信息分析聚类

May 5, 2015

Outline

- 贝叶斯定理与信息评估框架
- 聚类
- 结果

贝叶斯定理与信息评估框架

- 频率学派
- 贝叶斯学派

频率学派对概率的认识

从“自然”角度出发，试图直接为“事件”本身建模，即事件A在独立重复试验中发生的频率趋于极限P，那么P即为该事件的概率

贝叶斯学派对概率的认识

从“观察者”角度出发，认为不确定性来自于观察者知识的不完备。随机性并不源于事件本身是否发生，而只是描述观察者对该事件的知识状态。观察者又试图通过已经观察到的“证据”来推断这一事件的结果，因此只能靠猜，而概率论则是用来描述理性推断过程的数学语言

贝叶斯定理与信息评估框架

贝叶斯理论使用贝叶斯定理作为根据新的信息导出或者更新现有的置信度 的规则。在形式上，它描述了随机事件A和B的条件概率（$P(A|B)$,$P(B|A)$）与非条件概率（$P(A)$,$P(B)$）的关系：

$$P(A|B) = \frac{P(B|A)}{P(B)} \times P(A) \tag{1}$$

在贝叶斯定理中，上式每项都有约定俗成的名称：

Table: 贝叶斯定理各项意义

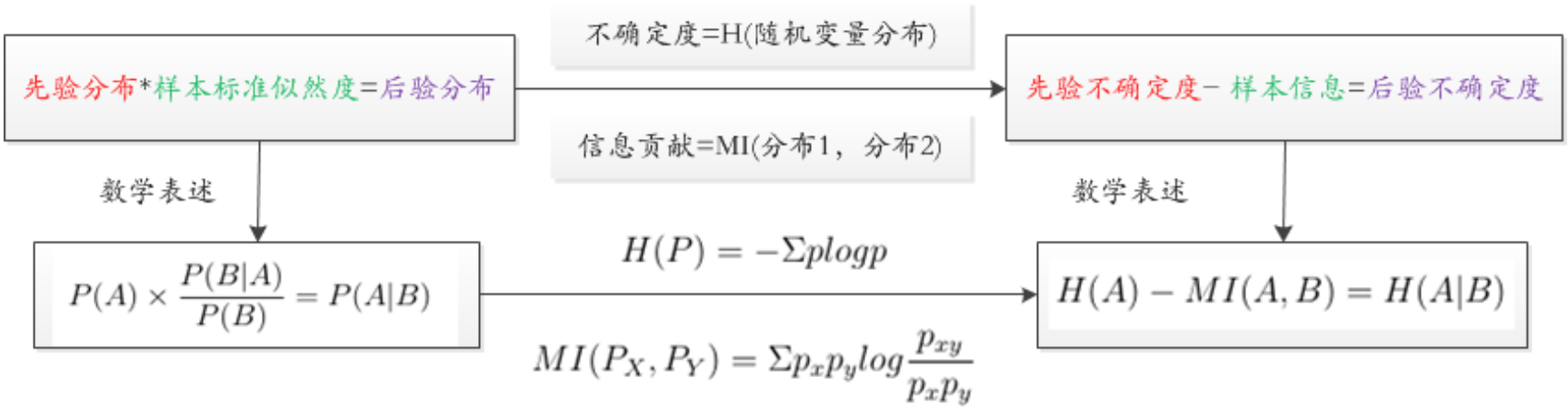
项	名称
$P(A)$	A的先验概率或边缘概率
$P(A\|B)$	已知B的取值，A的后验概率
$P(B\|A)$	已知A的取值，B的后验概率 / 相似度
$P(B)$	B先验概率 / 标准化常量（normalizing constant）
$P(B\|A)/P(B)$	标准相似度（standardised likelihood）

按这些术语，贝叶斯定理可表述为：

$$\text{后验概率} = \frac{\text{相似度}}{\text{标准化常量}} \times \text{先验概率} \tag{2}$$

或

$$\text{后验概率} = \text{标准相似度} \times \text{先验概率} \tag{3}$$


先验分布*样本标准似然度=后验分布
不确定度=H(随机变量分布)
先验不确定度− 样本信息=后验不确定度
信息贡献=MI(分布1，分布2)
数学表述
数学表述
$P(A) \times \frac{P(B|A)}{P(B)} = P(A|B)$
$H(P) = -\Sigma plogp$
$H(A) - MI(A,B) = H(A|B)$
$MI(P_X, P_Y) = \Sigma p_x p_y log \frac{p_{xy}}{p_x p_y}$

1,对方程 1两边取对数:

$$logP(A|B) = logP(A) + log\frac{P(AB)}{P(A)P(B)} \tag{4}$$

2, 方程两边各项乘以 $-P(A,B)$:

$$-P(A,B)logP(A|B) = -P(A,B)logP(A) - P(A,B)log\frac{P(AB)}{P(A)P(B)} \tag{5}$$

3,对方程两边各项在概率空间内求和或取积分

$$-\sum_A\sum_B P(A,B)logP(A|B) = -\sum_A\sum_B P(A,B)logP(A) - \sum_A\sum_B P(A,B)log\frac{P(AB)}{P(A)P(B)} \tag{6}$$

或

$$-\int\int P(A,B)logP(A|B)dAdB = -\int\int P(A,B)logP(A)dAdB - \int\int P(A,B)log\frac{P(AB)}{P(A)P(B)}dAdB \tag{7}$$

根据定义,方程6可简化为:

$$H(A|B) = H(A) - I(A,B) \tag{8}$$

根据定义,方程7可简化为:

$$h(A|B) = h(A) - I(A,B) \tag{9}$$

根据方程1和 8及方程9等号两边各项的对应关系，可以认为$H(A|B)$表示后验不确定度，$H(A)$表示先验不确定度，$I(A,B)$表示由数据间的信息贡献。
因此，按贝叶斯理论术语，信息熵与互信息关系可表述如下：

$$\text{后验不确定度} = \text{先验不确定度} - \text{数据信息贡献} \tag{10}$$

Table: 估算信息项

类别	估算项
观测	$h(X_o)$
	$I(X_o; Xi_{original}), I(X_o; Xi_{original}, Xi_{new})$
	$I(X_o; Xi_{current}), I(X_o; Xi_{former}, Xi_{current})$
模型	$I(X_o; X_s), I(X_o; S, Xi_{current})$

复杂性不是问题，因为研究对象本来就是复杂的，问题是复杂性带来的迷惑性.

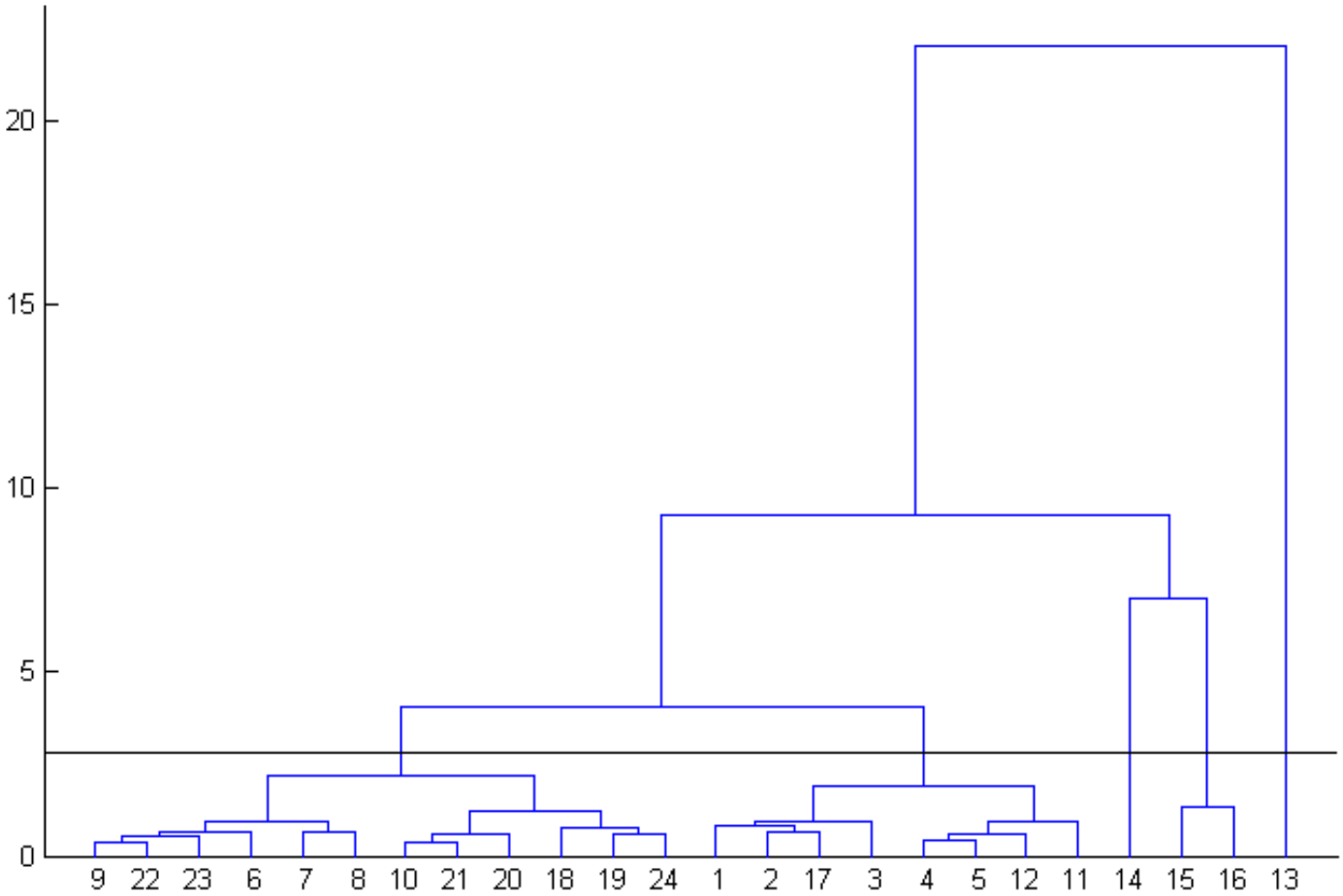
聚类

聚类分析（英语：Cluster analysis，亦称为群集分析）是无监督学习的一种，在许多领域受到广泛应用，包括数据挖掘，模式识别，图像分析以及生物信息。聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集（subset），这样让在同一个子集中的成员对象都有相似的一些属性，常见的包括在坐标系中更加短的空间距离等（wikipedia）.

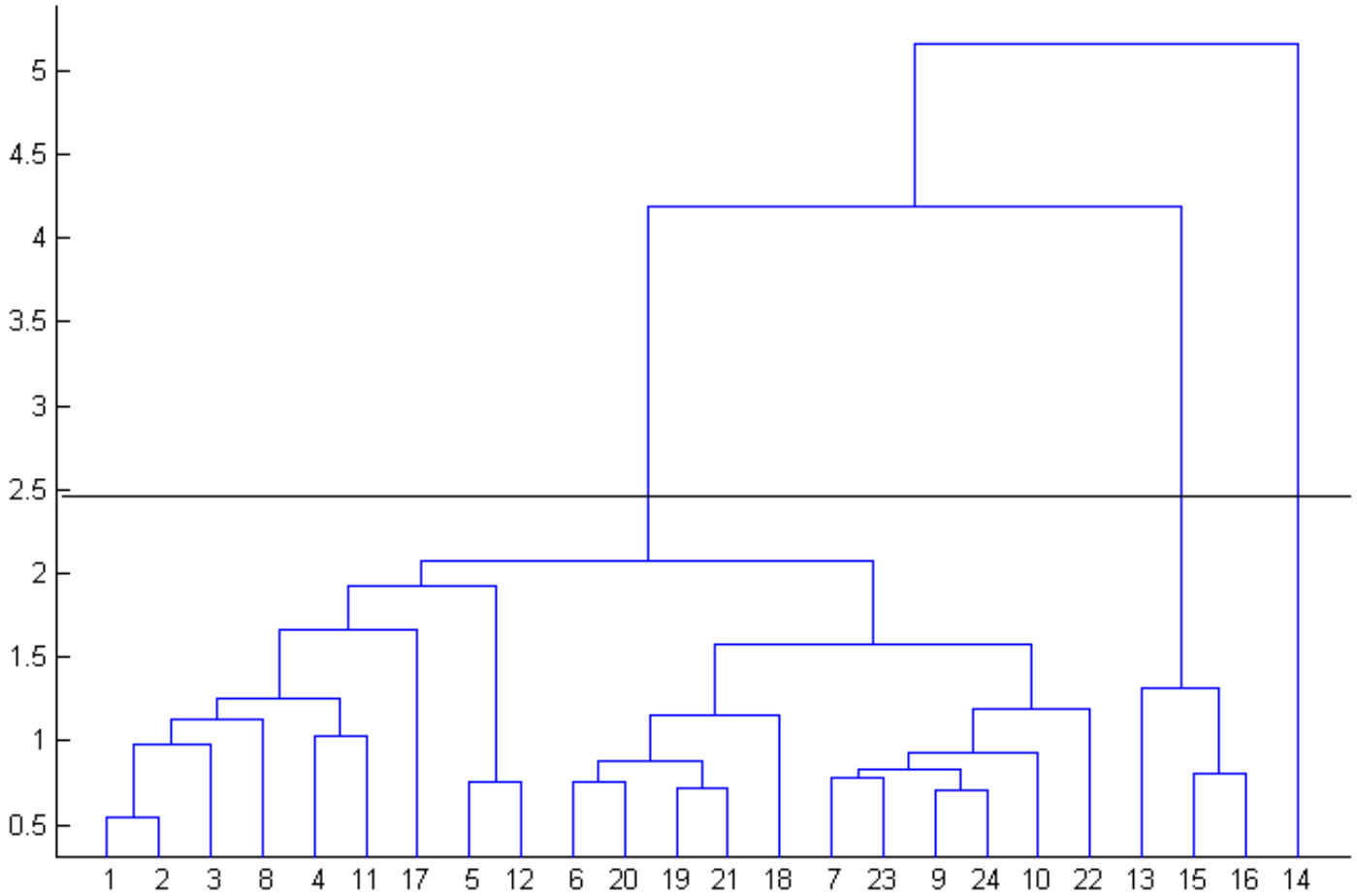
数据根据气候类型预分类

气候类型	编号	面积(km^2)	$P_{mean}(mm)$	$PE_{mean}(mm)$	$R_{mean}(mm)$
	1	215	1299	882	553
	2	611	1252	965	539
WA	3	2953	1321	1101	330
	4	9886	1452	1061	549
	5	6967	1440	1055	489
	6	3349	922	993	232
	7	2901	1001	1066	261
WS	8	9811	1006	959	303
	9	2344	948	1259	221
	10	5227	935	1303	160
	11	4338	1371	976	542
	12	1924	1442	1059	509
	13	290	522	1407	34
SA	14	1577	2748	751	2212
	15	2590	1613	681	1105
	16	3056	1287	775	872
	17	5317	1052	851	510
	18	4022	854	1017	254
	19	8472	839	984	224
	20	8912	794	998	117
SS	21	7268	808	1027	173
	22	1052	941	1110	228
	23	865	950	1186	236
	24	9889	877	1250	187

先验不确定度聚类结果



TPWB认知不确定度聚类结果



Budyko认知不确定度聚类结果