



# 大数据技术与城市计算

相关性、机器学习概述和公益课题“宝贝在哪儿”

---

姚尧 博士，副教授，高级工程师

地理与信息工程学院，地图制图学与地理信息工程

阿里巴巴集团，达摩院，访问学者

Email: [yaoy@cug.edu.cn](mailto:yaoy@cug.edu.cn)

办公地点：未来城校区地信楼522办公室





# 主要内容



- 1 机器学习发展与应用
- 2 相关分析和显著性
- 3 机器学习的基本任务
- 4 “宝贝在哪儿” 关联挖掘
- 5 公益课题 “宝贝在哪儿”
- 6 疫情风险分析



- 1.1 人工智能、机器学习和数据挖掘
- 1.2 机器学习的发展历程
- 1.3 机器学习的应用现状

## 人工智能(Artificial Intelligence):

人工制造出来的系统所表现的智能，所谓的智能，即指可以观察周围环境并据此做出行动以达到目的。

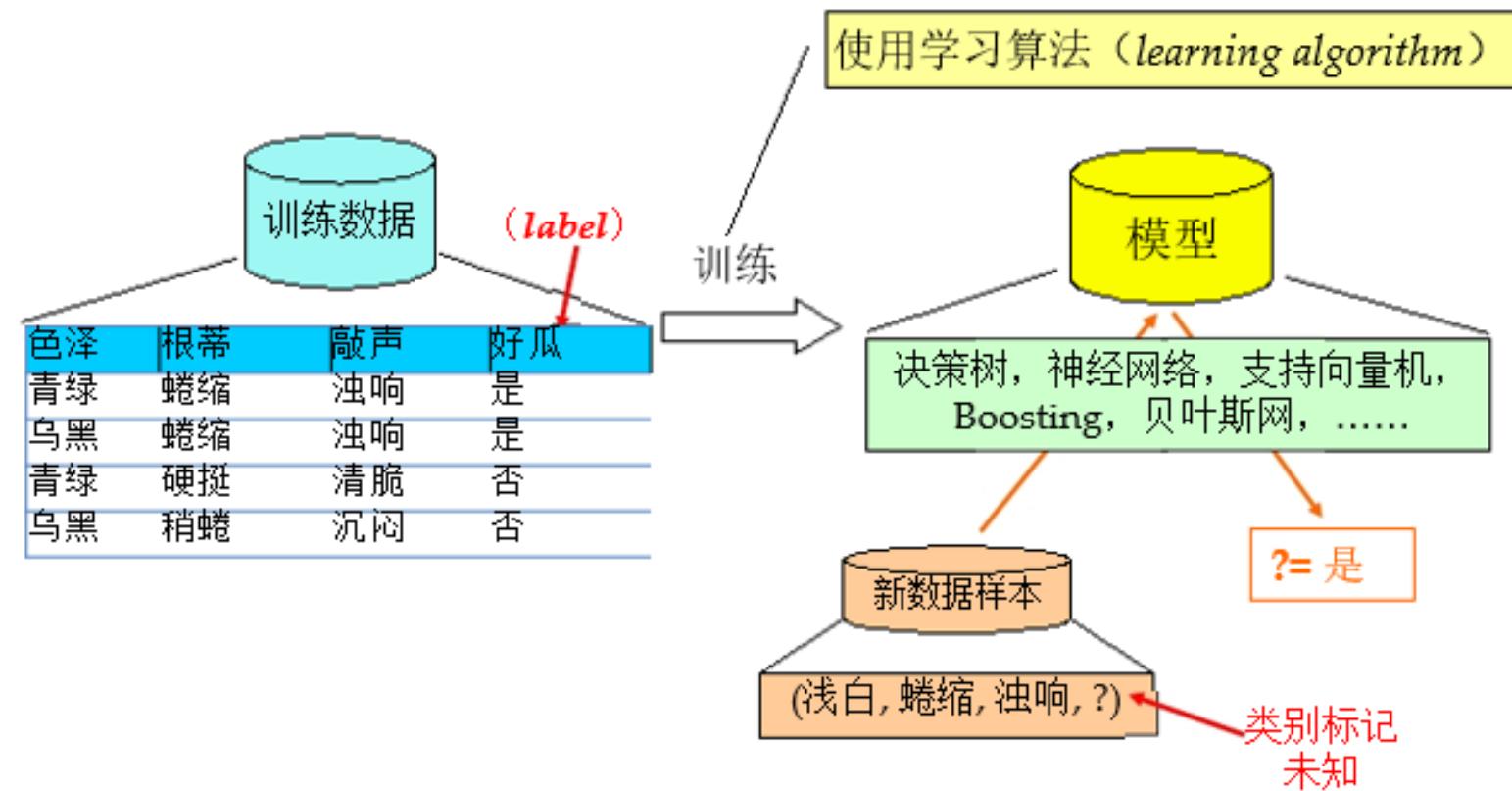
- 弱人工智能 (ANI)
- 强人工智能 (AGI)
- 超人工智能 (ASI)



约翰 麦卡锡  
(1927-2011)  
“人工智能之父”  
1971年图灵奖

## 机器学习(Machine Learning):

机器学习通过设计算法，让计算机能够自动地从数据中“学习”规律，并利用规律对未知数据进行预测。



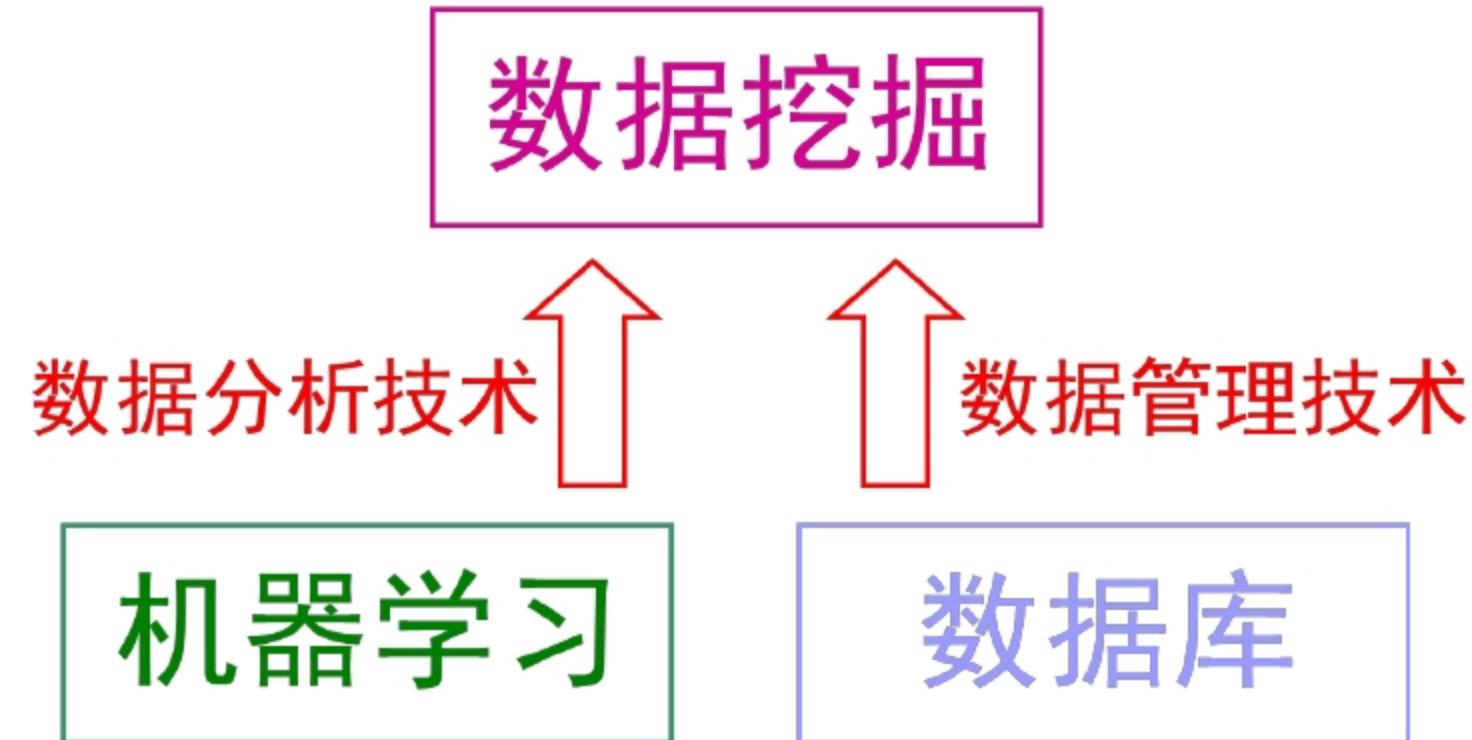
## 数据挖掘(Knowledge Discovery in Database):

数据的处理过程，就是从输入数据开始，对数据进行预处理，包括特征选择、规范化、降低维数、数据提升等，然后进行数据的分析和挖掘，再经过处理，例如模式识别、可视化等，最后形成可用信息的全过程。

# 1.1 | 人工智能、机器学习和数据挖掘



收集、传输、  
存储大数据的目  
的是为了“利用”  
大数据；没有机  
器学习技术分析  
大数据，“利用”  
无从谈起。



## 第一阶段：推理期 1956-1960s: Logic Reasoning

□ 出发点：“数学家真聪明！”

□ 主要成就：自动定理证明系统（例如，西蒙与纽厄尔的“Logic Theorist”系统）



赫伯特 西蒙  
(1916-2001)  
1975年图灵奖



阿伦 纽厄尔  
(1927-1992)  
1975年图灵奖

## 第二阶段：知识期 1970s -1980s: Knowledge Engineering

- 出发点：“知识就是力量！”
- 主要成就：专家系统（例如，费根鲍姆等人的“DENDRAL”系统）



赫伯特 西蒙  
(1936- )  
1994年图灵奖

## 第三阶段：学习期 1990s -now: Machine Learning

□出发点：“系统自己学习！”

□主要成就：.....

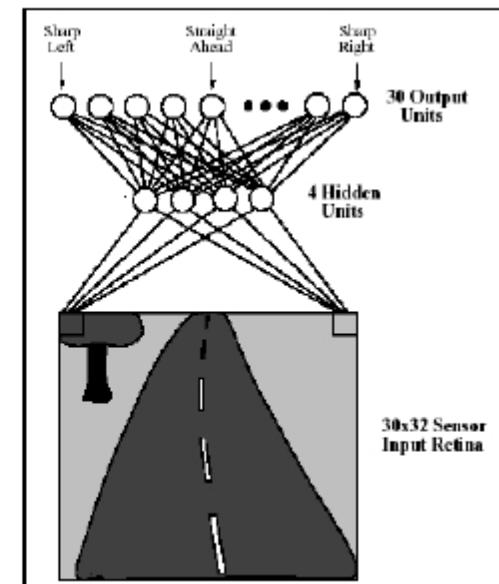
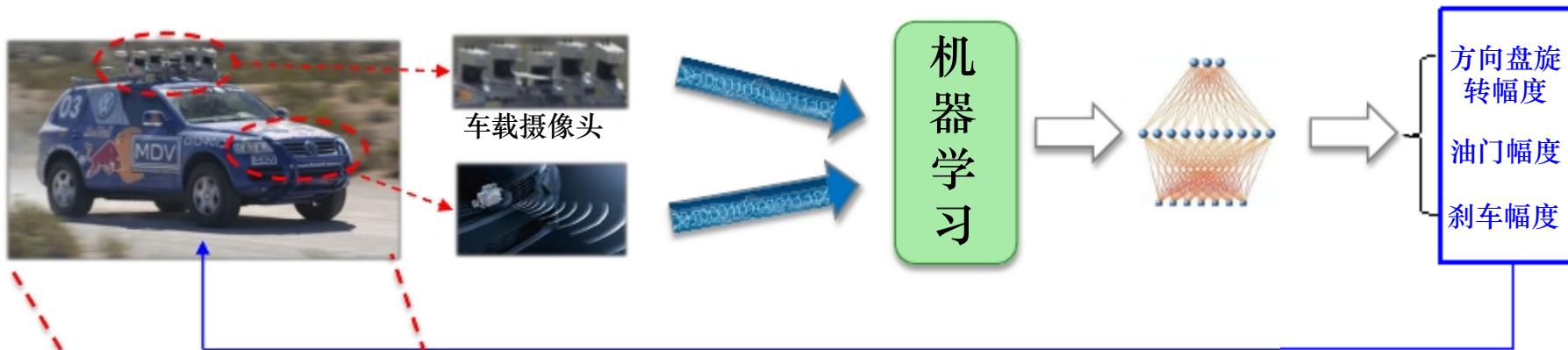
机器学习是作为“突破知识工程瓶颈”之利器而出现的

恰好在20世纪90年代中后期，人类发现自己淹没在数据的汪洋中，对自动数据分析技术——机器学习的需求日益迫切

# 1.3 | 机器学习的应用现状



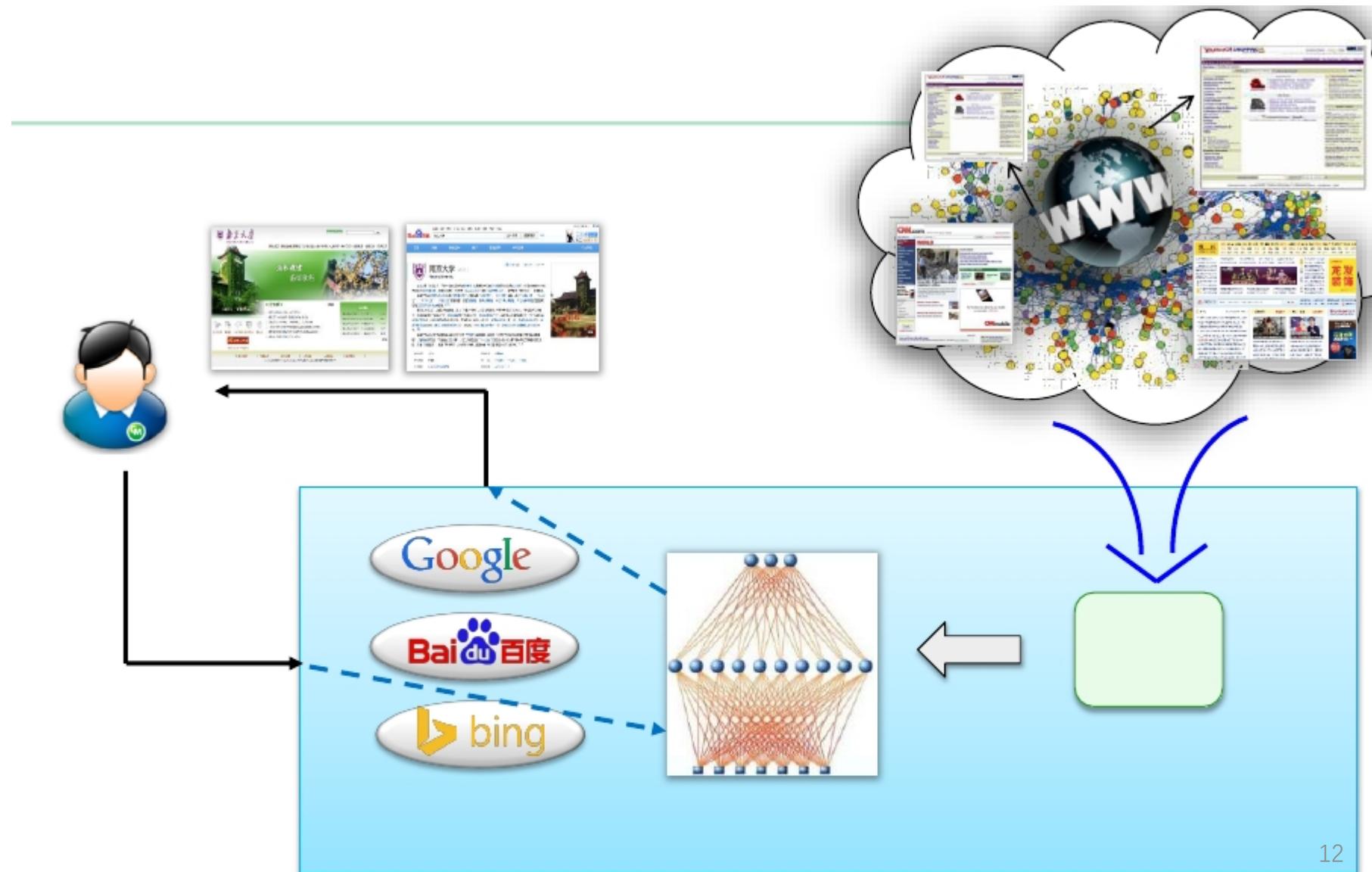
## 自动驾驶



# 1.3 | 机器学习的应用现状



## 互联网搜索



# 1.3 | 机器学习的应用现状



## 奥巴马胜选

### How Obama's data crunchers helped him win

TIME

By Michael Scherer

November 8, 2012 – Updated 1645 GMT (0045 HKT) | Filed under: Web

《时代》



## 奥巴马胜选

通过机器学习模型：

- 在总统候选人第一次辩论后，分析出哪些选民将倒戈，为每位选民找出一个最能说服他的理由；
- 精准定位不同选民群体，建议购买冷门广告时段，广告资金效率比2008年提高14%；
- 向奥巴马推荐，竞选后期应当在什么地方展开活动 —— 那里有很多争取对象；
- 借助模型帮助奥巴马筹集到创纪录的10亿美元。

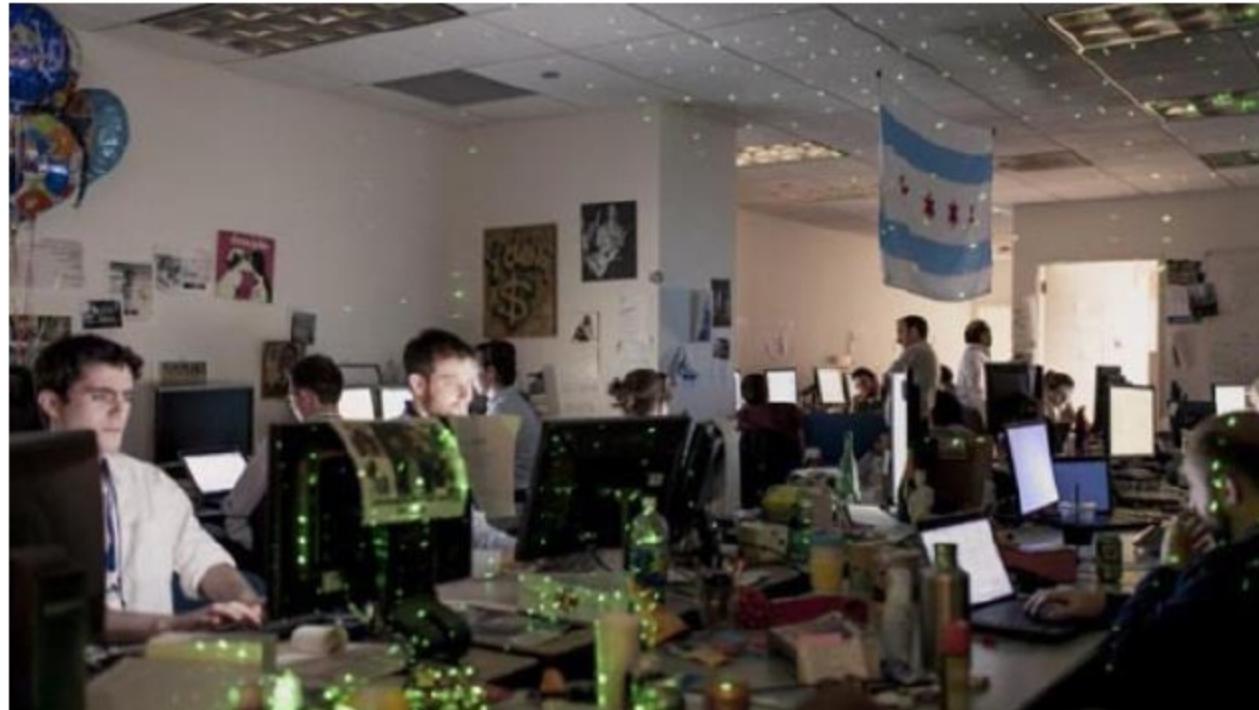
例如：利用模型分析出，明星乔治克鲁尼（George Clooney）对于年龄在40-49岁的美西地区女性颇具吸引力，而她们恰是最愿意为和克鲁尼/奥巴马共进晚餐而掏钱的人 …… 乔治克鲁尼为奥巴马举办的竞选筹资晚宴成功募集到1500万美元



# 1.3 | 机器学习的应用现状



## 奥巴马胜选

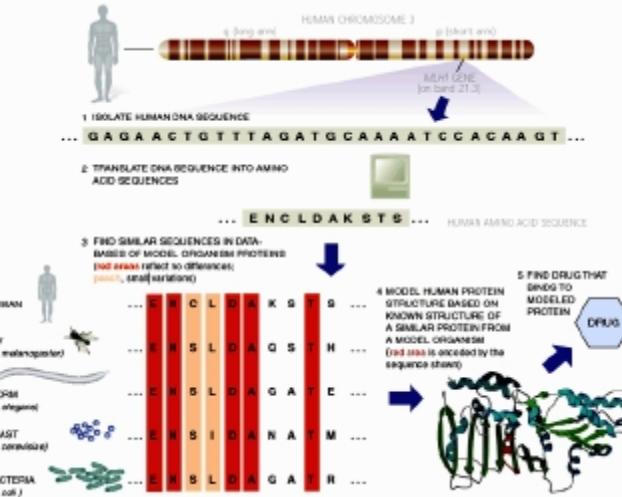


这个团队行动保密，定期向奥巴马报送结果；  
被奥巴马公开称为总统竞选的  
“核武器按钮” (“They are our nuclear codes”)

# 1.3 | 机器学习的应用现状



## 机器学习+大数据 无处不在





# 主要内容



- 1 机器学习发展与应用
- 2 相关分析和显著性
- 3 机器学习的基本任务
- 4 “宝贝在哪儿” 关联挖掘
- 5 公益课题 “宝贝在哪儿”
- 6 疫情风险分析



- 2.1 相关分析的意义
- 2.2 地理要素间相关关系的种类
- 2.3 地理相关程度的测度方法
- 2.4 相关系数的显著性检验

## 2.1 | 相关分析的意义



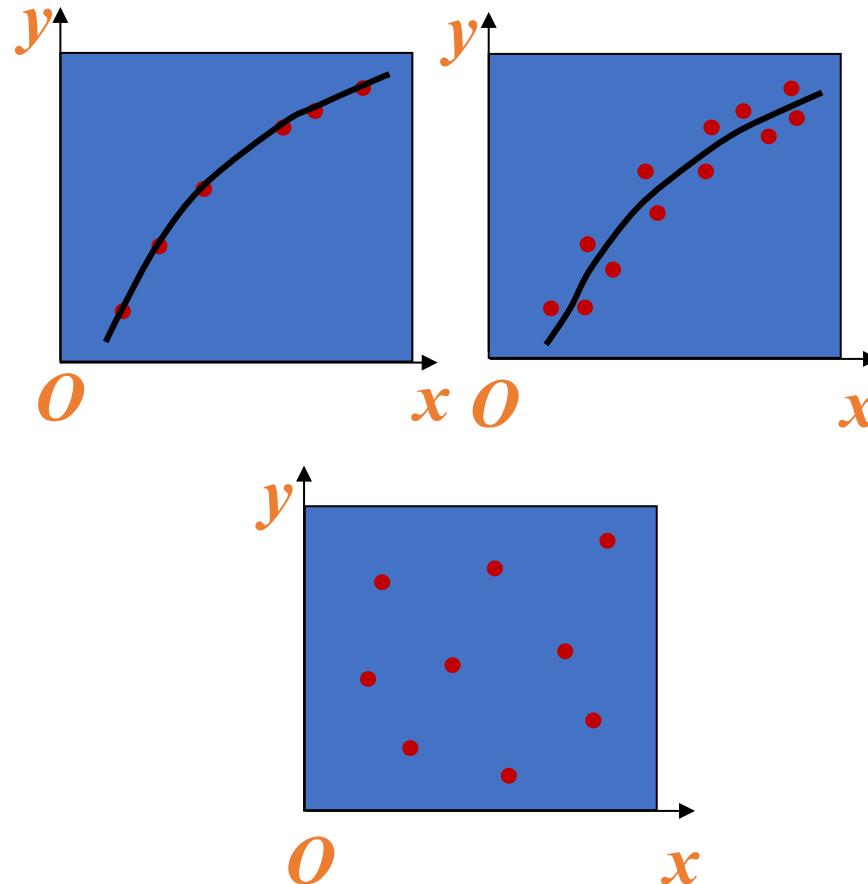
### 相关分析

自变量和因变量的判定

地理相关

要素之间关系的类型：

- 函数关系（完全相关）
- 相关关系（统计相关）
- 独立



## 2.1 | 相关分析的意义



函数关系与相关关系的联系和区别：

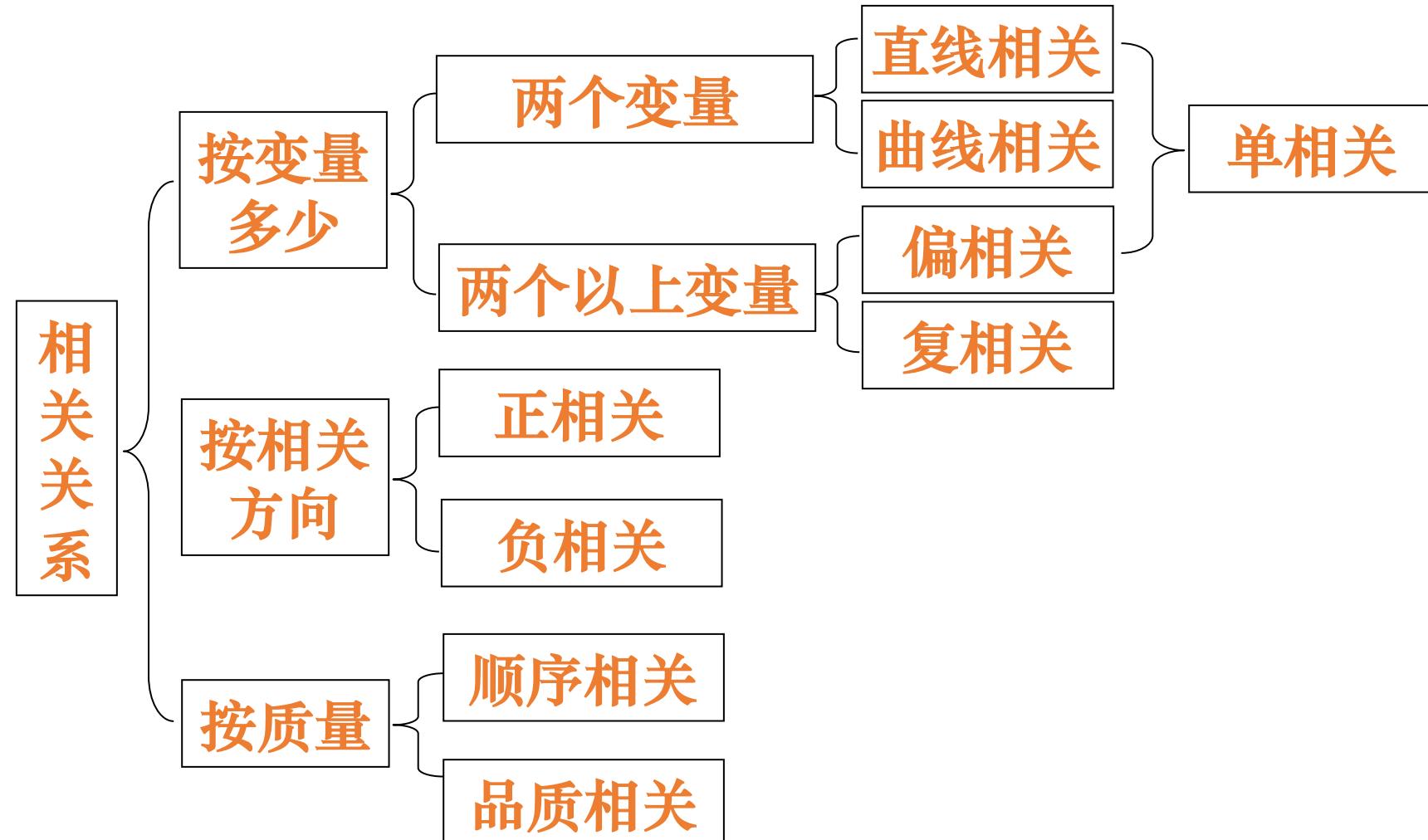
——联系：

在一定条件下是可以相互转化的。

——区别：

研究分析方法不同。

## 2.2 地理要素间相关关系的种类





- 2.3.1简单线性相关程度的测度
- 2.3.2多要素相关与相关矩阵
- 2.3.3偏相关与复相关



## 相关表

- 是一种显示变量之间相关关系的统计表。
- 将两个变量的对应值平行排列,且其中某一变量按其取值大小顺序排列,便可得到相关表。

某商店10名售货员的工龄和日工资的相关表

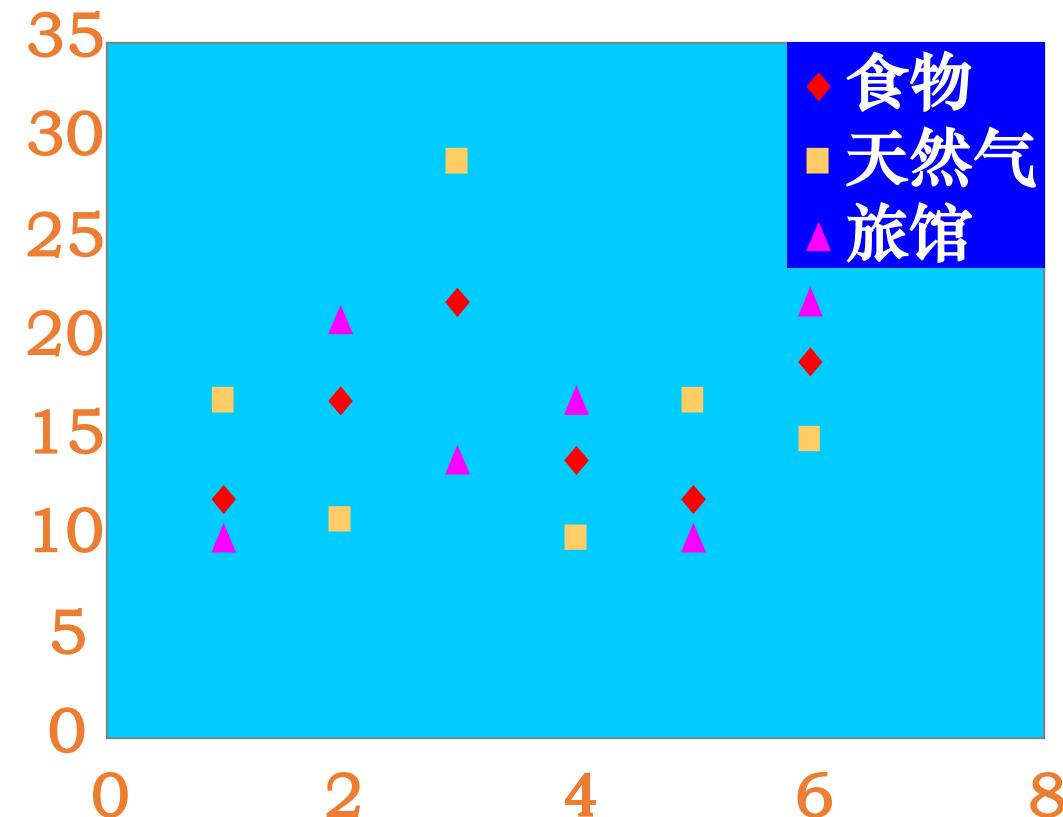
工龄 (年)	4	4	5	6	7	8	8	9	9	10
日工资 (元)	42	46	50	60	64	68	74	72	80	84

## 2.3 地理相关程度的测度方法



### 相关图(散点图)

□ 是将两个变量的对应值,在平面直角坐标系中用坐标点的形式描绘而成的图形。





### 2.3.1 简单线性相关程度的测度

#### ■ 相关系数：

用来度量直线相关程度和方向的指标。

- 1.一般常用的相关系数（ $r$ ）
- 2.顺序（等级）相关系数（ $rs$ ）



### 1.一般常用的相关系数 (pearson r)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

分子：两个变量的离差积和；

分母：两变量离差平方和之积的平方根。

## 2.3 地理相关程度的测度方法



### 简单线性相关

$$\begin{aligned} r &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} \\ &= \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left[ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] \left[ \sum y_i^2 - \frac{(\sum y_i)^2}{n} \right]}} \\ &= \frac{l_{xy}}{\sqrt{l_{xx} \cdot l_{yy}}} \end{aligned}$$

$$l_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \quad l_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \quad l_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

## 相关系数 $r$ 的性质：

### 简单线性相关

- ◆  $r \in [-1, 1]$ ；
- ◆ 当  $r > 0$  时为正相关， $r < 0$  时为负相关；
- ◆ 当  $|r| = 1$  时，则  $r = 1$  为完全正相关，  
 $r = -1$  为完全负相关；
- ◆ 当  $r = 0$  时，说明两变量之间完全无关；
- ◆ 当  $|r| \rightarrow 1$  时，说明两变量之间关系密切；
- ◆ 当  $|r| \rightarrow 0$  时，说明两变量之间相关程度差。



## 简单线性相关

相关系数  $r$  的性质：

$r = 0$	完全不相关；
$0 < r \leq 0.3$	微弱相关；
$0.3 \leq r \leq 0.5$	低度相关；
$0.5 \leq r \leq 0.8$	显著相关；
$0.8 \leq r < 1$	高度相关；
$r = 1$	完全相关。

例:某地区历年人均收入与商品销售额资料如下:

年份	人均收入 (百元)x	商品销售额 (百万元)y	xy	$x^2$	$y^2$
1998	24	11	264	576	121
1999	30	15	450	900	225
2000	32	14	448	1024	196
2001	34	16	544	1156	256
2002	38	20	760	1444	400
$\Sigma$	158	76	2466	5100	1198

要求计算x与y的相关系数，说明其相关方向和程度。

## 2.3 地理相关程度的测度方法



解：将计算表中的数值代入r计算公式得：

$$\begin{aligned} r &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{5 \times 2466 - 158 \times 76}{\sqrt{5 \times 5100 - 158^2} \sqrt{5 \times 1198 - 76^2}} \\ &= \frac{12330 - 12008}{\sqrt{536} \times \sqrt{214}} = 0.95 \end{aligned}$$

计算结果表明，人均收入与商品销售额之间存在高度的直线正相关关系。

## 2.3 地理相关程度的测度方法



### 2. 顺序（等级）相关系数 ( $r_s$ )

表示两个要素（变量）顺序间直线相关程度和方向的系数

公式： 
$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

设两个要素x和y有n对样本值，令 $T_1$ 代表要素x的序号（或位次）， $T_2$ 代表要素y的序号（或位次），则：

$$d_i^2 = (T_{1i} - T_{2i})^2$$

代表要素x和y的同一组样本位次差的平方。



**例：**全国1999年31个省(市、区)的总人口(x)和国内生产总值(y)及其位次列于下表中。试计算x与y之间的顺序相关系数。

省 (市、区)	总人口 (x) 及其位次		国内生产总值 (y) 及其位次		位次差的平方 $d_i^2 = (T_{1i} - T_{2i})^2$
	人口数 (万人)	位次 $T_1$	产值 (亿元)	位次 $T_2$	
北京	1257	26	2175.46	15	121
天津	959	27	1450.06	23	16
河北	6614	6	4569.19	6	0
...	...	...	...	...	...
$\Sigma$	124219	/	87671.13	/	962

## 2.3 地理相关程度的测度方法



解：将计算表中的数值代入 $r_s$ 计算公式得：

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$
$$= 1 - \frac{6 \times 962}{31(31^2 - 1)} = 0.806$$

即：总人口(x)与国内生产总值(y)之间的等级相关系数为0.806。计算结果表明，二者存在高度的直线正相关关系。



表1 秩相关系数检验的临界值

$n$	显著水平 $\alpha$		$n$	显著水平 $\alpha$	
	0.05	0.01		0.05	0.01
4	1.000	--	16	0.425	0.601
5	0.900	1.000	18	0.399	0.564
6	0.829	0.943	20	0.377	0.534
7	0.714	0.893	22	0.359	0.508
8	0.643	0.833	24	0.343	0.485
9	0.600	0.783	26	0.329	0.465
10	0.564	0.746	28	0.317	0.448
12	0.456	0.712	30	0.306	0.432
14	0.456	0.645	--	--	--

注:  $n$ 代表样本个数,  $\alpha$ 代表不同的置信水平, 也称 $r_\alpha$ 显著水平, 表中的数值为临界值。

## 2.3 地理相关程度的测度方法



在上例中， $n=31$ ，表中没有给出相应的样本个数下的临界值 $r_\alpha$ ，但是同一显著水平下，随着样本数的增大，临界值 $r_\alpha$ 减少。在 $n=30$ 时，查表得： $r_{0.01} = 0.432$ ，由于 $r'_{0.01}=0.8060 > r_{0.01} = 0.432$ ，所以在 $\alpha=0.01$ 的置信水平上来看，中国大陆各省（直辖市、自治区）人口规模与GDP是等级相关的。



## 2.3.2 多要素相关与相关矩阵

设有原始地理数据矩阵

$$\begin{matrix} & & 1 & 2 & \dots & n & \text{指标} \\ \begin{matrix} \text{要素} \\ 1 \\ 2 \\ \vdots \\ m \end{matrix} & \left[ \begin{matrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{matrix} \right] \end{matrix}$$

## 2.3 地理相关程度的测度方法



要测度两两要素之间的相关程度，公式为：

$$r_{ij} = \frac{\sum (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum (x_{ik} - \bar{x}_i)^2 \cdot \sum (x_{jk} - \bar{x}_j)^2}}$$

得到相关系数矩阵：

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix}$$

多要素相关矩阵的性质：

- 对角线上的元素均为1；
- 此矩阵为方阵；
- 沿对角线对称。

### 2.3.3 偏相关与复相关

① 定义：在多要素所构成的地理系统中，先不考虑其他要素的影响，而单独研究两个要素之间的相互关系的密切程度，这称为偏相关。用以度量偏相关程度的统计量，称为偏相关系数。

## 2.3 地理相关程度的测度方法



② 计算：  
3个要素的偏相关系数

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1-r_{14.3}^2)(1-r_{24.3}^2)}}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2}r_{34.2}}{\sqrt{(1-r_{14.2}^2)(1-r_{34.2}^2)}}$$

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}r_{43.2}}{\sqrt{(1-r_{13.2}^2)(1-r_{43.2}^2)}}$$

$$r_{23.14} = \frac{r_{23.1} - r_{24.1}r_{34.1}}{\sqrt{(1-r_{24.1}^2)(1-r_{34.1}^2)}}$$

## 2.3 地理相关程度的测度方法



例如：对于某4个地理要素 $x_1, x_2, x_3, x_4$ 的23个样本数据，经过计算得到了如下的单相关系数矩阵：

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0.416 & 0.346 & 0.579 \\ 0.416 & 1 & -0.592 & 0.950 \\ -0.346 & -0.592 & 1 & -0.469 \\ 0.579 & 0.950 & -0.469 & 1 \end{bmatrix}$$



利用公式计算一级偏向关系数，如表2所示：

表2 一级偏相关系数

$r_{12\cdot 3}$	$r_{13\cdot 2}$	$r_{14\cdot 2}$	$r_{14\cdot 3}$	$r_{23\cdot 1}$	$r_{24\cdot 1}$	$r_{24\cdot 3}$	$r_{24\cdot 1}$	$r_{34\cdot 2}$
0.821	0.808	0.647	0.895	-0.863	0.956	0.945	-0.875	0.371

利用公式计算二级偏相关系数，如表3所示：

表3 二级偏相关系数

$r_{12\cdot 34}$	$r_{13\cdot 24}$	$r_{14\cdot 23}$	$r_{23\cdot 14}$	$r_{24\cdot 13}$	$r_{34\cdot 12}$
-0.170	0.802	0.635	-0.187	0.821	-0.337

4个要素的一级偏相关系数有12个，这里给出了9个；二级偏相关系数有6个，这里全部给出来了。



### 偏相关系数的性质：

- ① 偏相关系数分布的范围在-1到1之间；
- ② 偏相关系数的绝对值越大，表示其偏相关程度越大；
- ③ 偏相关系数的绝对值必小于或最多等于由同一系列资料所求得的复相关系数，即  $R_{1\cdot 23} \geq |r_{12\cdot 3}|$ 。



## 相关系数的显著性检验

**目的：**判定相关系数是否有意义

**简单线性相关系数的显著性检验步骤**

(1) 计算出相关系数 $r$ 。

(2) 给定显著性水平 $\alpha$ ,按 $n-2$ 查相关系数临界值( $r_\alpha$ )表，查出相应的临界值  $r_\alpha$ 。

(3) 比较 $|r|$ 与 $r_\alpha$ 的大小：

当 $|r| \geq r_\alpha$ 时,说明两变量在 $\alpha$ 水平上达到显著性；

当 $|r| < r_\alpha$ 时,说明两变量在 $\alpha$ 水平上没有达到所要求的精度。



## 相关系数的显著性检验

相关系数是根据要素之间的样本值计算出来，它随着样本数的多少或取样方式的不同而不同，因此它只是要素之间的样本相关系数，只有通过检验，才能知道它的可信度。

检验是通过在给定的置信水平下，查相关系数检验的临界值表来实现的。

表4 检验相关系数  $\rho = 0$  的临界值 ( $r_\alpha$ ) 表

$$p\{|r| > r_\alpha\} = \alpha$$

$f$	0.10	0.05	0.02	0.01	0.001
1	0.987 69	0.996 92	0.999 507	0.999 877	0.999 998
2	0.900 00	0.950 00	0.980 00	0.990 00	0.999 000
3	0.805 4	0.878 3	0.934 33	0.958 73	0.991 160
4	0.729 3	0.811 4	0.882 2	0.917 20	0.974 06
5	0.669 4	0.754 5	0.832 9	0.874 5	0.950 74
6	0.621 5	0.706 7	0.788 7	0.834 3	0.924 93
7	0.582 2	0.666 4	0.749 3	0.797 7	0.898 2
8	0.549 4	0.631 9	0.715 5	0.764 6	0.872 1
9	0.521 4	0.602 1	0.685 1	0.734 8	0.847 1
10	0.497 3	0.576 0	0.658 1	0.707 9	0.823 3
11	0.476 2	0.552 9	0.633 9	0.683 5	0.801 0
12	0.457 5	0.532 4	0.612 0	0.661 4	0.780 0



在表1中， $f$  称为自由度，其数值为  $f=n-2$ ， $n$  为样本数；上方的 $\alpha$ 代表不同的置信水平；表内的数值代表不同的置信水平下相关系数 $\rho = 0$ 的临界值，即 $r_\alpha$ ；公式 $p\{|r| > r_\alpha\} = \alpha$ 的意思是当所计算的相关系数 $r$ 的绝对值大于在 $\alpha$ 水平下的临界值 $r_\alpha$ 时，两要素不相关（即 $\rho = 0$ ）的可能性只有 $\alpha$ 。



## 偏相关系数的显著性检验

偏相关系数的显著性检验，一般采用 $t$ 检验法。其统计量计算公式为

$$t = \frac{r_{12\cdot34\dots m}}{\sqrt{1-r^2_{12\cdot34\dots m}}} \sqrt{n-m-1}$$

式中： $r_{12\cdot34\dots m}$  为偏相关系数；  $n$  为样本数；  $m$  为自变量个数。



譬如，对于上例计算得到的偏相关系数  $r_{24.13} = 0.821$ ，由于  $n=23$ ,  $m=3$ ，故

$$t = \frac{0.821}{\sqrt{1 - 0.821^2}} \sqrt{23 - 3 - 1} = 6.268$$

查  $t$  分布表，在自由度为  $23-3-1=19$  时，  
 $t_{0.001}=3.883$ ，显然  $t > t_{\alpha}$  这表明在置信度水平  
 $\alpha=0.001$  上，偏相关系数  $r_{24.13}$  是显著的。



# 主要内容



- 1 机器学习发展与应用
- 2 相关分析和显著性
- 3 机器学习的基本任务
- 4 “宝贝在哪儿” 关联挖掘
- 5 公益课题 “宝贝在哪儿”
- 6 疫情风险分析



- 3.1 回归任务
- 3.2 分类任务
- 3.3 聚类任务

# 3.1 | 回归



**回归：**回归任务通常是用来预测一个连续值。

大小(平方英尺)	价格1000美元(y)	大小(平方英尺)	房间数量	层数	房龄	价格1000美元(y)
2104	460	2104	5	1	45	460
1416	232	1416	3	2	40	232
1534	315	1534	3	2	30	315
852	178	852	2	1	36	178
...	...	...	...	...	...	...

一元线性回归

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

多元线性回归

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

# 3.1 | 回归



一元线性回归：

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

repeat until convergence {  
     $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$   
    (for  $j = 1$  and  $j = 0$ )  
}

多元线性回归：

$$h_{\theta}(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

$$J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Repeat {  
     $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \dots, \theta_n)$   
}

**分类：**分类任务通常是用来预测一个离散的值。

在分类问题中，我们尝试预测的是结果是否属于某一个类（例如正确或错误）。分类问题的例子有：判断一封电子邮件是否是垃圾邮件；判断一次金融交易是否是欺诈；区别一个肿瘤是恶性的还是良性的。

$y \in \{0, 1\}$

0:	“负样本” (例: 良性肿瘤)
1:	“正样本” (例: 恶性肿瘤)

## 3.2 | 分类

分类任务常用评价指标：

真正(True Positive , TP): 预测为正的正样本

假正(False Positive , FP): 预测为正的负样本

假负(False Negative , FN): 预测为负的正样本

真负(True Negative , TN): 预测为负的负样本

准确率：

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

精确率：

$$\text{Precision} = \frac{TP}{TP+FP}$$

召回率：

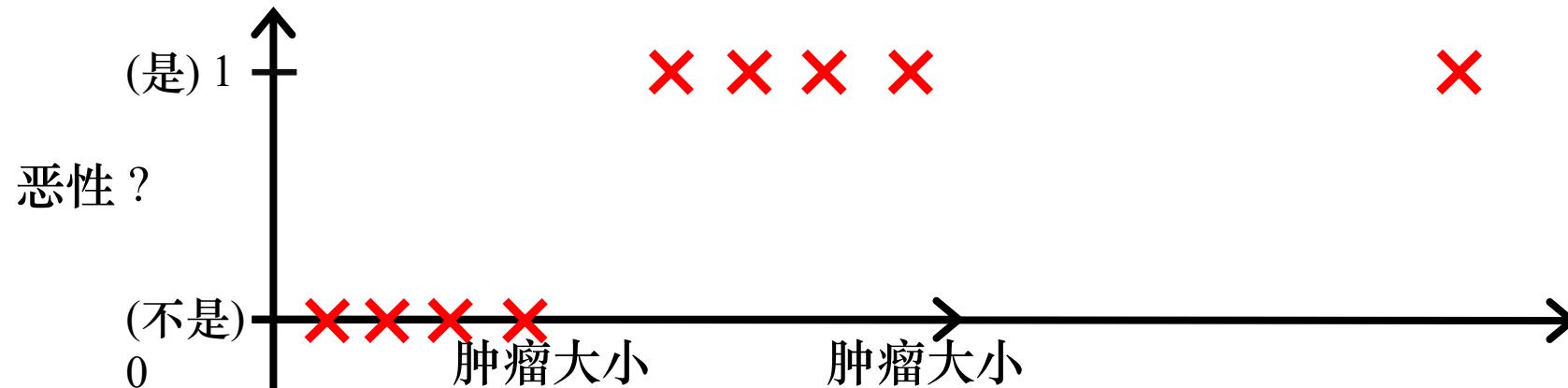
$$\text{Recall} = \frac{TP+TN}{TP+FN+FP+TN}$$

F1值：

$$\frac{2}{F1} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}$$

还可以通过ROC曲线、AUC曲线、PR曲线  
判断分类器的效果

## 3.2 | 分类-Logistic回归



输出数据分类阈值  $h_{\theta}(x)$  为 0.5:

如果  $h_{\theta}(x) \geq 0.5$  , 预测 “y = 1”

如果  $h_{\theta}(x) < 0.5$  , 预测 “y = 0”

## 3.2 | 分类-Logistic回归



### 逻辑回归：

真正类别:  $y = 0 \text{ or } 1$

$$h_{\theta}(x) = \theta^T x$$

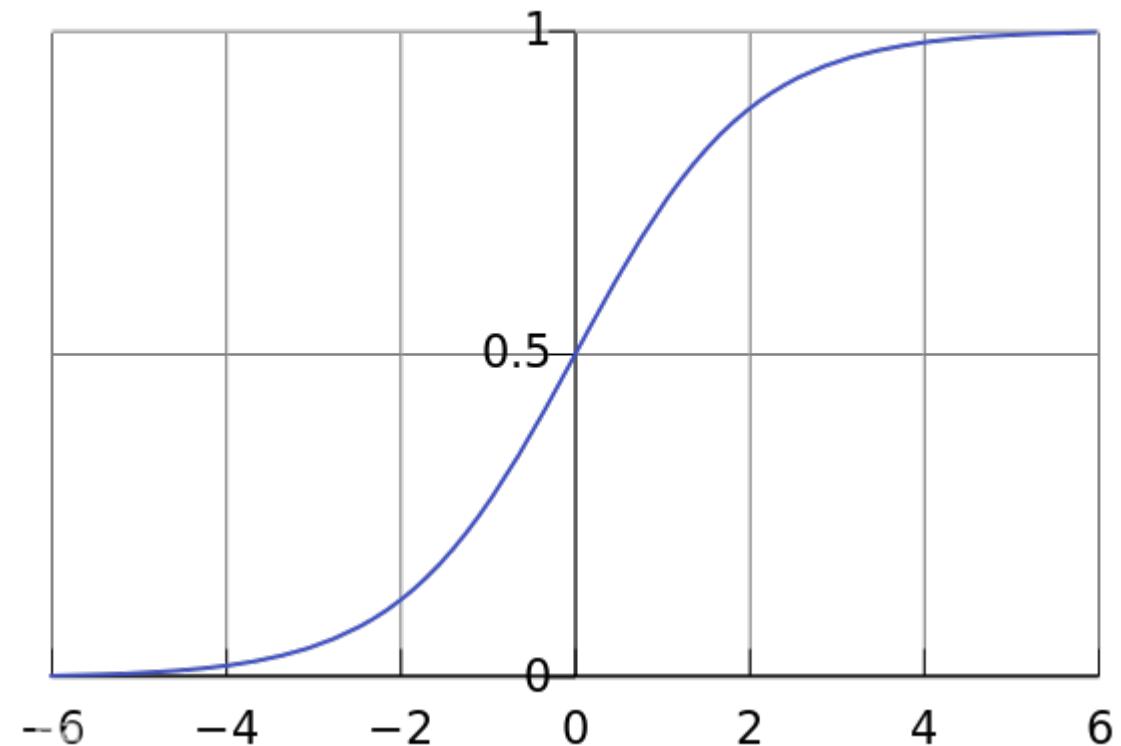
但是预测值  $h_{\theta}(x)$  可能 $> 1$  or  $< 0$

$$h_{\theta}(x) = g(\theta^T x)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

$$0 \leq h_{\theta}(x) \leq 1$$

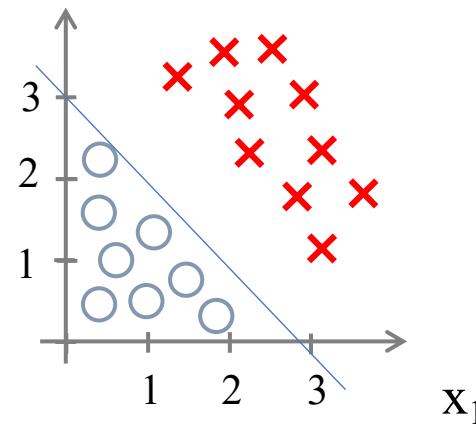
Logistic function:  $g(z) = \frac{1}{1 + e^{-z}}$



## 3.2 | 分类-Logistic回归



判别边界：

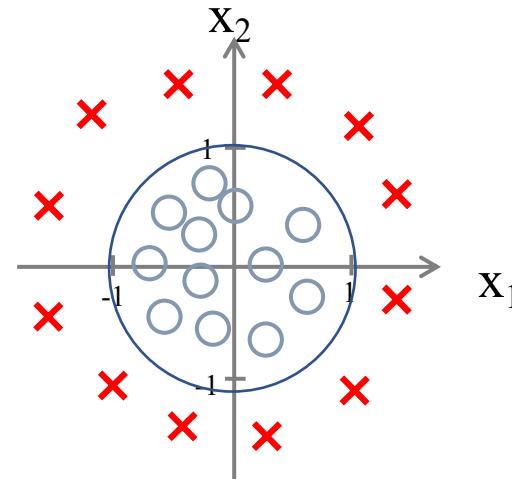


$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

预测 “ $y = 1$ ” 如果  $-3 + x_1 + x_2 \geq 0$

预测 “ $y = 1$ ” 如  $-1 + x_1^2 + x_2^2 \geq 0$



## 3.2 | 分类-决策树

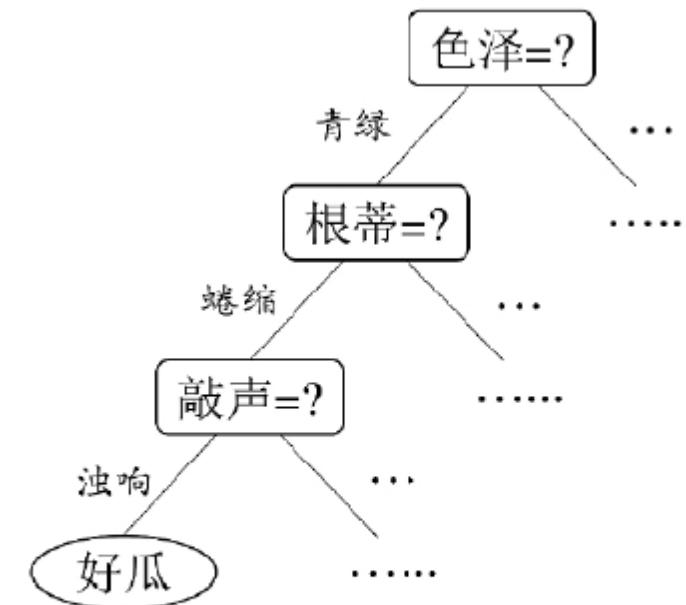
### 决策树：

决策树基于“树”结构进行决策

- 每个“内部结点”对应于某个属性上的“测试”(test)
- 每个分支对应于该测试的一种可能结果（即该属性的某个取值）
- 每个“叶结点”对应于一个“预测结果”

学习过程：通过对训练样本的分析来确定“划分属性”（即内部结点所对应的属性）

预测过程：将测试示例从根结点开始，沿着划分属性所构成的“判定测试序列”下行，直到叶结点



## 3.2 | 分类-决策树



- 第一个决策树算法：CLS (Concept Learning System)

[E. B. Hunt, J. Marin, and P. T. Stone's book "*Experiments in Induction*" published by Academic Press in 1966]

- 使决策树受到关注、成为机器学习主流技术的算法：ID3

[J. R. Quinlan's paper in a book "*Expert Systems in the Micro Electronic Age*" edited by D. Michie, published by Edinburgh University Press in 1979]

- 最常用的决策树算法：C5.5

[J. R. Quinlan's book "*C5.5: Programs for Machine Learning*" published by Morgan Kaufmann in 1993]

- 可以用于回归任务的决策树算法：CART (Classification and Regression Tree)

[L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone's book "*Classification and Regression Trees*" published by Wadsworth in 1984]

- 基于决策树的最强大算法：RF (Random Forest)

[L. Breiman's MLJ'01 paper "*Random Forest*"]

策略：“分而治之” (divide-and-conquer)

自根至叶的递归过程

在每个中间结点寻找一个“划分” (split or test) 属性

三种停止条件：

- (1) 当前结点包含的样本全属于同一类别，无需划分；
- (2) 当前属性集为空，或是所有样本在所有属性上取值相同，无法划分；
- (3) 当前结点包含的样本集合为空，不能划分.

## 信息增益

信息熵 (entropy) 是度量样本集合 “纯度” 最常用的一种指标

假定当前样本集合  $D$  中第  $k$  类样本所占的比例为  $P_k$   
则  $D$  的信息熵定义为：

$$Ent(D) = \sum_{k=1}^n P_k \log P_k$$

$Ent(D)$  越小，  $D$  的纯度越高

信息增益直接以信息熵为基础，计算当前划分对信息熵所造成的变化

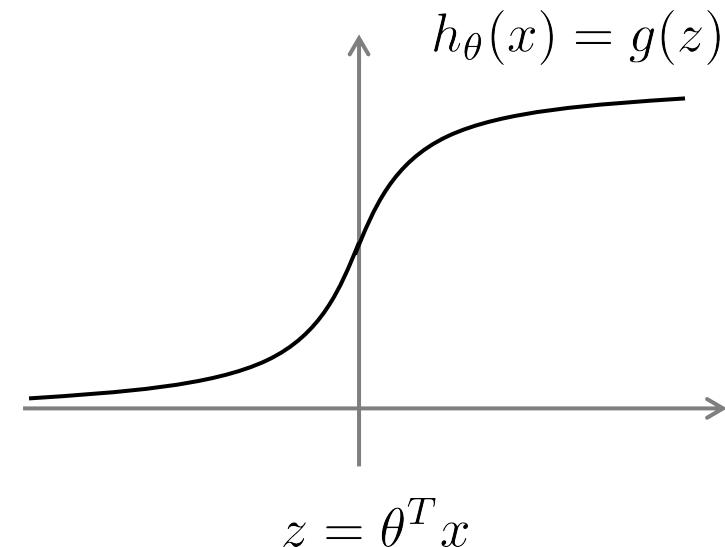
## 3.2 | 分类-支持向量机



支持向量机：

逻辑回归：

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$



If  $y = 1$ ,

$h_{\theta}(x) \approx 1$ ,  $\theta^T x \gg 0$

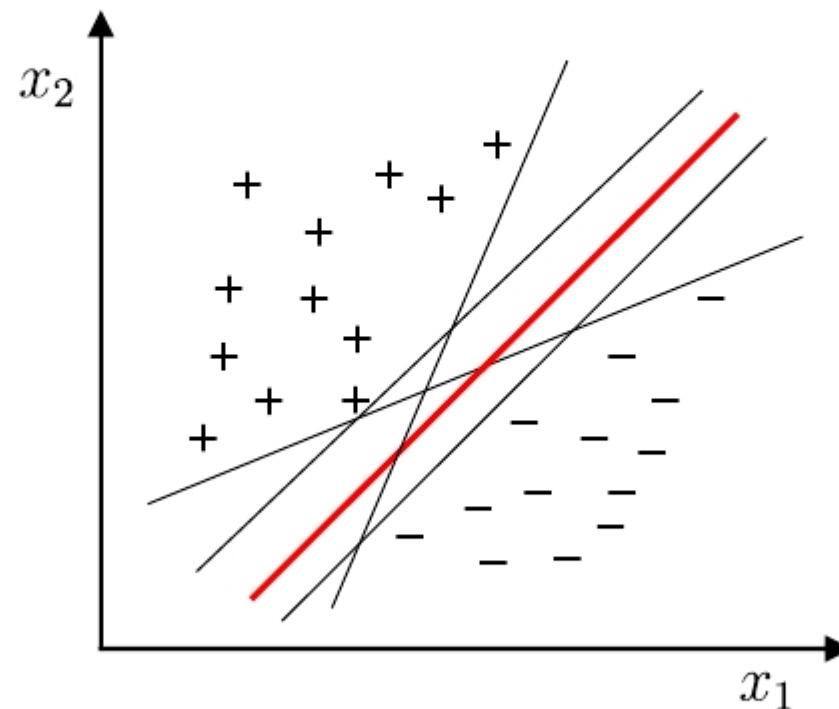
If  $y = 0$ ,

$h_{\theta}(x) \approx 0$ ,  $\theta^T x \ll 0$

## 3.2 | 分类-支持向量机



将训练样本分开的超平面可能有很多，哪一个更好呢？



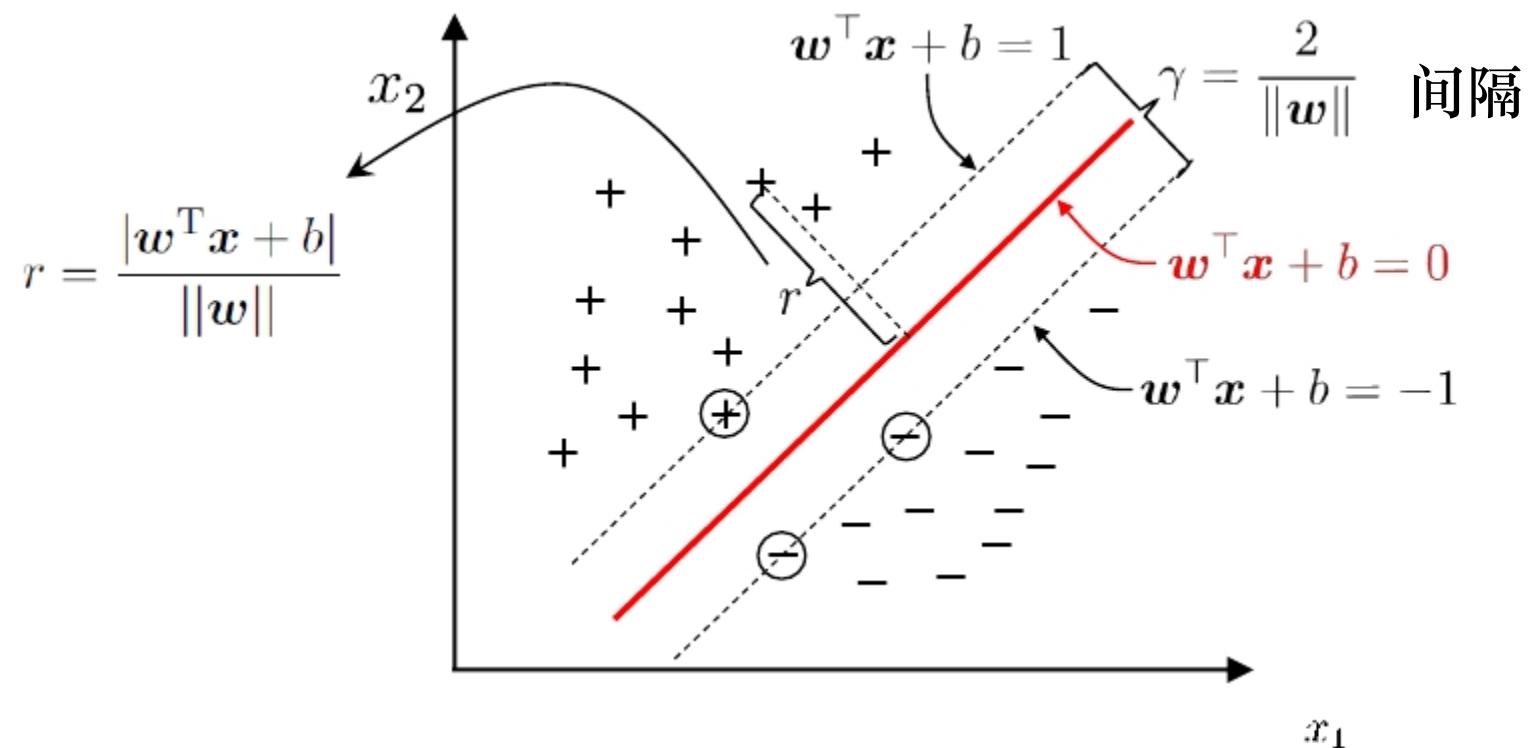
“正中间”的：鲁棒性最好，泛化能力最强

## 3.2 | 分类-支持向量机



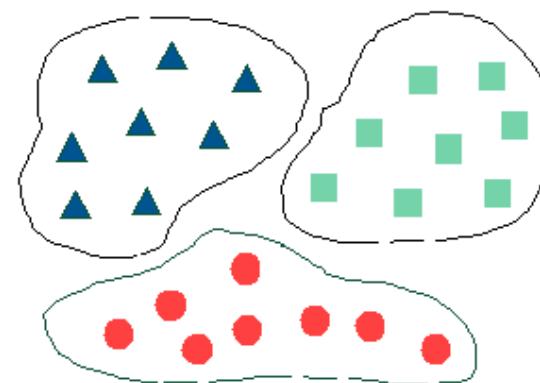
### 间隔 (margin)与支持向量 (support vector)

超平面方程 : $\mathbf{w}^\top \mathbf{x} + b = 0$



## 聚类：

将数据样本划分为若干个通常不相交的“簇”



既可以作为一个单独过程（用于找寻数据内在的分布结构）也可作为分类等其他学习任务的前驱过程

## 聚类：

聚类性能度量，亦称聚类“有效性指标” (validity index)

### □ 外部指标 (external index)

将聚类结果与某个“参考模型” (reference model) 进行比较  
如 Jaccard 系数，FM 指数，Rand 指数

### □ 内部指标 (internal index)

直接考察聚类结果而不用任何参考模型  
如 DB 指数，Dunn 指数等

基本想法：

- “簇内相似度” (intra-cluster similarity) 高，且“簇间相似度” (inter-cluster similarity) 低

## 距离计算：

距离度量 (distance metric) 需满足的基本性质：

非负性： $dist(x_i, x_j) \geq 0$ ;

同一性： $dist(x_i, x_j) = 0$  当且仅当  $x_i = x_j$ ;

对称性： $dist(x_i, x_j) = dist(x_j, x_i)$ ;

直递性： $dist(x_i, x_j) \leq dist(x_i, x_k) + dist(x_k, x_j)$ ;

常用距离形式：闵可夫斯基距离 (Minkowski distance)

$$dist_{mk}(x_i, x_j) = \left( \sum_{u=1}^n |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

$p = 1$ : 曼哈顿距离(Manhattan distance)    $p = 2$ : 欧氏距离(Euclidean distance)

## 常见聚类方法：

### □ 原型聚类

亦称“基于原型的聚类” (prototype-based clustering)

假设：聚类结构能通过一组原型刻画

过程：先对原型初始化，然后对原型进行迭代更新求解

代表：**k均值聚类，学习向量量化 (LVQ)，高斯混合聚类**

### □ 密度聚类

亦称“基于密度的聚类” (density-based clustering)

假设：聚类结构能通过样本分布的紧密程度确定

过程：从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇

代表：**DBSCAN, OPTICS, DENCLUE**

### □ 层次聚类 (hierarchical clustering)

假设：能够产生不同粒度的聚类结果

过程：在不同层次对数据集进行划分，从而形成树形的聚类结构

代表：**AGNES (自底向上), DIANA (自顶向下)**

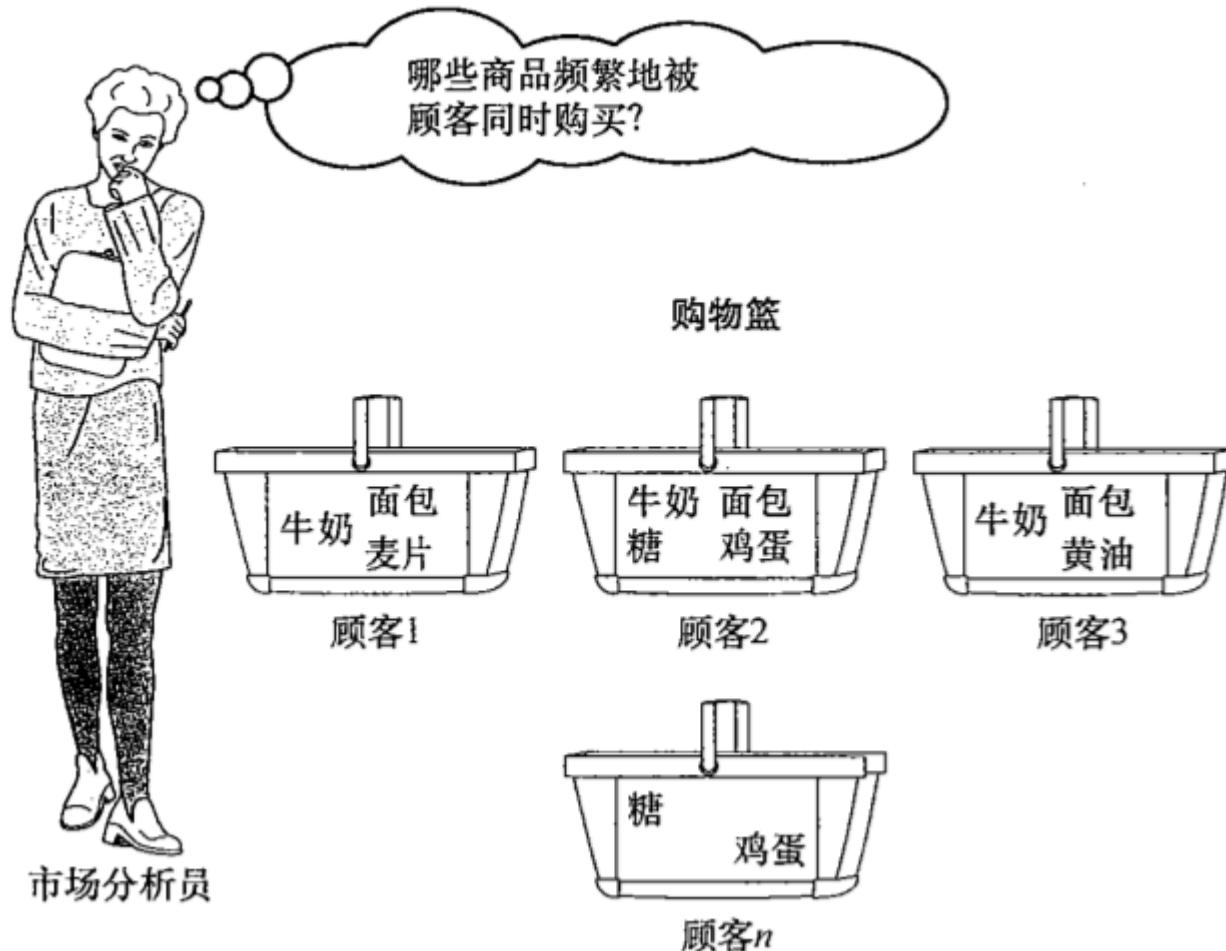


# 主要内容



- 1 机器学习发展与应用
- 2 相关分析和显著性
- 3 机器学习的基本任务
- 4 “宝贝在哪儿” 关联挖掘
- 5 公益课题 “宝贝在哪儿”
- 6 疫情风险分析

# 4.1 | 关联分析



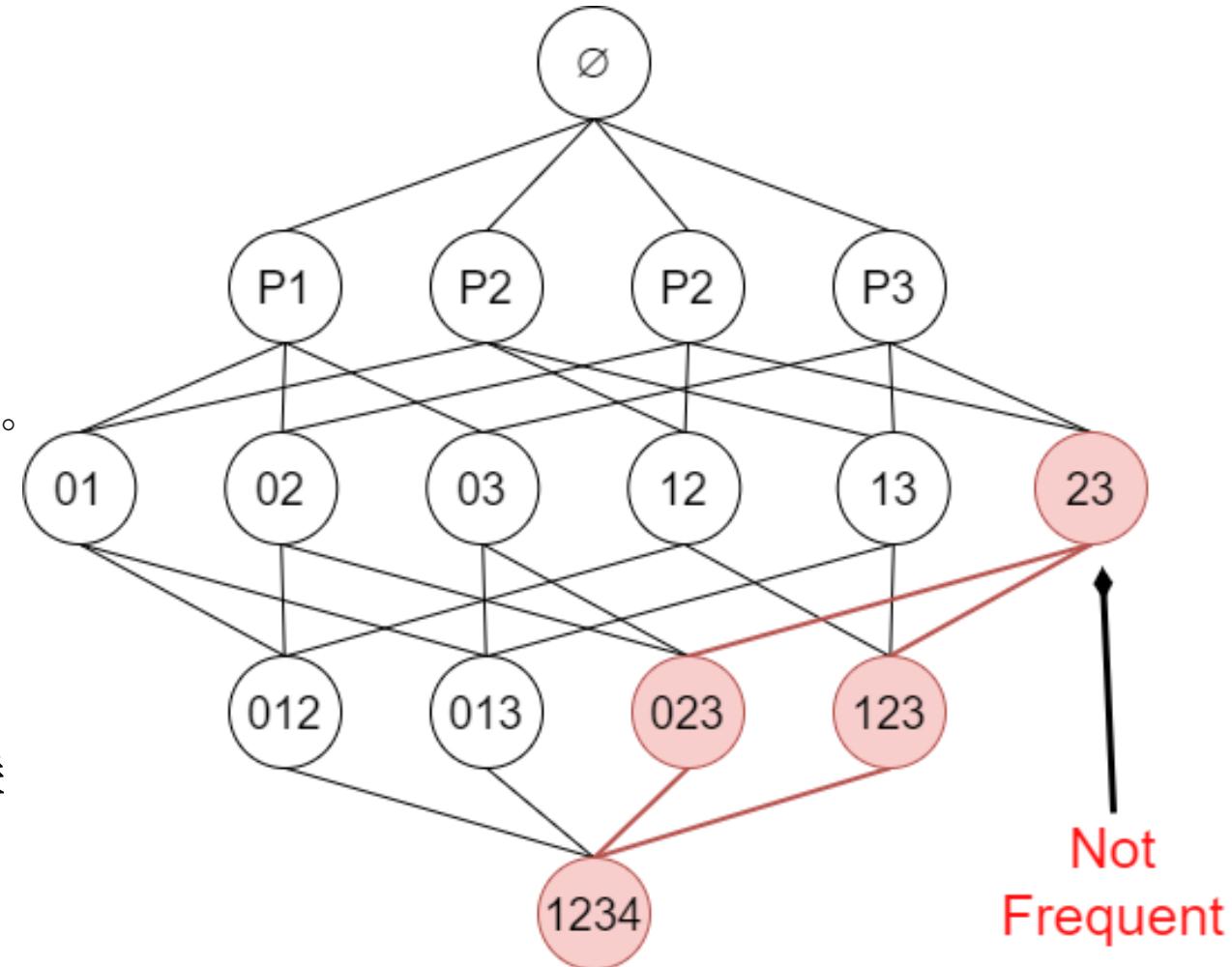
1. 从大规模数据集中寻找物品间的隐含关系被称作关联分析或者关联规则学习。
2. 在关联分析中，最重要的是计算频繁项集和关联规则。其中频繁项集的支持度可以用来表示该项集是否频繁；关联规则的置信度可以用来表示挖掘到的关联规则是否可信。
3. 计算支持度和置信度的原理很简单，统计每种商品集合的排列组合，求其占全部项集的比例即可。但对于成千上万的购买规则来说这种方式非常的慢，这里就需要引入Apriori规则减少计算量。

## 4.2 | 关联分析-Apriori

Apriori原理为

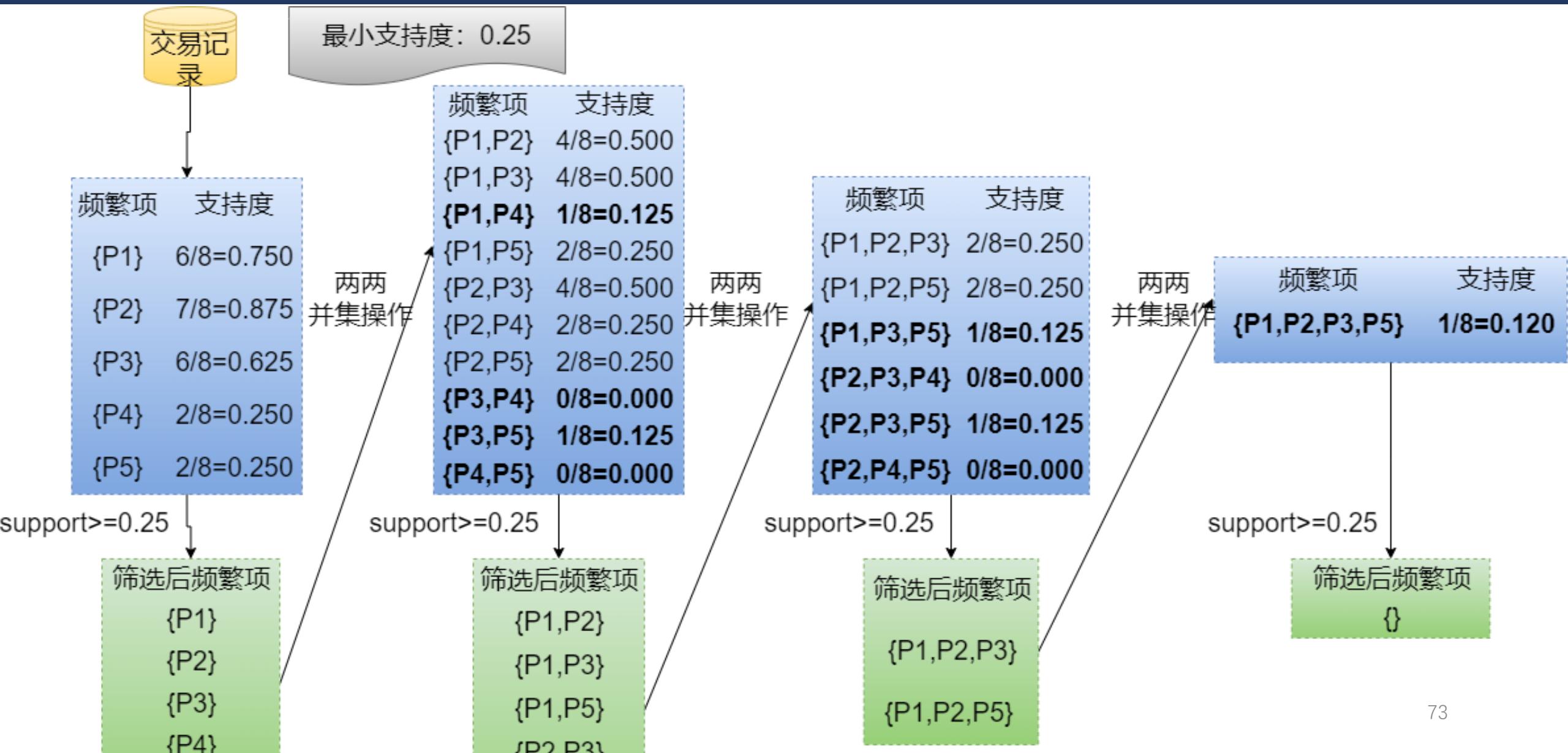
- 某个项集是频繁的，那他的子集也是频繁的
- 某个项集是非频繁的，那他的所有超集也是非频繁的  
(超集：包含某子集的集合都叫做超集)

右图中，由于{2,3}是非频繁的，因此{0, 2, 3}, {1, 2, 3}, {0, 1, 2, 3}也都是不频繁的，因此无需计算他们的支持度。  
这就是Apriori算法的剪枝过程。



接下来将会分别介绍频繁项集支持度的计算方法和关联规则置信度的计算方法。

## 4.2 | 关联分析-支持度



## 4.2 | 关联分析-置信度

根据频繁项集，我们可以挖掘出关联规则。如我们可以从频繁项集{P1, P2, P3}中挖掘出以下六种可能的关联规则

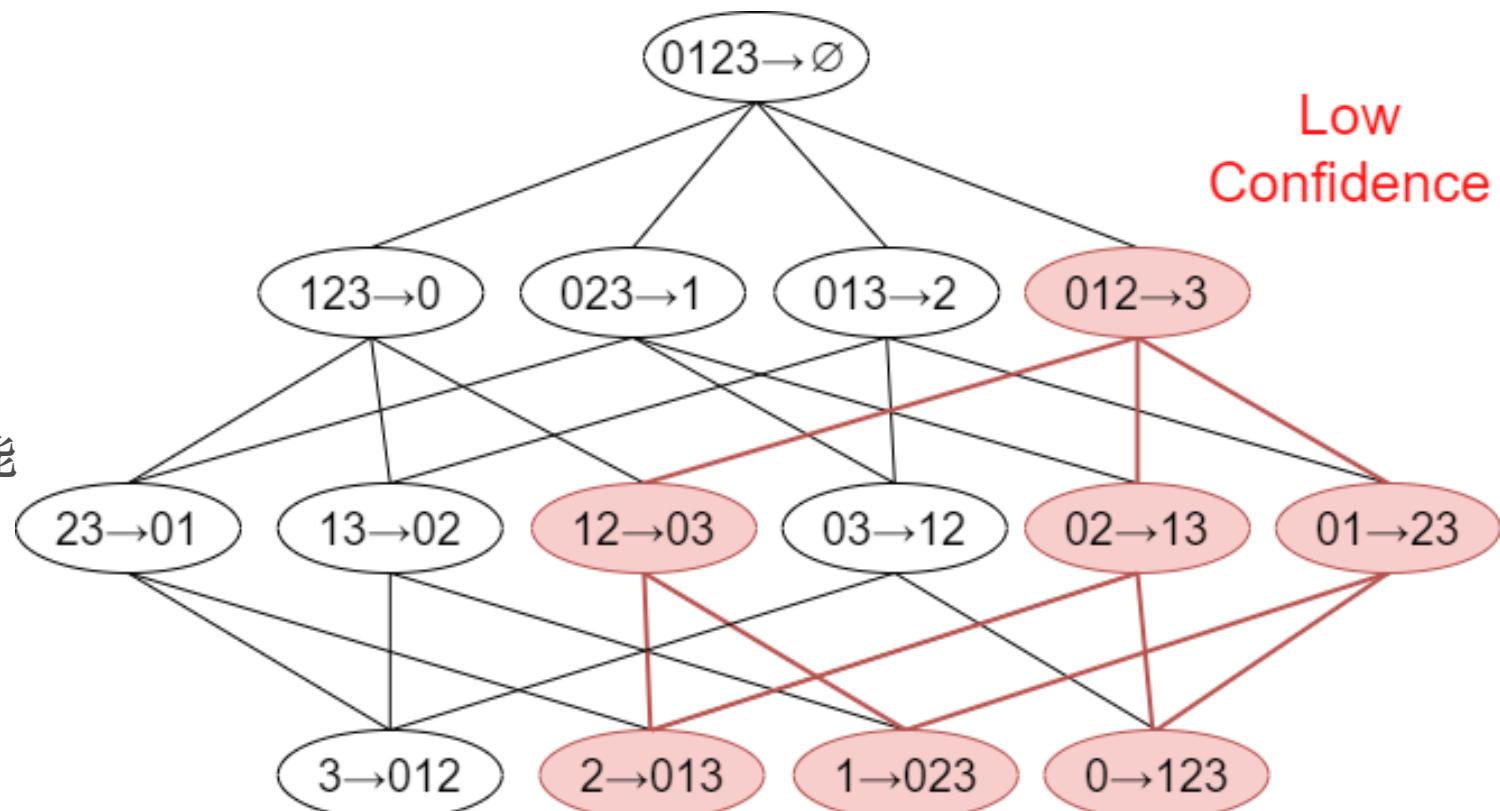
- $\{P1\} \rightarrow \{P2, P3\}$
- $\{P2\} \rightarrow \{P1, P3\}$
- $\{P3\} \rightarrow \{P1, P2\}$
- $\{P1, P2\} \rightarrow \{P3\}$
- $\{P2, P3\} \rightarrow \{P1\}$
- $\{P1, P3\} \rightarrow \{P2\}$

一个具有N个元素的频繁项集，共有M个可能的关联规则

$$M = \sum_{N=1}^{i=1} C_N^i$$

下图是一个频繁4项集的所有关联规则网格示意图，其中也存在类似于支持度的剪枝过程。

$$M = C_4^1 + C_4^2 + C_4^3 = 14$$



## 4.2 | 关联分析-置信度

频繁项	出现次数
{P1,P2}	4
{P1,P3}	4
{P1,P5}	2
{P2,P3}	4
{P2,P4}	2
{P2,P5}	2

频繁项	出现次数
{P1,P2,P3}	2
{P1,P2,P5}	2

首先找到所有  
频繁子项

频繁项集子项

- {P1,P2}
- {P1,P3}
- {P2,P3}
- {P1,P2}
- {P1,P5}
- {P2,P5}

最终关联关系

- P1→P2
- P2→P1
- P1→P3
- P5→P1
- P2→P3
- P3→P2
- P4→P2
- P5→P2

0.5 →

Confidence $\geq$ 0.5

最终关联关系  
 $\{P1\} \rightarrow \{P1, P2\}$

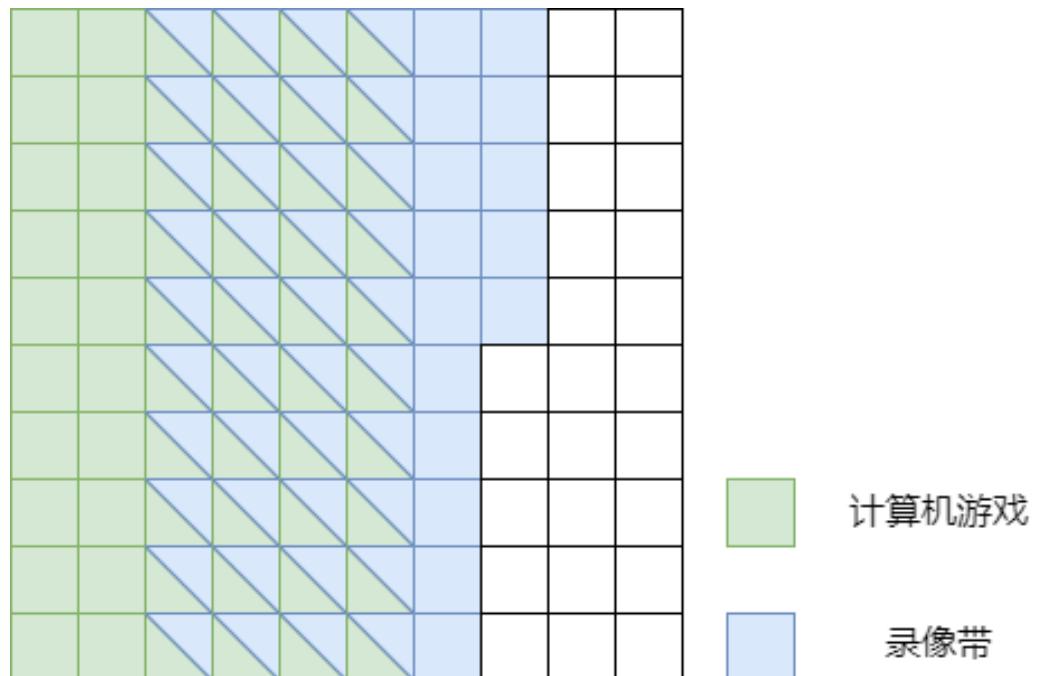
生成  
关联规则

关联规则	置信度
$\{P3\} \rightarrow \{P1, P2\}$	$2/6 \approx 0.333$
$\{P2\} \rightarrow \{P1, P3\}$	$2/7 \approx 0.288$
$\{P1\} \rightarrow \{P2, P3\}$	$2/6 \approx 0.333$
$\{P5\} \rightarrow \{P1, P2\}$	$2/2 = 1.000$
$\{P2\} \rightarrow \{P1, P5\}$	$2/7 \approx 0.288$
$\{P1\} \rightarrow \{P2, P5\}$	$2/6 \approx 0.333$

## 4.2 | 关联分析-提升度



支持度和置信度足够高的规则并不一定是有趣的。以下面的购买记录为例。



在所要分析的事务中，数据显示6000个顾客购买了计算机游戏，7500个购买了录像带，4000个同时购买了计算机游戏和录像带。设置最小支持度30%，最小置信度60%，则可以发现下列的关联规则

$\text{buys(“计算机游戏”)} \Rightarrow \text{buys(“录像带”)}$   
[support=40%, confidence=66%]

- Support =  $4000/1000 = 0.40$
- Confidence =  $4000/6000 = 0.66$

该规则是强关联规则，但却是一种误导。因为单独购买录像带的概率是75%，比66%还要高。实际上，计算机游戏和录像带的购买是负相关的，购买计算机游戏后实际上降低了购买录像带的可能性。因此需要一套计算方法衡量关联度的可信程度，才会减少作出错误决定的概率。



## 4.2 | 关联分析-提升度

提升度(lift)是一种简单的相关性度量。计算公式如下

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

其中P代表概率，比如 $P(A \cup B)$ 就是AB同时出现在一个事务中的概率。其计算出来的结果含义如下。

- 大于1: A和B是正相关的
- 等于1: A和B是独立的
- 小于1: A和B是负相关的

- $P(\text{'计算机游戏'}) = 0.60$
- $P(\text{'录像带'}) = 0.75$
- $P(\text{'计算机游戏, 录像带'}) = 0.40$
- $Lift = 0.40 / (0.75 * 0.60) = 0.89$

由于lift小于1，因此购买计算机游戏和购买录像带呈现负相关，因此应该舍弃该规则。

# 4.3 | “宝贝在哪儿”关联挖掘

注意：本站所有寻亲帮助均为免费，宝贝回家不会以任何理由收取费用，请勿上当受骗。 年龄人像APP下载

 宝贝回家 www.baobeihuijia.com



宝贝回家  
公益寻亲  
部警

首页 寻亲登记 志愿者登记 社会新闻 相关视频 紧急求助 论坛 党建平台

【简介】宝贝回家寻子网

国务院总理李克强与张宝艳握手

家寻宝贝 宝贝寻家 流浪乞讨 活动报道 打拐政策 志愿者指南

寻找1986年出生1996年失踪上 2013-08-31

寻找1990年出生1992年送养山 2020-11-08

更多

爱心企业 举报信箱 关于我们

特别鸣谢

爱心风向标 已有357059位爱心人士 爱心凝聚力量 奉献成就团圆

特别鸣谢

“宝贝回家”网站是隶属于宝贝回家志愿者协会的公益性寻人网站。它为失踪儿童、家长提供免费的寻人帮助。因为与公安机关进行合作，其社会认可度较高。

“宝贝回家”网站提供了许多栏目信息。其中“宝贝寻家”栏目是对于长大后发现自己是拐卖的孩子们设计的。他们可以在此登记自己的基本身份信息、失踪地点、现到达的地点等内容。这样便于网友和志愿者协助其找到自己的亲生父母。

接下来展示如何获取到上述信息，进行关联度分析，挖掘数据背后的表现。

## 4.3

# “宝贝在哪儿”关联挖掘



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	寻亲类别	寻亲编号	姓名	性别	出生日期	失踪时身高	失踪时间	失踪人所	Unnamed	失踪地点	Unnamed	寻亲者特征	其他资料	注册时间	跟进志愿者
2	宝贝寻家	295125	叶玲	女	1928年1月1日	未知	1928年11月1日	广西壮族	钦州市	广东省	广州市	不祥	我奶奶是	2017/10/2	旋转的回忆
3	宝贝寻家	295130	王雪莲	女	1988年10月17日	15厘米左右	1988年10月17日	河北省	承德市	河北省	承德市	亲生父母	没有	2017/10/2	卫兵
4	宝贝寻家	296516	耿晨博	男	2013年3月3日	50厘米左右	2013年3月3日	河南省	洛阳市	河南省	洛阳市	男孩，捡到时被花	2017/11/8	章儿	
5	宝贝寻家	294238	徐红苗	女	1993年1月1日	30厘米左右	1993年3月1日	河南省	安阳市	广西壮族	柳州市	脚上有个小胎记	2017/10/2	闲云清烟	
6	宝贝寻家	294363	马洪伟	男	1987年2月9日	未知	1988年2月9日	河北省	沧州市	河北省	沧州市	小的时候胖乎乎的。	2017/10/2	漂泊	
7	宝贝寻家	294386	江志敏	女	1971年7月7日	50厘米左右	1977年8月20日	河南省	南阳市	浙江省	嘉兴市	单眼皮，工具养父回	2017/10/2	追逐梦想	
8	宝贝寻家	294392	唐春艳	女	1989年3月5日	未知	1989年3月5日	重庆市	重庆市	四川省	南充市	我是三岁半左右被拖	2017/10/2	小河小鱼	
9	宝贝寻家	294418	时敏	女	1982年10月15日	未知	1982年10月24日	安徽省	淮北市	江苏省	南京市	听说是在南京福利院	2017/10/2	天高云淡	
0	宝贝寻家	294542	郑女士	女	1984年2月1日	167厘米左右	1985年10月25日	上海市	上海市	山东省	临沂市	手臂有种过水痘留下	2017/10/2	点赞	
1	宝贝寻家	294625	张燕	女	1981年10月25日	未知	1989年10月25日	安徽省	淮南市	四川省	成都市	无	2017/10/2	放飞心情	
2	宝贝寻家	48050	不知道姓	女	1984年11月13日	未知	1984年12月10日	湖北省	孝感市	湖北省	孝感市	无	(湖北-孝	2012/5/26	湖北-孝感
3	宝贝寻家	296813	刘新	男	1997年8月13日	50厘米左右	1997年10月27日	河北省	廊坊市	河北省	廊坊市	我就听我	2017/11/1	小河小鱼	
4	宝贝寻家	47439	袁晓亮	男	1975年12月11日	65厘米左右	1976年3月12日	福建省	福州市	上海市	上海市	我听我养	2012/5/24	小秀才	
5	宝贝寻家	296506	李沫	女	1996年4月22日	未知	1997年1月1日	福建省	福州市	福建省	南平市	不明确，	2017/11/8	若邻郎	
6	宝贝寻家	296848	姚慧娟	女	1993年3月18日	80厘米左右	2003年1月1日	河南省	周口市	河南省	周口市	我是呗爸	2017/11/1	惠琴	
7	宝贝寻家	296556	张建伟	男	1989年6月28日	未知	1991年1月1日	福建省	厦门市	福建省	莆田市	听爸妈说，据家里人	2017/11/8	淡雅宁静	
8	宝贝寻家	296557	宇	女	1971年8月8日	153厘米左右	1971年8月20日	四川省	成都市	四川省	成都市	养家说是家里子女多	2017/11/8	琳闵	
9	宝贝寻家	296584	小涵	女	2000年10月4日	49厘米左右	2000年11月12日	广东省	惠州市	广东省	惠州市	养女17岁，本人当年	2017/11/9	乔峰	
0	宝贝寻家	296599	李加强	男	1991年7月21日	100厘米左右	1992年1月9日	安徽省	合肥市	福建省	莆田市	手脚趾头比较短	2017/11/9	网事如烟	
1	宝贝寻家	296631	李木春	男	1994年10月13日	140厘米左右	2003年10月6日	湖北省	十堰市	河南省	驻马店市	寻亲人系新疆乌鲁木	2017/11/9	追逐梦想	
2	宝贝寻家	296672	王女士	女	1992年6月22日	45厘米左右	1992年6月22日	江苏省	盐城市	江苏省	盐城市	臀部左边有硬币大小	2017/11/9	西岭熊猫	
3	宝贝寻家	296674	郑灵芝	女	1983年6月19日	未知	1983年6月19日	河南省	南阳市	河南省	南阳市	寻找生父生母本人，	2017/11/9	张爱莉	
4	宝贝寻家	296767	张丽	女	1984年5月19日	未知	1984年5月19日	安徽省	滁州市	安徽省	蚌埠市	未知	安徽省蚌	2017/11/1	诺言
5	宝贝寻家	296784	张兆存	男	1989年6月28日	未知	1989年7月4日	河南省	濮阳市	河南省	濮阳市	右眼左上方有颗痣，	2017/11/1	闲云清烟	
6	宝贝寻家	296785	田莉莉	女	1984年3月21日	未知	1984年3月21日	湖北省	荆门市	湖北省	武汉市	养父	2017/11/1	闲云清烟	
7	宝贝寻家	296793	耿富荣	女	1964年5月16日	未知	1964年5月16日	山东省	济南市	内蒙古自治区	包头市	左脚第四、五脚趾有	2017/11/1	情缘	
8	宝贝寻家	296811	付文明	男	1989年6月10日	50厘米左右	1989年6月12日	山东省	临沂市	山西省	太原市	那个时候	2017/11/1	守护你一生	
9	宝贝寻家	296447	林秋莎	女	1987年6月28日	未知	1987年6月28日	福建省	莆田市	福建省	莆田市	不知道	不知道	2017/11/8	梦恩

使用selenium + request + BeautifulSoup 这些常用的python爬虫库进行网页数据爬取

对于每条寻亲信息，提取到性别，出生日期，失踪时身高，失踪时间，失踪人所在地，失踪地点这些关键信息

最终将信息保存为CSV文件

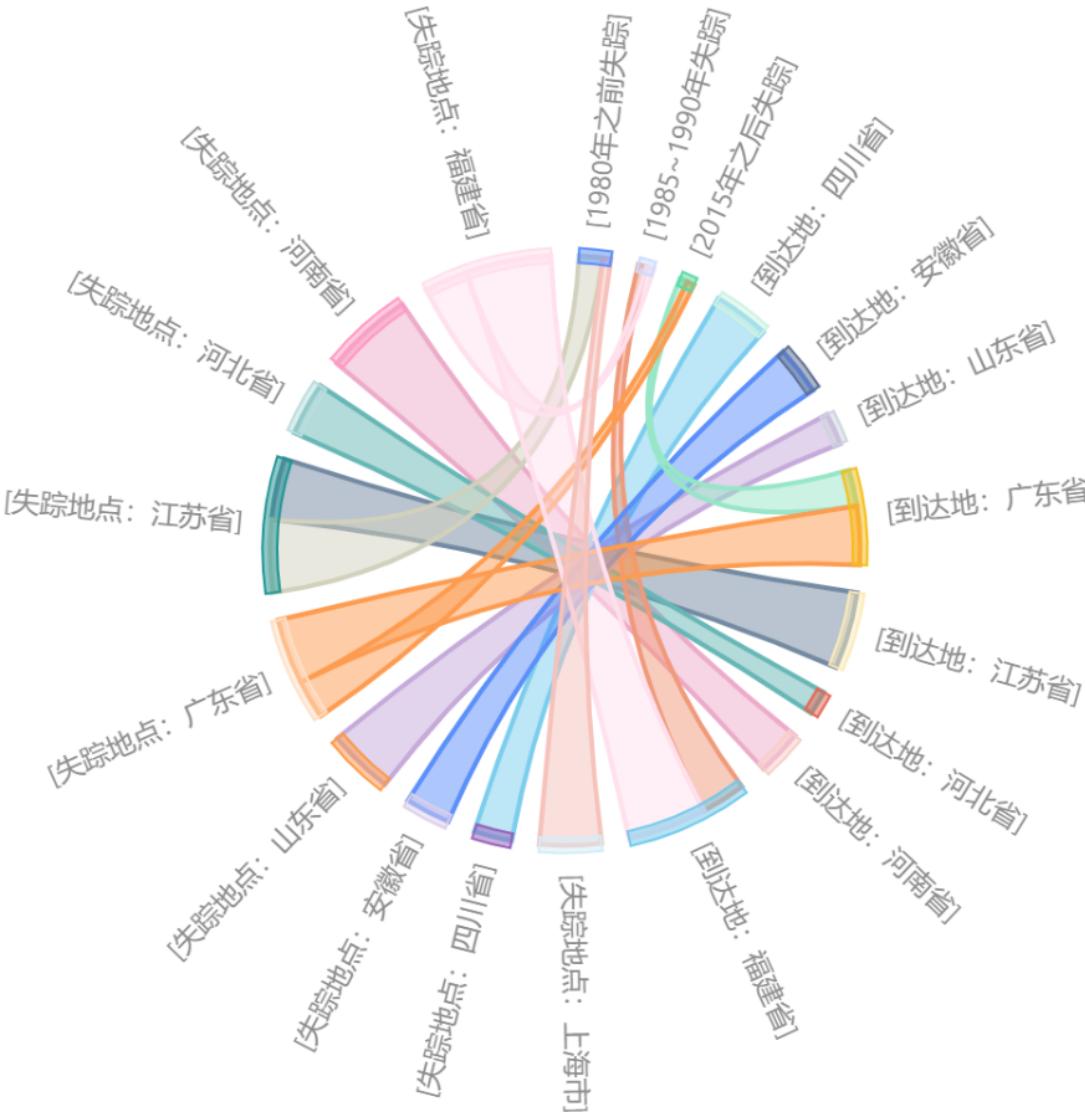
# 4.3 | “宝贝在哪儿” 关联挖掘

## 剔除的规则

- 地区描述中是否有[‘市’，‘自治州’，‘自治区’，‘县’，‘盟’]这几个区域。没有的话可以剔除
- 丢失日期早于出生日期的应该剔除
- 身高出现异常值的应该剔除
- 根据失踪地点描述，提取出所在“县”和“市”的两层行政区划

	H	I	J	K	L
时间	失踪人所在地	Unnamed	失踪地点	Unnamed	寻
性	广西壮族自治区	钦州市	广东省	广州市	不
女	河北省	承德市	河北省	承德市	亲
i女	河南省	洛阳市	河南省	洛阳市	男
男	河南省	安阳市	广西壮族自治区	柳州市	脚
男	河北省	沧州市	河北省	沧州市	小
男	河南省	南阳市	浙江省	嘉兴市	单
男	重庆市	重庆市	四川省	南充市	我
男	安徽省	淮北市	江苏省	南京市	听
男	上海市	上海市	山东省	临沂市	手
男	安徽省	淮南市	四川省	成都市	无
	湖北省	孝感市	湖北省	孝感市	无
	河北省	廊坊市	河北省	廊坊市	我
	福建省	福州市	上海市	上海市	
	福建省	福州市	福建省	南平市	不
	河南省	周口市	河南省	周口市	我
	福建省	厦门市	福建省	莆田市	左

# “宝贝在哪儿”关联挖掘

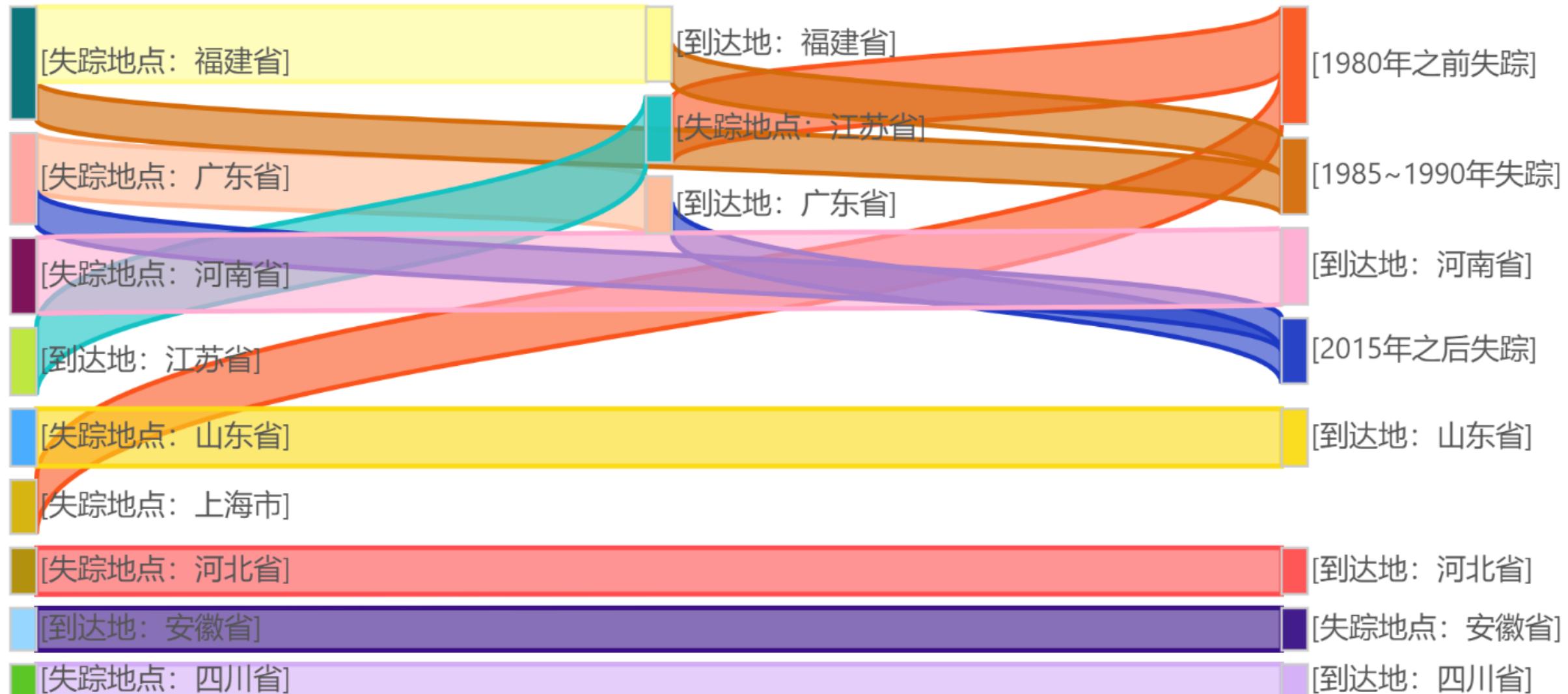


频繁项支持度设置为0.05  
置信度设置为50%

可以得到以下的结论

- 福建，河南，河北，广东，山东，四川，江苏，安徽这几个省的省内转移非常严重，大部分被拐卖的孩子最终都流向了本省
- 福建，江苏省的失踪儿童在1990年之前较多，在那之后较少。推断当时可能存在一些在当地的犯罪集团。
- 广东省在2015年之后省内转移的失踪儿童较多，可能也存在犯罪团伙或别的因素。

# “宝贝在哪儿”关联挖掘





# 主要内容



- 1 机器学习发展与应用
- 2 相关分析和显著性
- 3 机器学习的基本任务
- 4 “宝贝在哪儿” 关联挖掘
- 5 公益课题 “宝贝在哪儿”
- 6 疫情风险分析



## • 研究背景

拐卖妇女儿童罪是指以出卖为目的，拐骗、绑架、收买、贩卖、施诈、接送、中转妇女、儿童的行为。（《刑法》，第二百四十条）

收入差距、失业率、贫困程度会对犯罪行为的发生有较大影响。

教育程度对犯罪情况的影响是相当大的。

## ■ 研究数据

登记信息

寻亲类别: 宝贝寻家

寻亲编号: 287501

姓 名: 陈志伟

性 别: 男

出生日期: 1999年06月22日

失踪时身高: 80厘米左右

失踪时间: 2002年08月08日

失踪人所在地: 河北省,沧州市

失踪地点: 云南省,昭通市,

寻亲者特征描述: 偏瘦身高很矮

其他资料:

注册时间: 2017/9/7 17:40:46

跟进志愿者: 男人如山

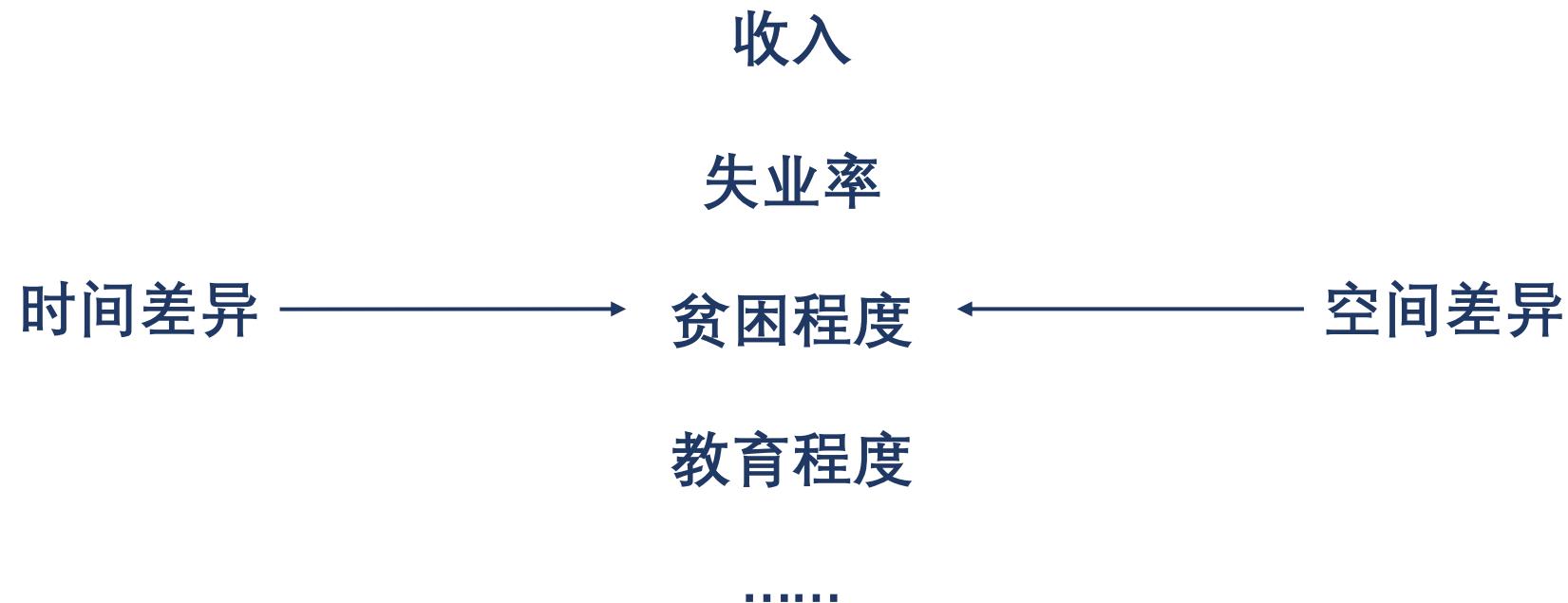
辅助性特征

时间特征

空间特征

图1 宝贝寻家资料卡  
(图片来源: 宝贝回家寻子网 <http://www.baobeihuijia.com>)

## ■ 研究内容



## ■ 省级行政区尺度分析

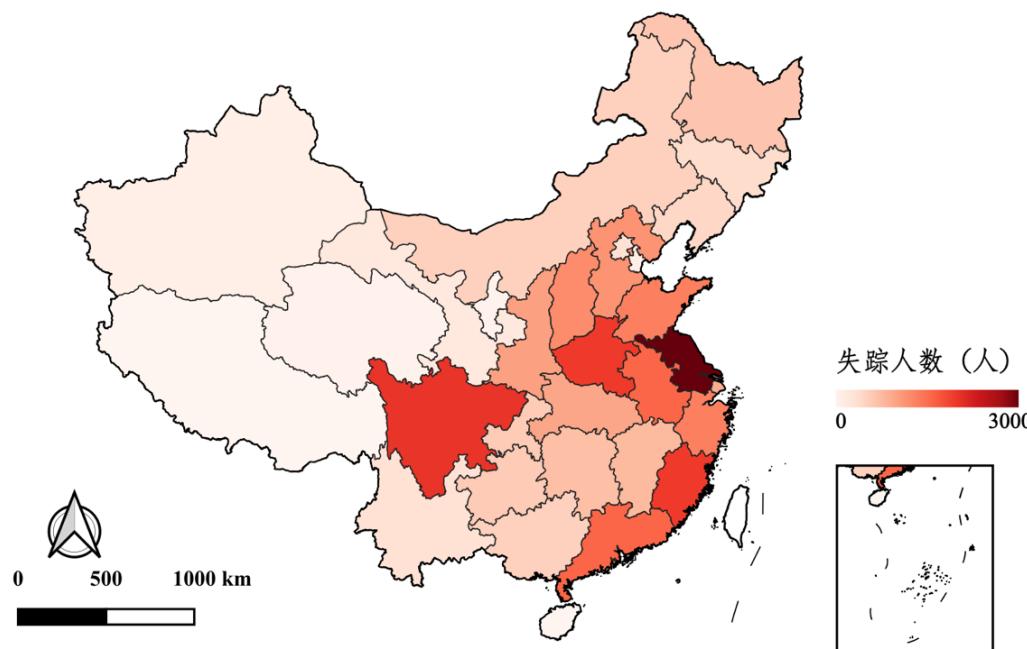


图2 各省失踪人口示意图

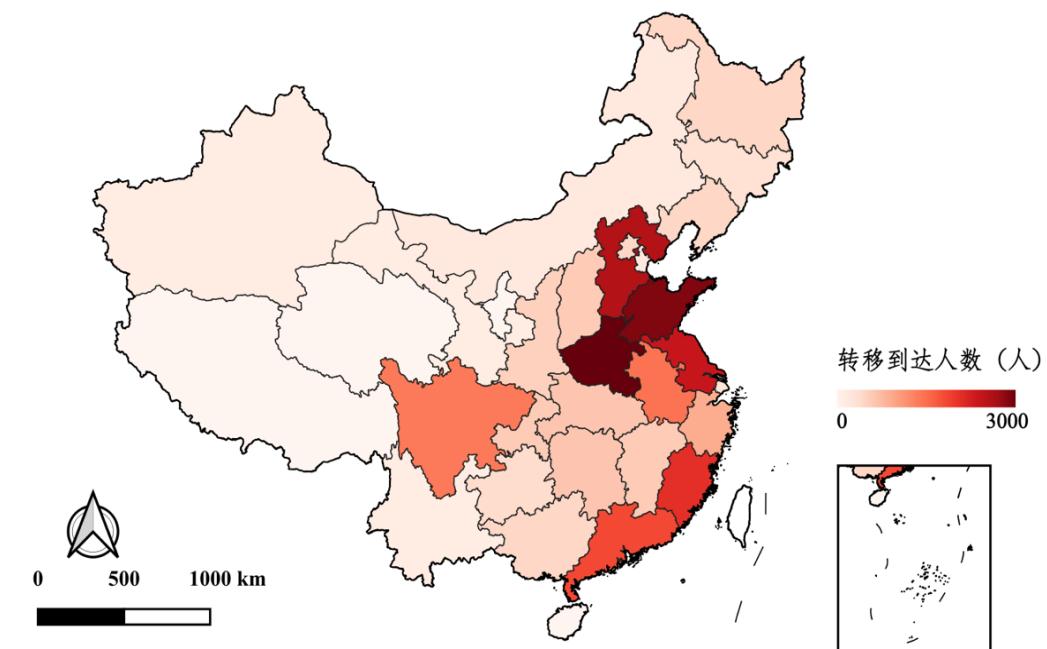


图3 转移到达各省人口示意图

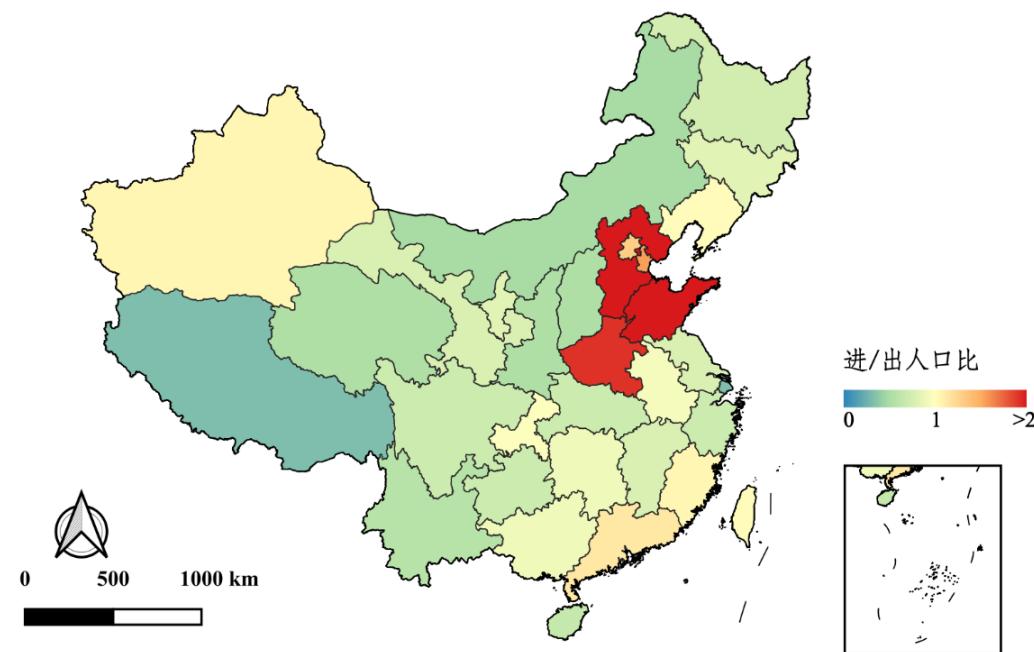


图4 各省进出失踪人口比示意图

## ■ 地级行政区尺度分析

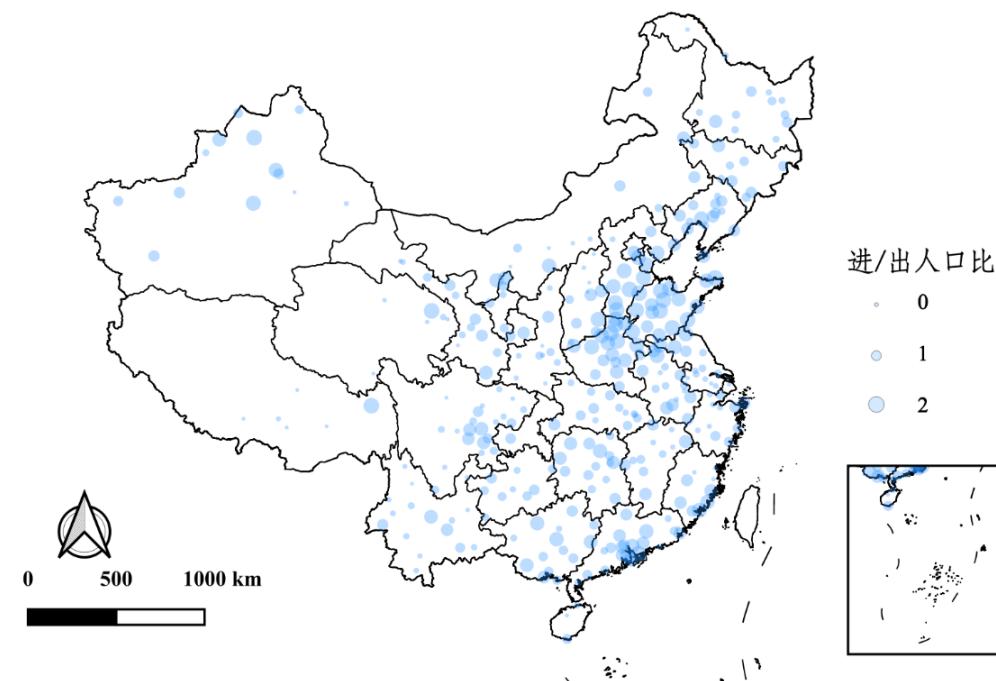


图5 各地市人口进出比示意图

## ■ 转移路径分析



图6 失踪人口转移路径专题图

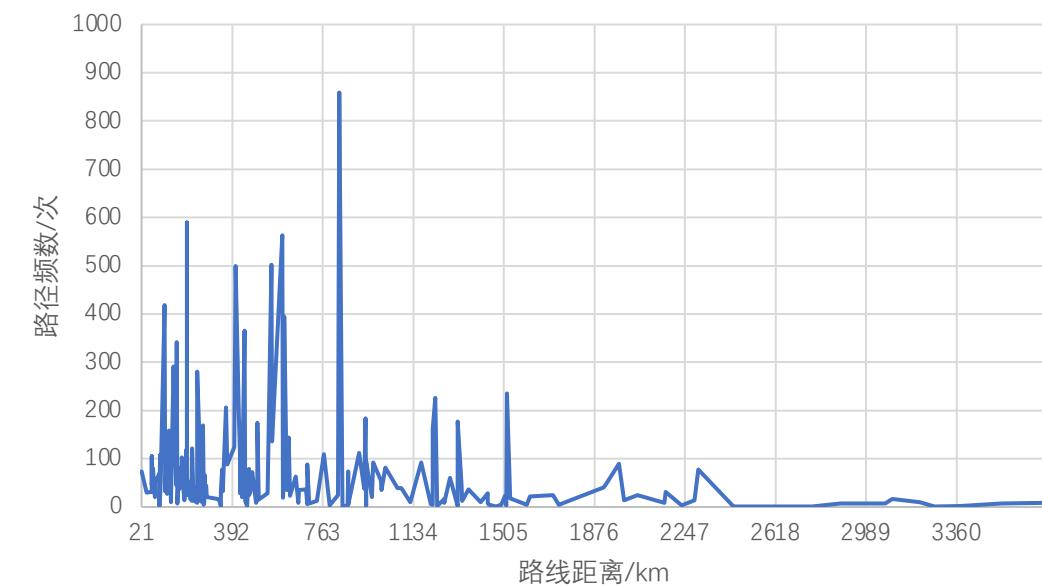


图7 转移路径距离分布折线图

## ■ 年份尺度分析

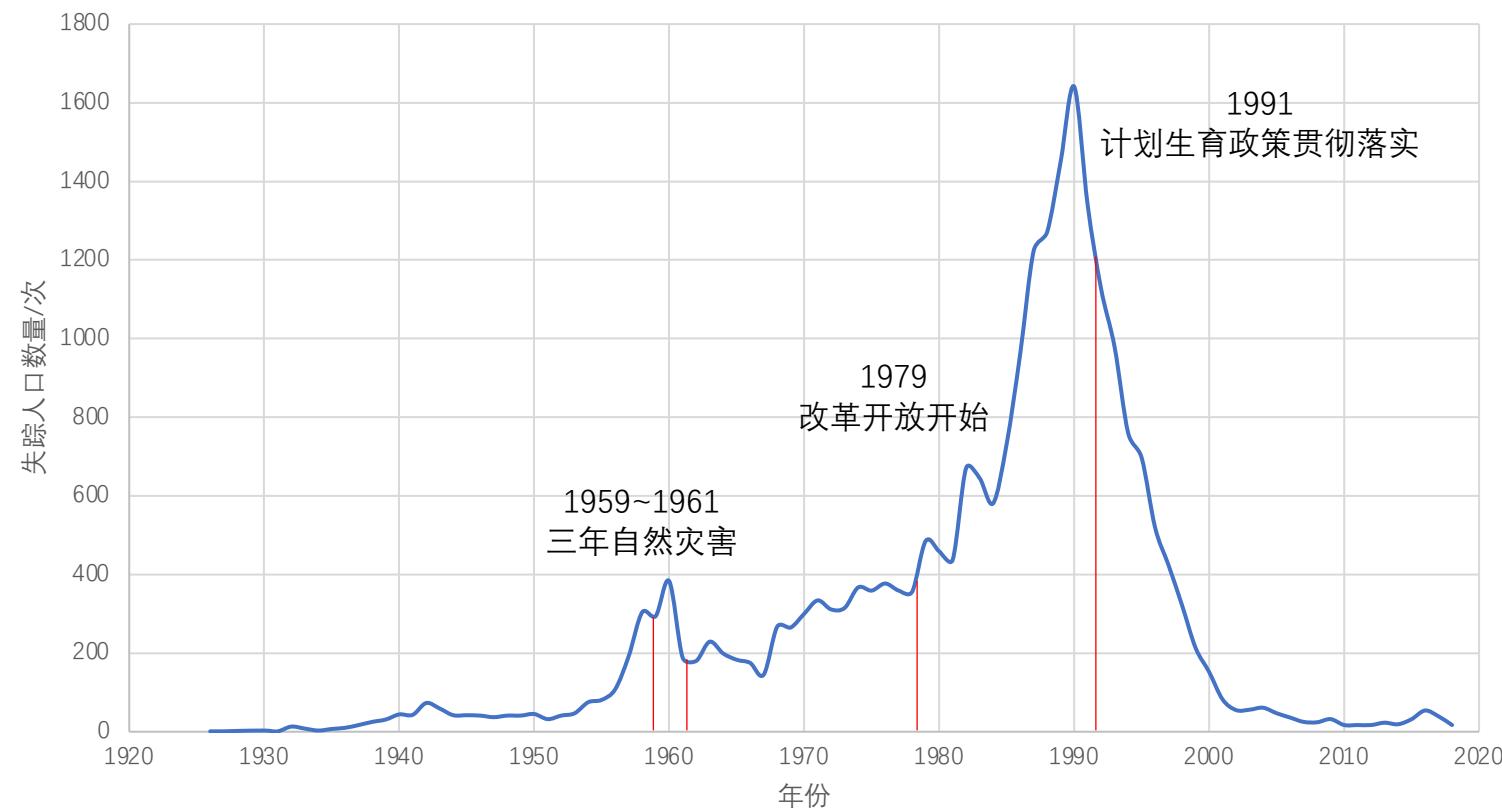


图8 各年失踪人数分布折线图

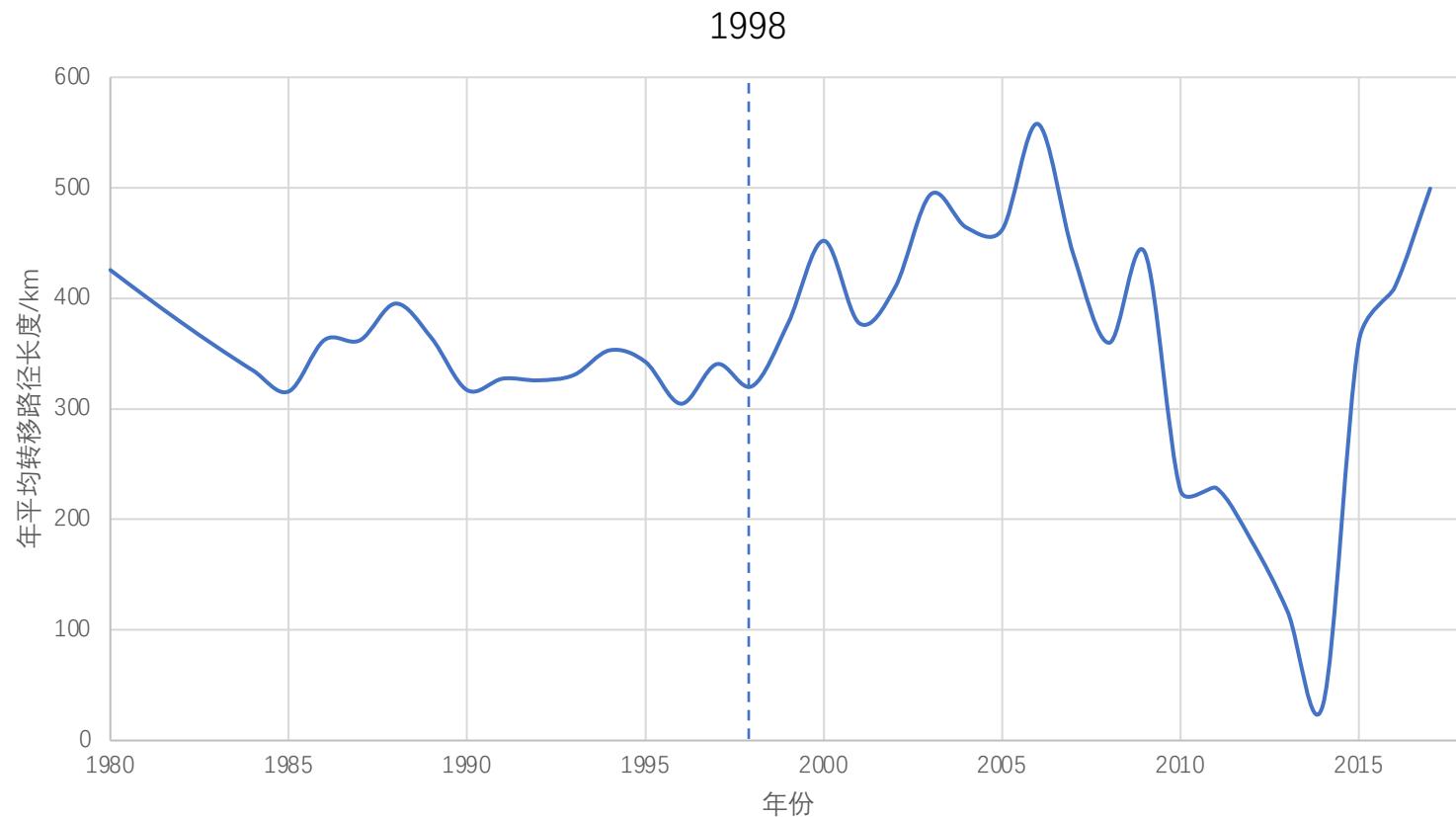


图9 年份相关的转移路径长度均值分布折线图

## ■ 月份尺度分析

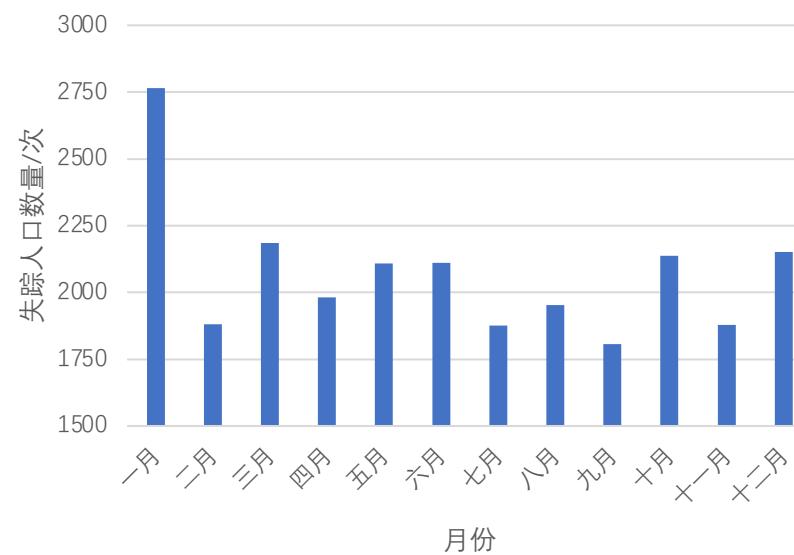


图10 各月失踪人数分布柱状图

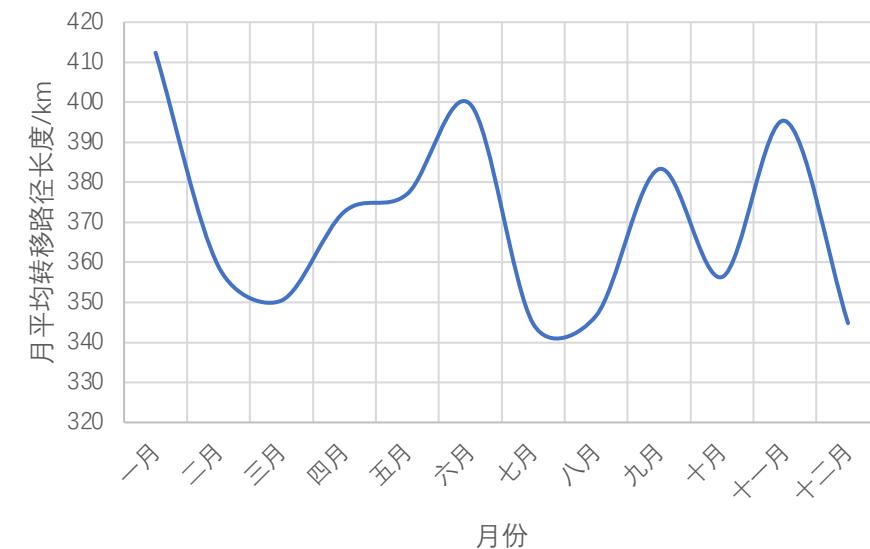


图11 月份相关的转移路径长度均值分布折线图

## ■ 预测实现机制

在何处失踪

何时失踪

失踪时年龄

失踪者身高

失踪者性别

.....





## ■ 独立模型构造

{年龄,性别,身高,失踪省份,失踪年份,失踪月份,失踪区块} -> {到达省份}

{年龄,性别,身高,失踪省份,失踪年份,失踪月份,失踪区块} -> {转移相对距离}

{年龄,性别,身高,失踪省份,失踪年份,失踪月份,失踪区块} -> {转移相对方位}

{年龄,性别,身高,失踪省份,失踪年份,失踪月份,失踪区块} -> {到达区块}

## ■ 独立模型构造

{年龄,性别,身高,失踪}

{年龄,性别,身高,失踪}

{年龄,性别,身高,失踪}

{年龄,性别,身高,失踪}

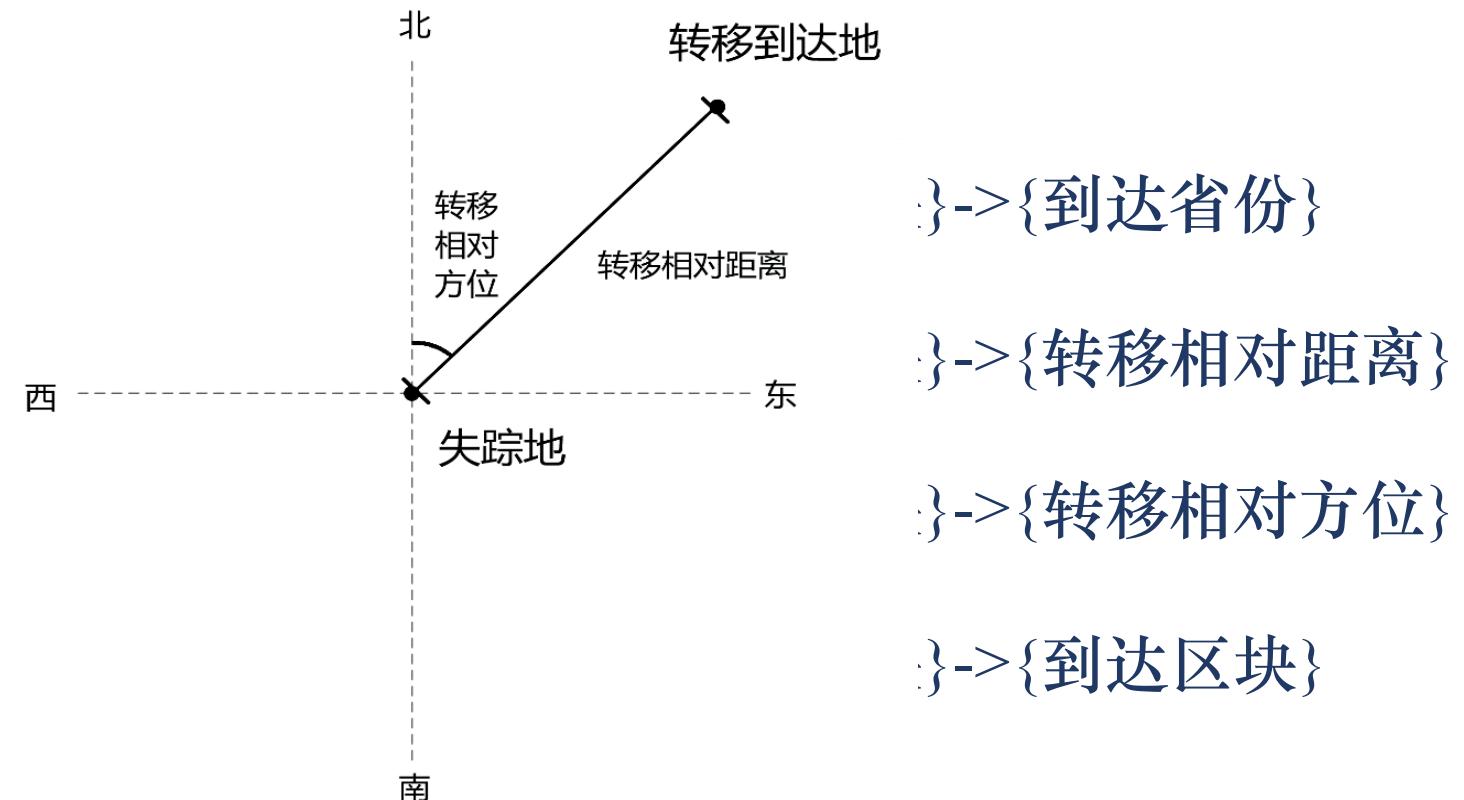


图13 转移相对距离与转移相对方位

## ■ 区块划分

{年龄,性别,身

{年龄,性别,身

{年龄,性别,身

{年龄,性别,身

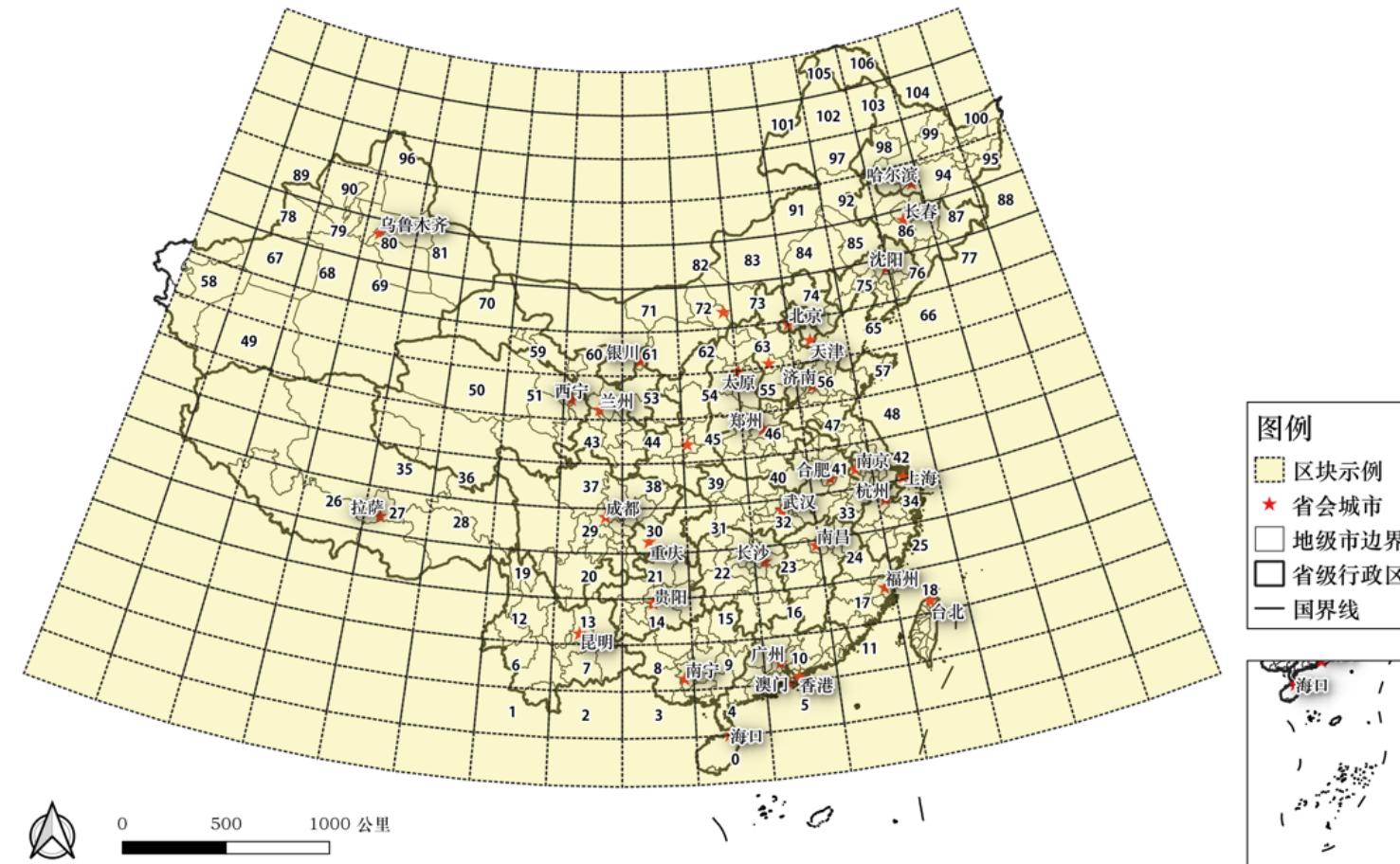


图14 区块划分示意图

## ■ 独立模型构造

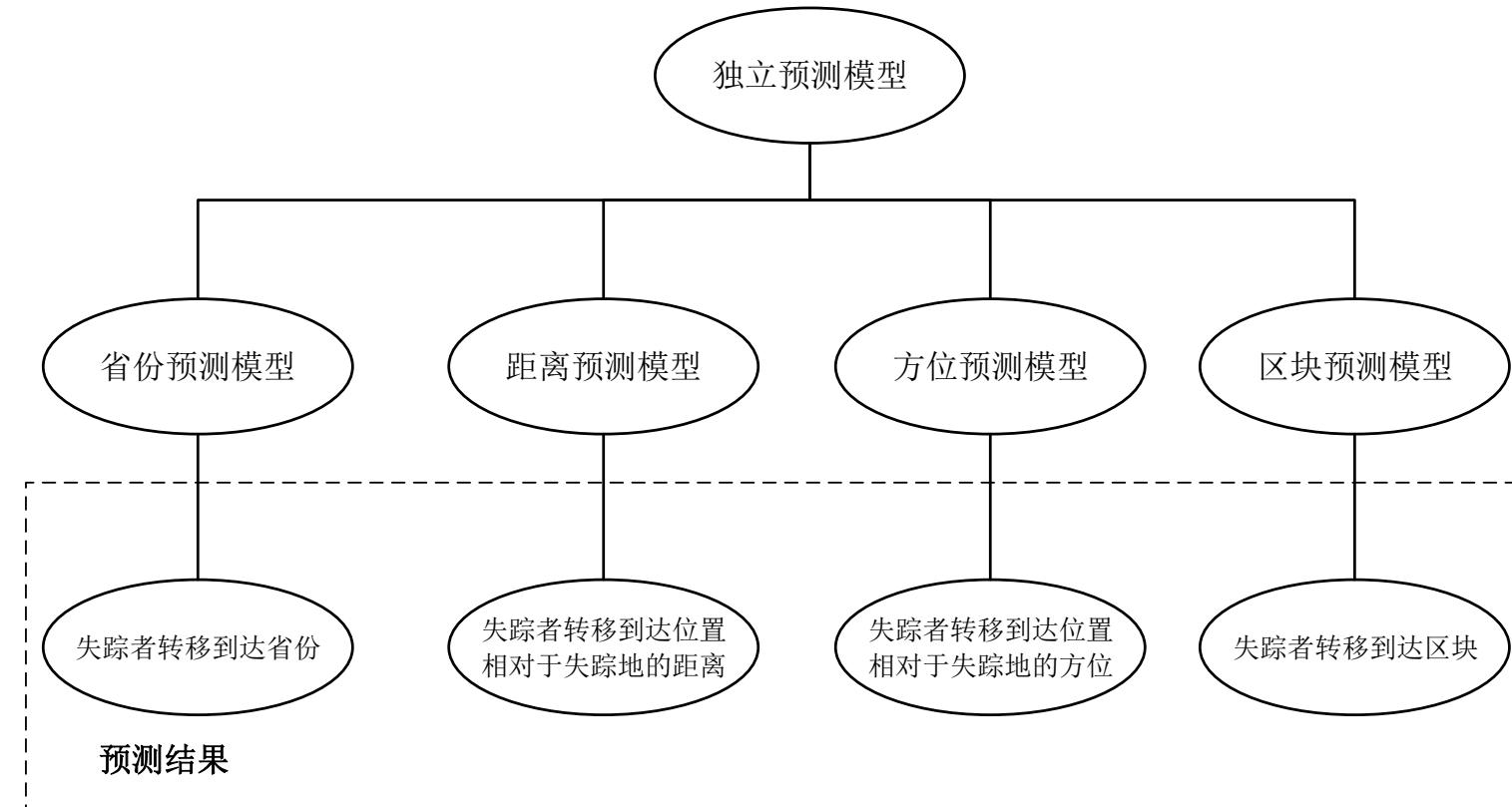


图15 独立预测模型

## ■ 综合预测算法

$$Freq_{region_i} = \frac{Qty_{region_i}}{\sum_{i=1}^n Qty_{region_i}}$$

$$Pr_{region} = Pr_{dire} \cdot Pr_{dist}$$

$$Pr_{city_{prov_i}} = Pr_{prov} \cdot Freq_{prov_i} \cdot OA_{prov}$$

$$Pr_{city} = \frac{Pr_{city_{region}} \cdot Kappa_{region} + Pr_{city_{prov}} \cdot Kappa_{prov} + Pr_{city_{block}} \cdot Kappa_{block}}{Kappa_{region} + Kappa_{prov} + Kappa_{block}}$$

## ■ 误差检验

表1 各独立预测模型的精度检验

模型	总体精度	Kappa 系数
省份	0.800	0.786
区块	0.757	0.745
距离	0.639	0.581
方位	0.656	0.608

表2 综合预测算法结果集的精度评价

评价标准	召回率
第一召回	39.50%
前 3 内召回	63.56%
前 5 内召回	73.02%
前 10 内召回	82.99%

## ■ 转移去向驱动因素分析

表3 各独立预测模型参数权重

模型	性别	年龄	身高	失踪年份	失踪月份	失踪省份	失踪区块
省份	0.107	0.115	0.091	0.110	0.106	0.247	0.224
区块	0.115	0.125	0.099	0.120	0.114	0.194	0.233
距离	0.129	0.150	0.118	0.142	0.136	0.163	0.161
方位	0.126	0.147	0.116	0.141	0.133	0.160	0.176



## ■ 中国失踪人口时空预测服务平台



### 预测简报

城市级别的综合预测结果显示，失踪者可能到达广州市、深圳市、郑州市、东莞市、新乡市、安阳市、周口市、洛阳市、开封市、南阳市等城市。具体可以参考独立预测模型结果。

预测内容仅供参考，衷心祝福家庭早日团聚！

详细信息

重新预测

5.2

# 失踪人口转移去向预测



The screenshot displays the China Missing Persons Space-Time Prediction Platform. At the top, there is a navigation bar with links to '使用说明' (Usage Instructions), '实现原理' (Implementation Principle), and '关于' (About). The main content area includes:

- 预测简报** (Prediction Summary): A large text box stating that city-level comprehensive prediction results show that the missing person may arrive in Wuhan, Xiangyang, Huanggang, Yichang, Ezhou, Jingzhou, and other cities.
- 城市级别的综合预测结果显示，失踪者可能到达武汉市、孝感市、黄冈市、襄阳市、黄石市、九江市、咸宁市、十堰市、岳阳市、荆州市等城市。具体可以参考独立预测模型结果。**
- 预测内容仅供参考，衷心祝福家庭早日团聚！**
- 详细信息** and **重新预测** buttons.
- Copyright © 2019 Liu Yifei and Dr. Yao Yao**
- 失踪人员基本信息**: Gender: Male, Height: 100 cm, Birth Date: 2002-10-30, Disappearance Date: 2007-10-01, Disappearance Location: East longitude 113.8127, North latitude 22.8195, Description: Located at No. 210, Dezheng Middle Road, Dongguan City, Guangdong Province. Prediction Time: 2019-4-18 20:36:19.
- 失踪人口去向预测报告**: A map of Chongqing showing predicted arrival locations with probability values (e.g., 0.01, 0.02, 0.82, 0.12).
- 预测简报**: A summary statement about the comprehensive prediction results for other cities.
- 独立模型预测结果**: A pie chart showing the probability of arrival by province: Guangdong (0.93), Jiangxi (0.01), Hunan (0.02), and Henan (0.04).
- 去向省份预测模型**: A line graph showing the cumulative distance distribution of predicted arrival provinces.
- Copyright © 2019 Liu Yifei and Dr. Yao Yao**



# 主要内容



- 1 机器学习发展与应用
- 2 相关分析和显著性
- 3 机器学习的基本任务
- 4 “宝贝在哪儿” 关联挖掘
- 5 公益课题 “宝贝在哪儿”
- 6 疫情风险分析



# 本章总结

本章解释了机器学习的**基本概念**，并介绍了机器学习的**发展历程与应用现状**。在未来，机器学习将成为日常生活中不可或缺的部分。

**相关性分析**是指对两个或多个具备相关性的变量进行分析，本章介绍了地理要素间相关关系的种类、地理相关程度的测度方法和相关系数的显著性检验。

**回归、分类、聚类**是机器学习的基本任务。**一元线性回归、多元线性回归**等是回归分析的重要方式；**Logistic回归、决策树、支持向量机**是分类常用的基本手段；聚类算法有多种聚类方式包括**原型聚类、密度聚类**还有**层次聚类**等。

**关联分析**是从大量数据集中挖掘数据中的关联性或相关性，**Apriori**算法是最著名的关联规则挖掘算法之一。关联分析是数据挖掘的重要手段。

机器学习技术可与多源数据耦合，在**社会公益**和**流行病防控**等方面有广泛的应用前景。



# Discussion !

---

**姚尧** 博士，副教授，高级工程师

地理与信息工程学院，地图制图学与地理信息工程

阿里巴巴集团，达摩院，访问学者

Email: [yaoy@cug.edu.cn](mailto:yaoy@cug.edu.cn)

办公地点：未来城校区地信楼522办公室

