Sys.getenv 查看环境变量。
  - names(Sys.getenv()) 可以看到所有能获取的环境变量的信息，要多于系统管理界面中的信息。
- 注意事项
  - 一些函数和功能在不同的操作系统中可能会有差异，通常会使用 .Platform 函数来识别操作系统，并进行分别处理。
  - 不同的操作系统甚至管理权限，会导致默认的包目录不同，可以通过 .libPaths 函数进行确认。

异常处理

- 关于异常处理
  - 在工程的应用中，对于程序健壮性的要求是一个非常重要的问题，尤其是在很多自动执行的场景下，针对各种异常中断的可能性都需要提前进行处理。
  - 各种编程语言都包含异常处理的机制，可以针对不同的错误进行识别和处理，不影响整个程序的运行。
- try 函数
  - 可以吞掉错误而不影响正常运行。
  - 如果没有错误，返回正常值，如果报错，返回 try-error 类型的对象。
  - silent = TRUE 可以设置不显示警告。
- tryCatch 函数
  - interrupt，打断后的处理，包裹 ctrl+c 或者 esc 打断。
  - error，错误后的处理。
  - warning 警告后的处理。
  - finally，无论如何也要运行的代码。

程序调试

- 调试工具
  - 在函数体内部手工添加 `browser()` 可以设置断点。
  - 重新对函数赋值，再次运行时在断点处中断。
  - 可以查看当前的对象。
  - 回车表示继续执行。
- browser 生效时的命令
  - 回车表示继续执行
  - `c` 表示退出 `browser` 并执行后面的语句。
  - `f` 表示停止当前循环或函数。
  - `n` 表示单步到下一句命令。
  - `Q` 表示退出调试环境。
- 图形工具
  - Rstudio 0.98 以上的版本支持图形化的调试。

目 录

R 的使用方式

- R Script
  - 能直接执行的代码。
  - 通常保存为后缀为 .R 的文本文件。
  - 可以通过复制粘贴到 R 的控制台执行。
  - 也可以在操作系统的命令行使用 Rscript 来执行整个文件。
- R Function
  - 将要实现的功能封装成函数。
  - 通过参数来输入，通过函数返回值来输出。
  - 使用时调用某一个函数，可以避免复制粘贴大量代码，便于重用。
- R Package
  - 可以将自定义的函数保存成 .R 格式的文本，利用 R 包的开发机制将多个 R 函数做成一个 R 包。
  - 使用时通过 R 包的方式进行安装和加载即可，与 CRAN 上的第三方包的使用方式完全相同。
  - 在工程中应该尽量使用 R 包，便于重用和维护。

为什么需要 R 包？

- 方便调用
  - 无需复制粘贴大量代码。
  - 加载包之后可以直接使用包中的函数。
- 命名空间
  - 同一个包中的函数名不用担心与其他包的重名。
  - 函数的作用域也可以得到控制。
- 多人协作
  - 便于多人协作开发大量的代码。
  - 可以自动测试和自动编译。
- 发布和分享
  - 将自己开发的 R 包发布到 CRAN 或者 Github，就可以与全世界的 R 用户进行分享。

R 包的开发环境

- 基础环境
  - R 环境。
  - 编译文档需要 LaTeX，建议安装 TeXLive。
- 开发工具
  - Windows 下需要 RTools。
  - Mac 和 Linux 需要 GCC 环境。
- 其他依赖环境
  - 有些 R 包需要依赖其他的开发环境，比如 Java、数据库等，根据文档进行安装即可。
  - 有些 R 包只能安装在特定的操作系统，使用时也需要注意。
- 操作系统及 R 包的输出结果
  - Linux 下生成 `.tar.gz` 的源码包。
  - Mac 下生成 `.tgz` 的二进制包。
  - Windows 下生成 `.zip` 二进制包。

R 包结构简介

- 必需的目录
  - R 文件夹，包含函数的 .R 脚本。
  - man 文件夹，.rd 格式的帮助文档，近于 LaTeX 格式。
  - DESCRIPTION 文件，固定格式的介绍文档。
  - NAMESPACE 文件，定义导出函数的文件。
- 可选的目录
  - data，Lazy Load 的数据集。
  - inst，任何保留原格式的文件或文件夹。
  - src，C 或者 Fortran 的源码。
  - demo，Demo 代码，可以通过 demo 函数来调用。
  - inst/doc，Sweave 或者 knitr 格式的文档，自动编译成 Vignettes。
  - inst/java，Jar 包。

编译 R 包

- 直接使用命令行
  - `R CMD build`。
  - `R CMD INSTALL --build`。
- devtools
  - `build` 函数。
  - 包含很多很方便的工具。
- RStudio
  - RStudio 提供了编译 R 包的命令。
- 使用脚本
  - Windows 下可将编译过程写成 `.bat` 文件，然后自动执行（通过计划任务）。
  - 推荐使用跨平台的 ANT。

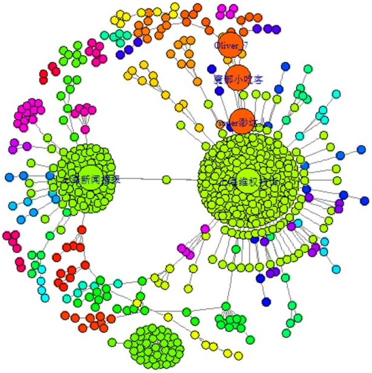
目 录

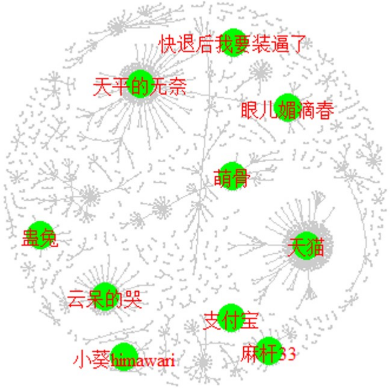
目 录

电商评论分析


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
20
2
3
18
6
1
13
8
7
10
11
17
19
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (11.2% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
超市
天猫
东西
方便
不用
活动
便宜
购物
真的
家里
价格
优惠
现在
实惠
经常
翻牌
划算
喜欢
累计评价 61930
与描述相符
4.9
大家都写到
物流快(810)
质量很好(810)
便宜(750)
服务好(186)
尺码正(61)
不贵(54)
质量一般(69)
全部
追评 (607)
图片 (1241)
有内容
按默认
又好又快，昨晚拍，今天就到了。本来只想买够88，后来禁受不住优惠券的诱惑又囤了一些。
03.12
颜色分类：绿色
尺码：M
萧***4 (匿名)
超级会员
Chuang et. al (2012)
y (2014)
货品收到，包装严实，在天猫超市购物有瘾，物流很快，货品和在当地超市购买的一样，棒棒哒。
颜色分类：绿色
东***h (匿名)

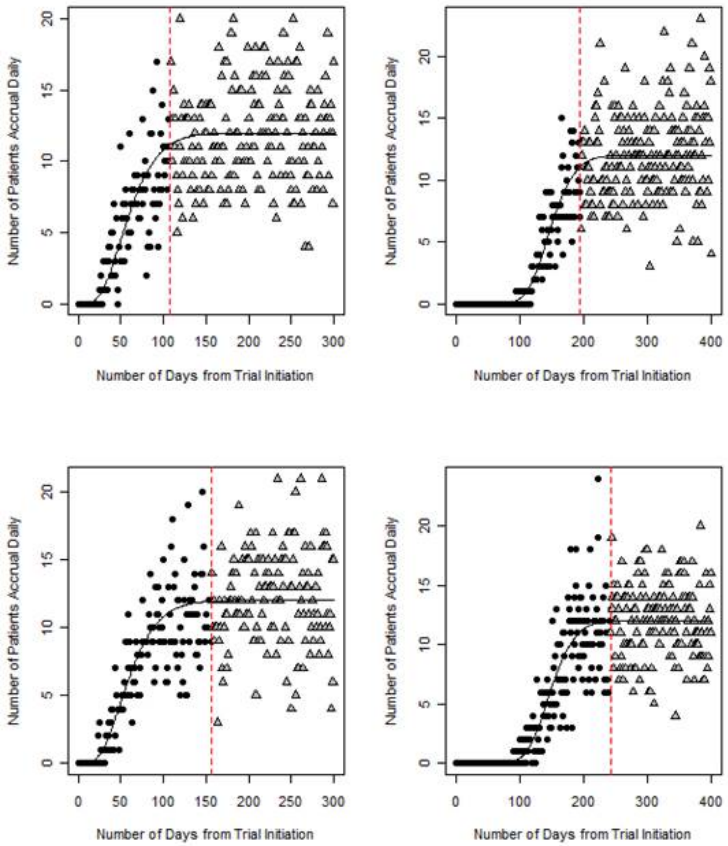
舆情追踪

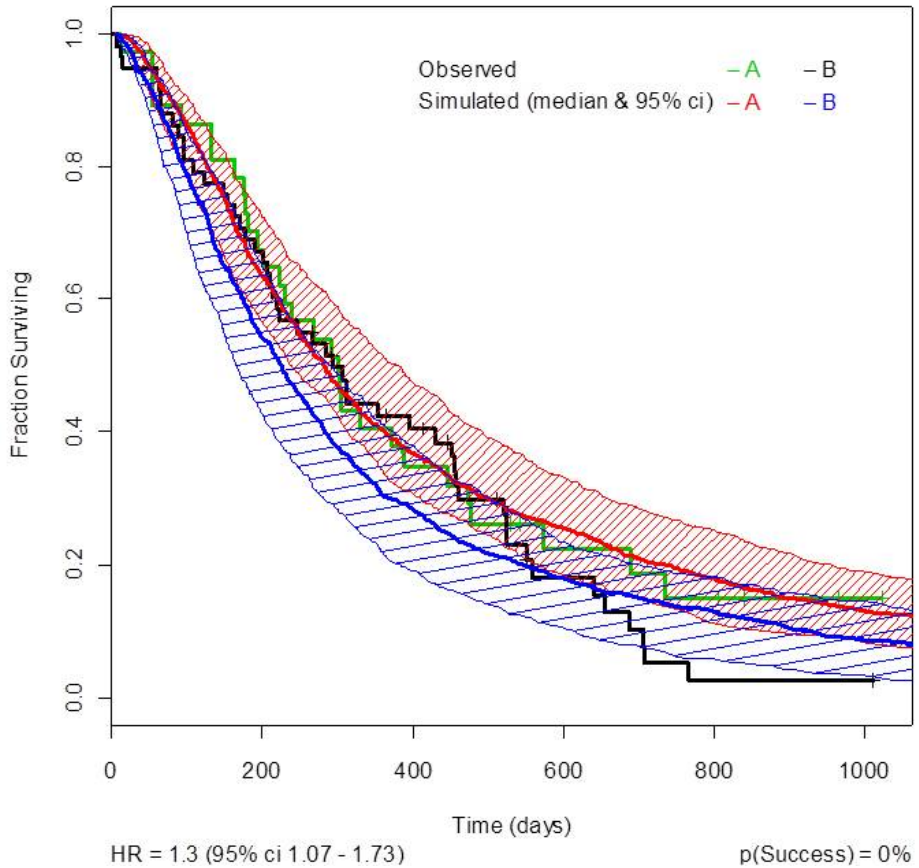




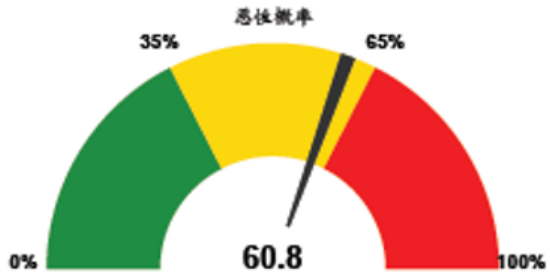
目 录

临床试验和建模模拟

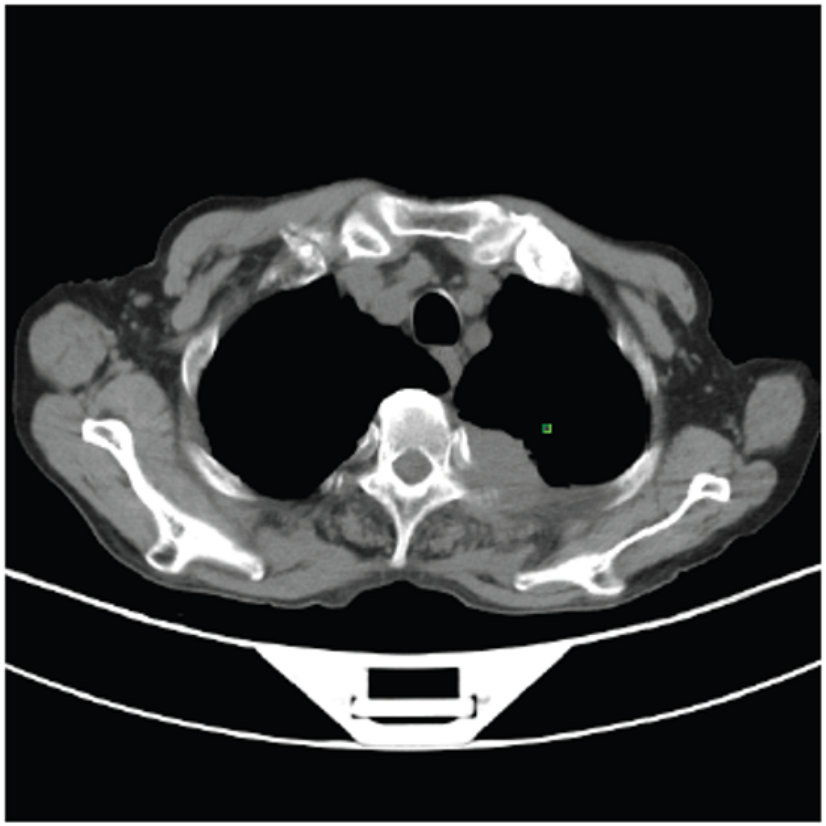

Number of Patients Accrual Daily
Number of Days from Trial Initiation


Observed – A – B
Simulated (median & 95% ci) – A – B
Fraction Surviving
Time (days)
HR = 1.3 (95% ci 1.07 - 1.73)
p(Success) = 0%

人工智能自动诊断

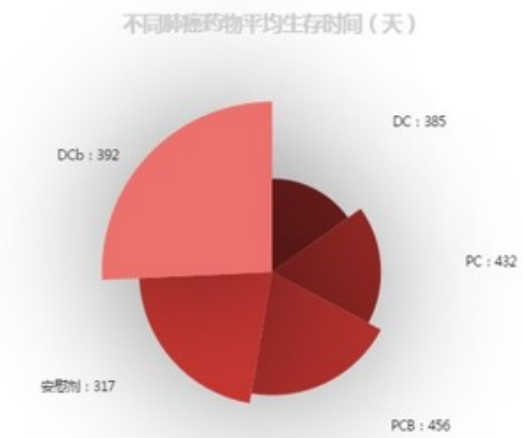


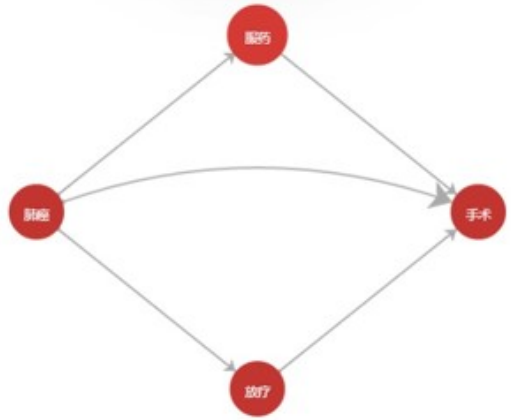
图像信息	
面积	5.89 mm^2
最长直径	1.49 mm
最短直径	1.35 mm
平均直径	1.40 mm
像素个数	14
直径标准差	0.07



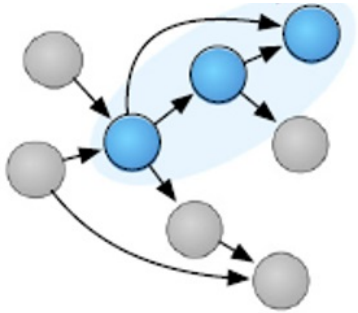
医疗大数据分析

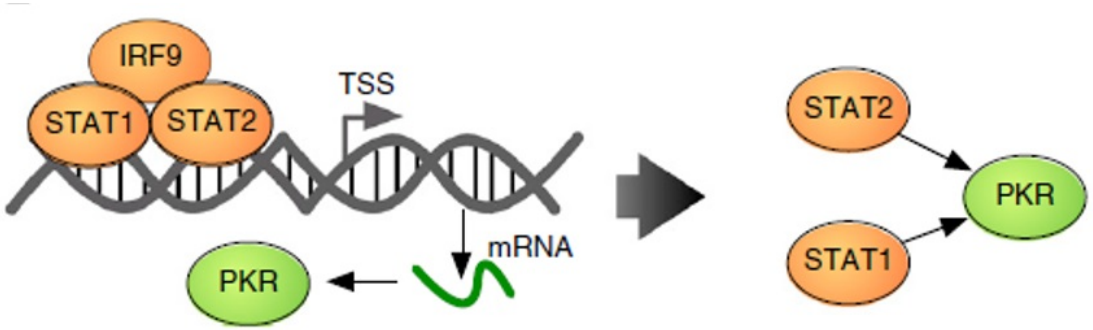

转诊流向
北京
西安
成都
重庆
武汉
南昌
上海
广州


不同肺癌药物平均生存时间（天）
DC：385
DCb：392
PC：432
安慰剂：317
PCB：456


服药
手术
放疗

文献研究和知识图谱





Treatment	M_BASE (cm)	M_SR (1/week)	M_PR (cm/week)	ω_{BASE}
PCB	9.1 (0.33)	0.06 (0.004)	0.13 (0.02)	59% (0.04)
PC	8 (03)	0.038 (0.01)	0.14 (0.04)	63% (0.06)
DC	8.7 (0.31)	0.052 (0.01)	0.16 (0.02)	57% (0.07)
DCb	9.2 (0.38)	0.047 (0.005)	0.16 (0.02)	64% (0.09)
VC	8.5 (0.28)	0.063 (0.01)	0.17 (0.02)	50% (0.04)
DT	8.5 (0.82)	0.033 (0.01)	0.13 (0.02)	77% (0.23)
PT	7.4 (0.47)	0.023 (0.01)	0.25 (0.05)	70% (0.11)
PB[a]	8.6 (0.44)	0.0047 slow (0.001) 0.13 fast (0.004)	0.20 (0.02)	66% (0.08)
ET[a]	8.4 (0.32)	0.0045 slow (0.001) 0.11 fast (0.05)	0.058 (0.02)	67% (0.09)

目 录

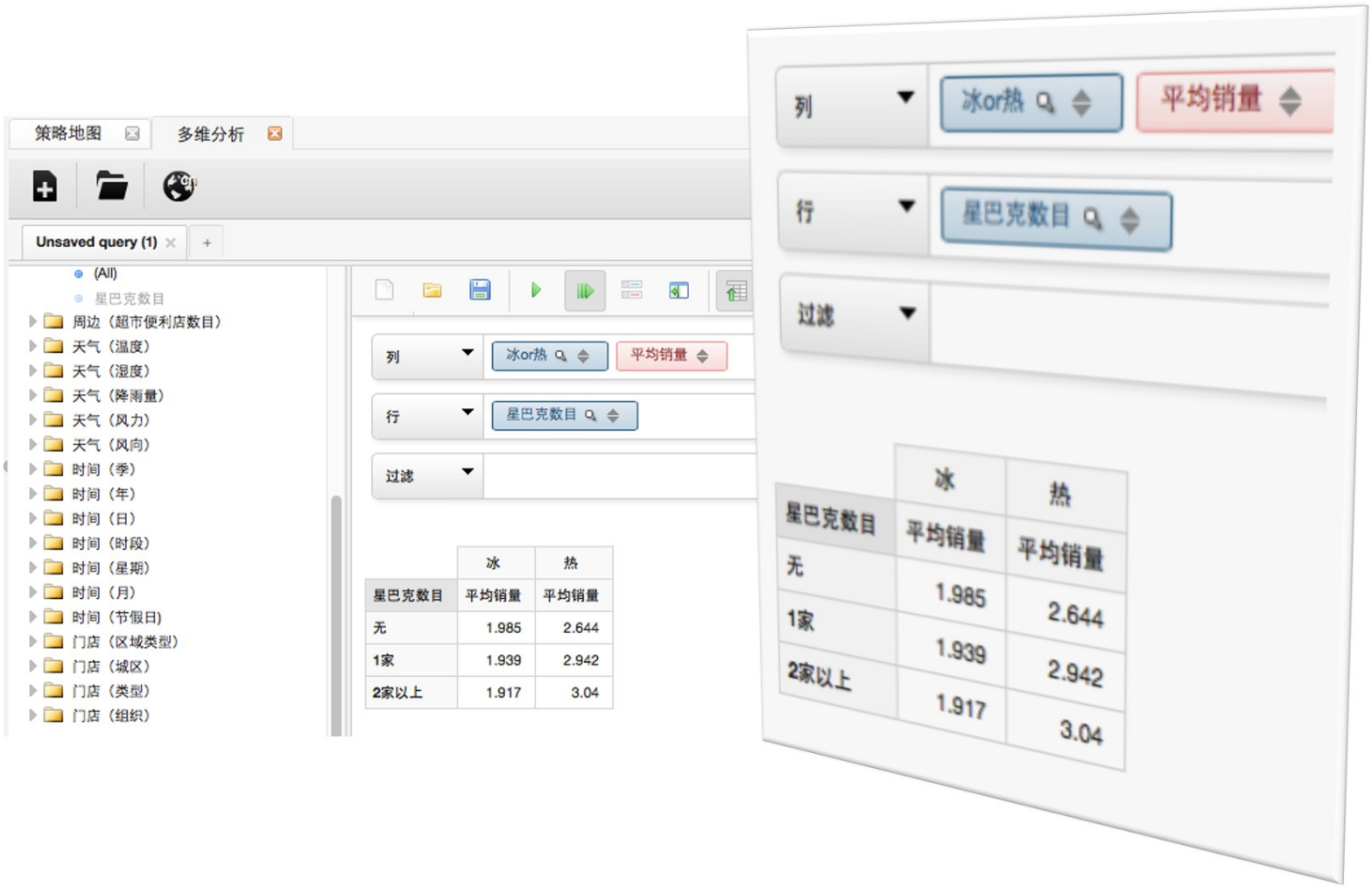
动态地图与新店选址


中环路
大柏树
大华
S126
上海西站
甘泉路
宜川路
虹口区
杨浦
陆家嘴
外滩
上海博物馆
静安区
长宁区
长宁路
华山医院高架直升机场
董家渡
南浦大桥
上海交通大学徐汇校区
古北
徐汇区
Navi & 道道通
2
2
5


淮海公园
大会址
瑞金一路
绍兴公
南京西路
车站路


中山医院
汇联商厦
斜土路
汇区政府
梅林路
东安路
龙华

销售情况的 BI 分析

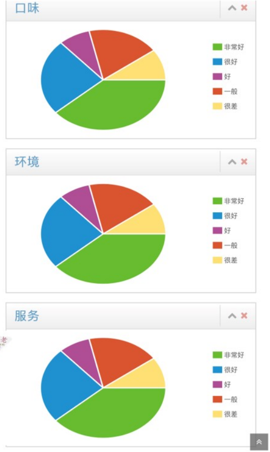

策略地图
多维分析
Unsaved query (1)
(All)
星巴克数目
周边（超市便利店数目）
天气（温度）
天气（湿度）
天气（降雨量）
天气（风力）
天气（风向）
时间（季）
时间（年）
时间（日）
时间（时段）
时间（星期）
时间（月）
时间（节假日）
门店（区域类型）
门店（城区）
门店（类型）
门店（组织）
列
冰or热
平均销量
行
星巴克数目
过滤
冰
热
星巴克数目
平均销量
平均销量
无
1.985
2.644
1家
1.939
2.942
2家以上
1.917
3.04

餐饮用户画像


» 投诉最高菜品
查询条件
菜品词云
全家福
红豆糕
云吞面
全家福馄饨
蟹黄馄饨
荠菜虾仁馄饨
投诉最高菜品

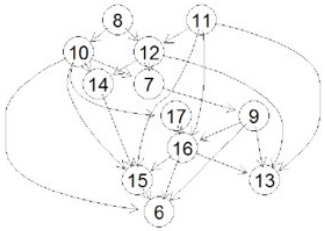

大歌星
九毛九
keywear
影城
LOVELOVE
仙踪林
老西安
芙蓉楼
大玩家
泰道
HMOPTICAL


量不足
太老
腥味很大
偏淡
量不大
咸了
没有了


口味
环境
服务
非常好
很好
好
一般
很差

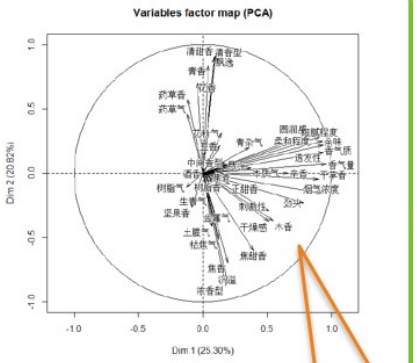
目 录

食品研发



2-吡咯甲醛
4-甲基喹啉
3-苯基吡啶
2, 3-联吡啶

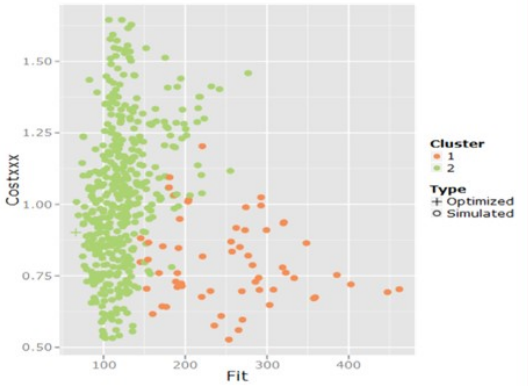
理化属性分析



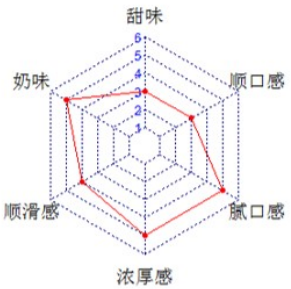
木味
金属味
涩味
苦味

感官分析

FGD 分析



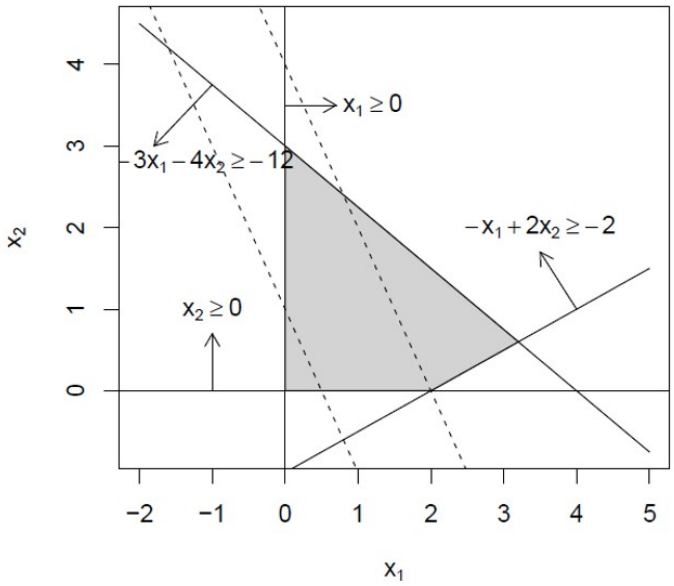
生产最优化

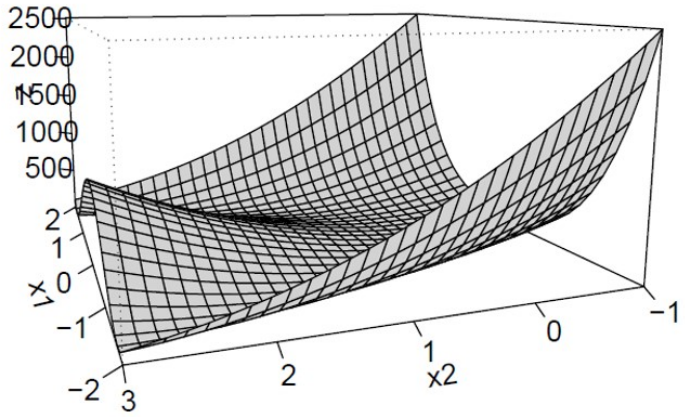


口味测试

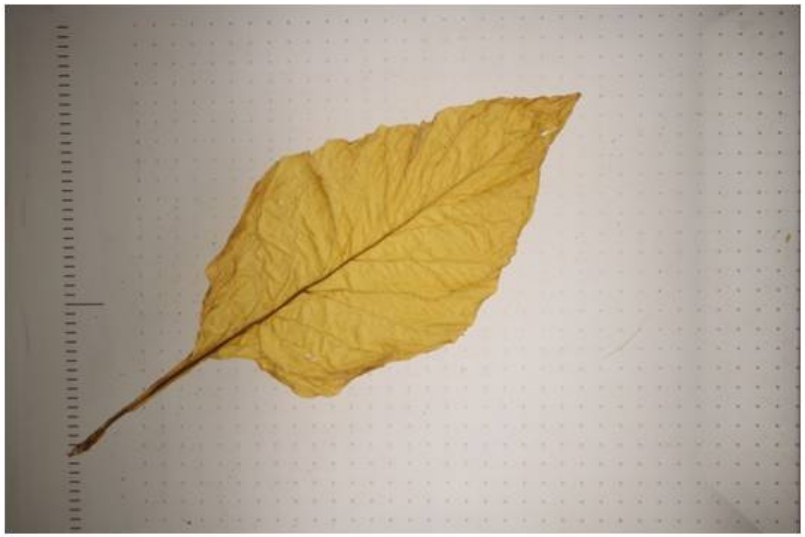
51

最优配比与生产制造


$x_1 \geq 0$
$3x_1 - 4x_2 \geq -12$
$-x_1 + 2x_2 \geq -2$
$x_2 \geq 0$
x_1
x_2

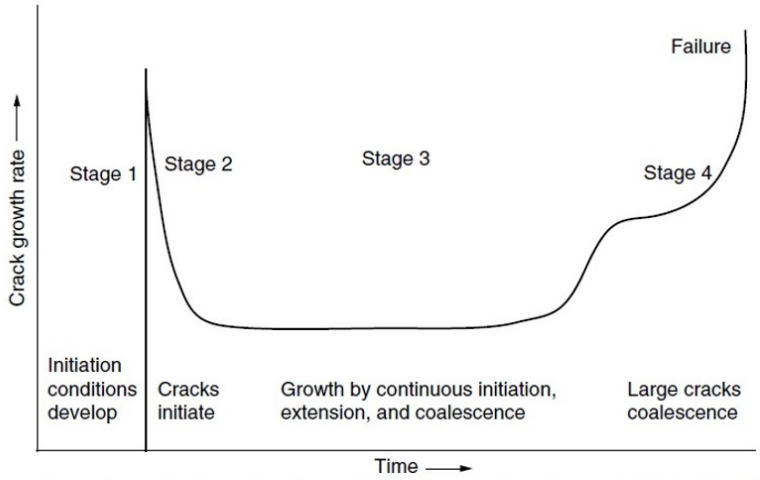

2500
2000
1500
1000
500
x1
x2

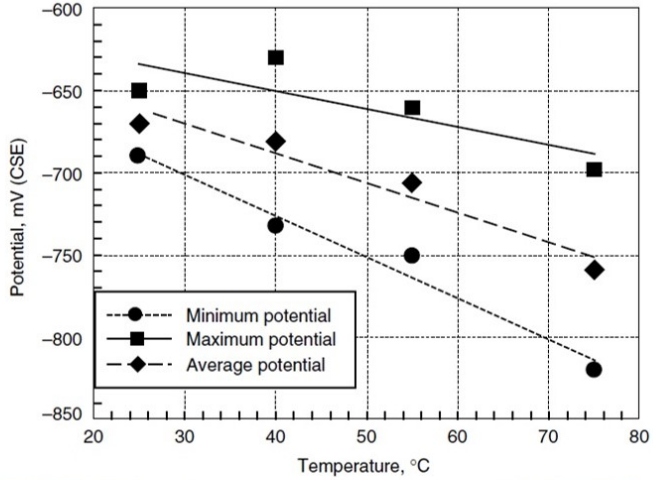
烟叶图像识别



	ID	总面积(cm2)	梗面积(cm2)	最大长度(cm)	R值	G值	B值	描述	区域	等级	年份	报告时间
1	IMG_0008.jpg	711.38	3.10	58.15	0.76	0.57	0.15	IMG_0008				2013-01-17 11:58:42
2	IMG_0009.jpg	719.20	0.02	48.03	0.77	0.60	0.24	IMG_0009				2013-01-17 11:58:54
3	IMG_0010.jpg	532.30	9.44	60.84	0.74	0.51	0.06	IMG_0010				2013-01-17 11:59:28

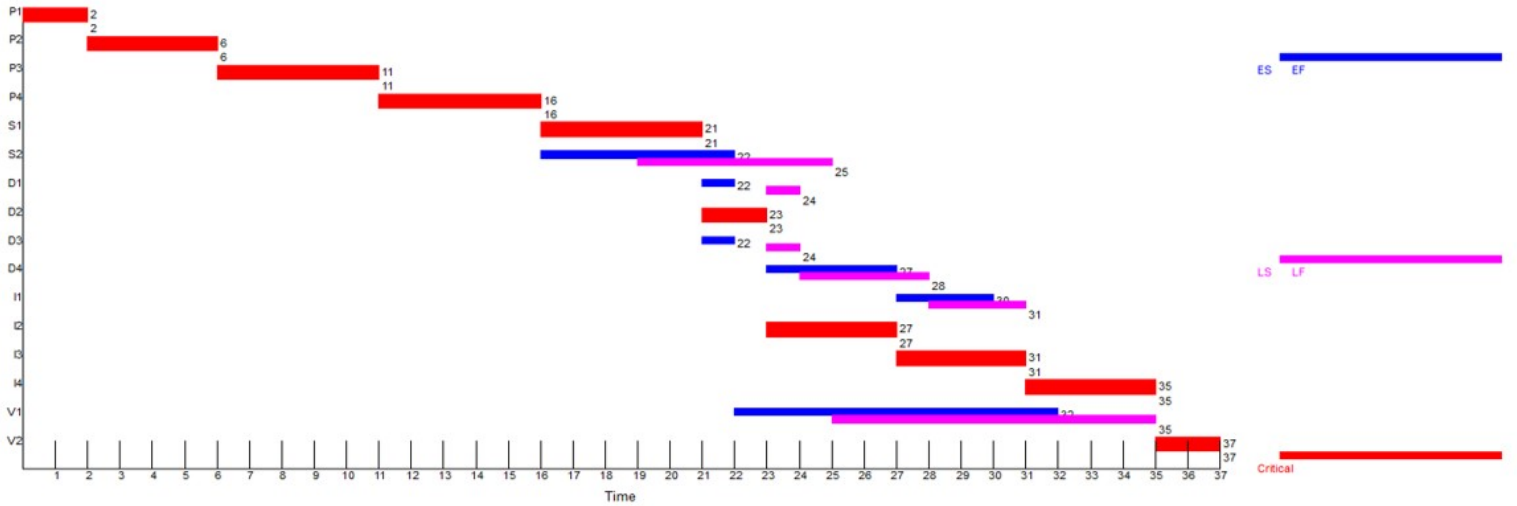
管道腐蚀的预测


Crack growth rate
Failure
Stage 1
Stage 2
Stage 3
Stage 4
Initiation conditions develop
Cracks initiate
Growth by continuous initiation, extension, and coalescence
Large cracks coalescence
Time


Potential, mV (CSE)
−600
−650
−700
−750
−800
−850
20
30
40
50
60
70
80
Temperature, °C
Minimum potential
Maximum potential
Average potential

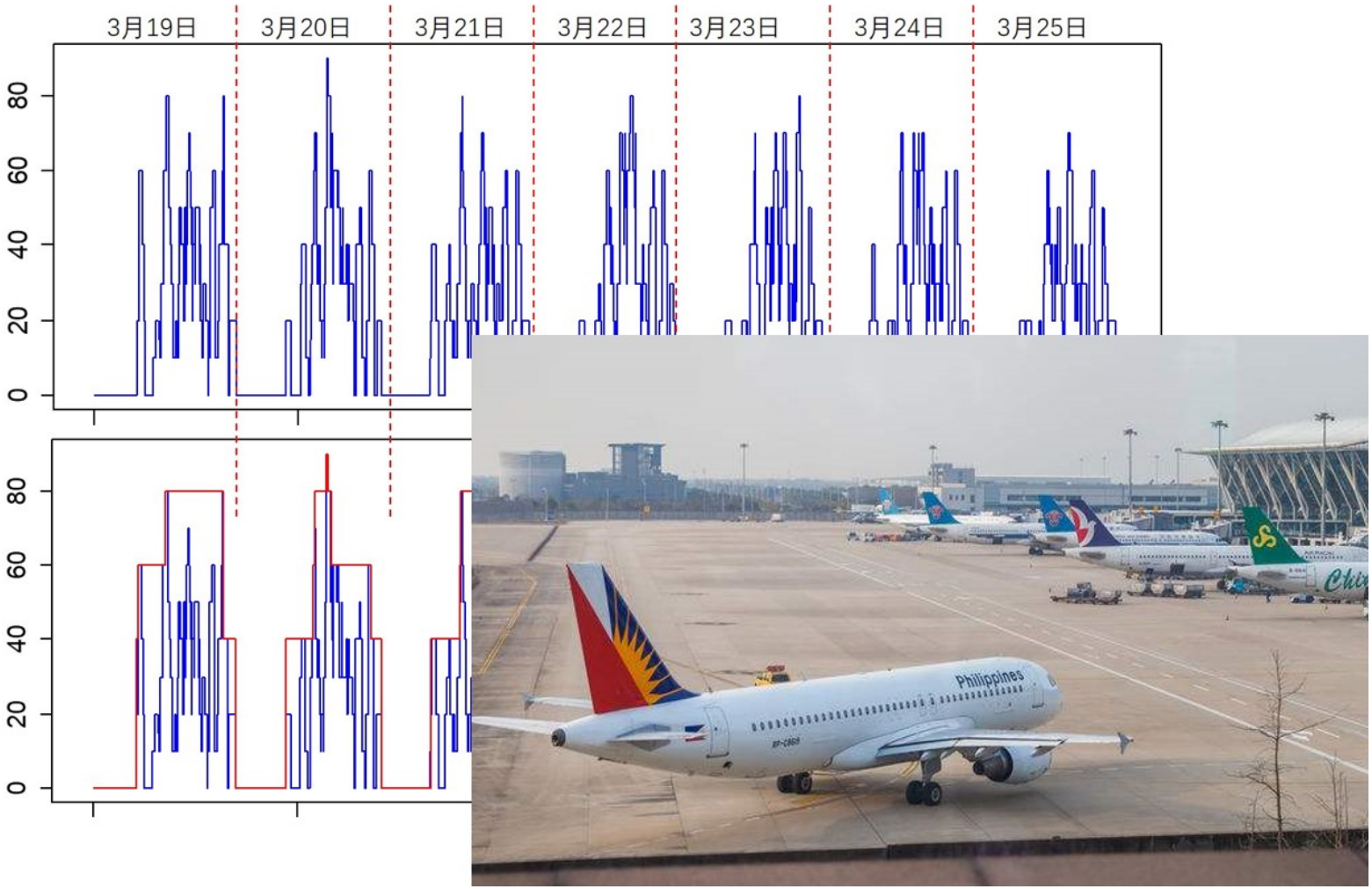
生产排程的优化

Activity	Activity time	Predecessor 1	Predecessor 2	Predecessor 3	Predecessor 4	Predecessor 5	Predecessor 6	Predecessor 7
P1	2							
P2	4	P1						
P3	5	P2						
P4	5	P3						
S1	5	P4						
S2	6	P4						
D1	1	S1						
D2	2	S1						
D3	1	S1						
D4	4	D1	D2	D3				
I1	3	D4						
I2	4	D2						
I3	4	I2						
I4	4	I1	I3					
V1	10	S1	S2					
V2	2	I4	V1					

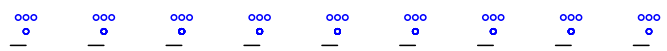


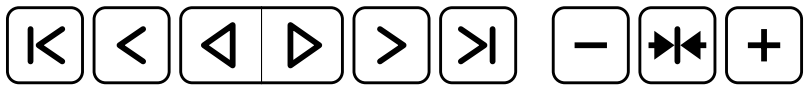
目 录

机场排班优化


3月19日
3月20日
3月21日
3月22日
3月23日
3月24日
3月25日
0
20
40
60
80
Philippines

车站的规划设计

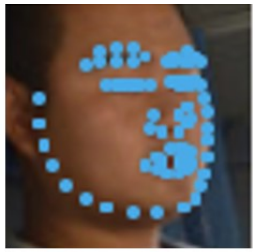


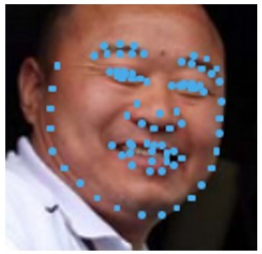


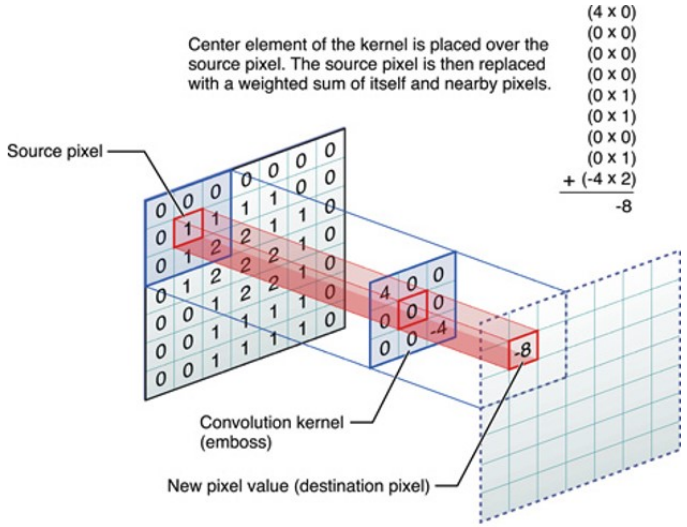
运输风险监控

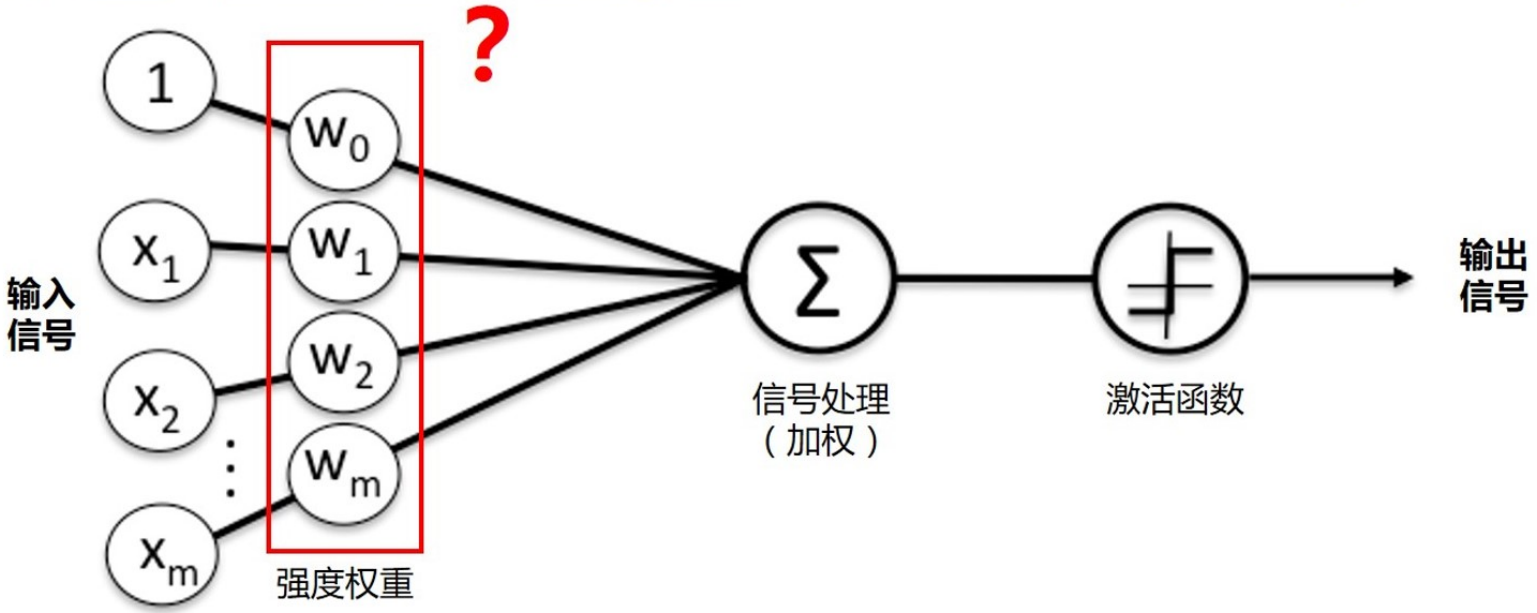

今日安全总体得分
97
昨日运输情况占比图
30%
在线
70%
离线
运行平均里程
最近30天
本月
3,100
2,800
2,500
2,300
2,100
09-01 09-05 09-10 09-15 09-20 09-25 09-30
运行平均时长
最近30天
本月
3,100
2,800
2,500
2,300
2,100
09-01 09-05 09-10 09-15 09-20 09-25 09-30
运行走势
最近30天
本月
Km
h
250
200
150
50
0
250
200
150
100
0
09-01 09-05 09-10 09-15 09-20 09-25 09-30

疲劳驾驶检测和人工智能









车险的定损和理赔


货物
发现
被保险人
受损
破损
造成
玻璃
损坏
损失
北京
导致
运输
爆裂
财产
漏水
途中
发生
外包装
有限公司
水管
时间
破碎
维修
原因
地点
不慎
现场
地板
短量
保险
人员
太保
严重
部分
业主
破裂
服务
需承担
车辆
丢失
顾客
我司
进行
员工
内件
医院
分公司
地下
目的地
公司
出现
设备
手机
卫生间
估计
变形
过程
致使
客户
电梯
一件
金额
物业
浸泡
出险
挤压
查勘
到达
包装
左右
上述
家中
雨水
砸伤
大厦
碰撞
天花板
延保
份额
房间
共计
商品
报案
机场
情况
墙面
材料
航班
一台
期间
暴雨
事故
及时
地面

Thank you!

李舰 Email: jian.li@188.com