

第1章 概论

- 1.1 模式识别的基本概念
- 1.2 模式识别系统
- 1.3 模式识别的基本方法
- 1.4 模式识别的应用
- 1.5 模式识别的基本问题

1.1 模式识别的基本概念

- 模式识别的定义(Definition)

Pattern recognition is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns.
(Anil K. Jain)

模式识别与机器学习的关系：模式识别~机器学习。两者的主要区别在于前者是从工业界发展起来的，后者则主要源自计算机学科。在著名的《Pattern Recognition And Machine Learning》一书中，Christopher M. Bishop在开头是这样说的“模式识别源自工业界，而机器学习来自于计算机学科。不过，它们中的活动可以被视为同一个领域的两个方面，同时在过去的10年间，它们都有了长足的发展”。

● 机器学习的定义

目前机器学习还没有一个准确、统一的定义。

经典定义：计算机程序如何随着经验的积累自动改善自身的性能

[T.Mitchell, Machine Learning. CMU Book 97]。

Definition: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. (对于某类任务T和性能度量P，若一个计算机程序在T上以P衡量的性能随着经验E而自我完善，那么我们称该计算机在从经验E中学习。)---从学术的角度定义

例：手写文字识别学习问题：

- 任务T：识别和分类图像中的手写体文字
- 任务性能标准P：分类的正确率
- 训练经验E：已知分类的手写文字数据库

“Machine Learning” as an Engineering Paradigm: Use data and examples, instead of expert knowledge, to automatically create systems that perform complex tasks. ---从工程的角度定义

只要是有数据的地方，就会对数据进行分析(当前热门的研究领域：大数据分析)，机器学习就无处不在(Machine Learning Everywhere)，随着该领域的发展，机器学习主要做智能数据分析。机器学习的子类----深度学习可用于大数据分析。

机器学习与相关学科的关系：

- (1) **数据挖掘**：数据挖掘=机器学习+数据库。大部分数据挖掘中的算法是机器学习算法在数据库中的优化。
- (2) **计算机视觉**：计算机视觉=图像处理+机器学习。如百度识图、指纹识别、人脸识别等。
- (3) **语音识别**：语音识别=语音处理+机器学习。语音识别技术一般不会单独使用，一般会结合自然语言处理的相关技术，目前的相关成功应用有苹果语音助手siri等。
- (4) **自然语言处理**：自然语言处理=文本处理+机器学习。自然语言处理技术主要是让机器(即计算机)理解人类语言的一门学科领域。在自然语言处理技术中，大量使用了计算机编译原理相关的技术，例如词法分析、语法分析等，除此之外，在理解层面则使用了语义理解、机器学习等技术。

1.1 模式识别的基本概念

模式(pattern)：具有某种特定性质的观察对象。广义地，存在于时间、空间中可观察的事物，具有时间或空间分布的信息。

观察对象举例：一个数字、一句话、一张照片等都是观察对象，都能成为模式识别中的识别对象。

模式识别(Pattern Recognition)：用计算机实现人对各种事物或现象的分析、描述、判断、识别。

模式类：具有相似特性的模式的集合。

（模式与模式类的关系相当于集合论中的元素与集合的关系）

1.1 模式识别的基本概念

模式识别与图像处理、图像识别的关系：模式识别是模拟人的某些功能。

模拟人的视觉：计算机+图像传感器

模拟人的听觉：计算机+声音传感器

模拟人的嗅觉和触觉：计算机+嗅觉/触觉传感器

模式识别技术当前主要是对视觉和听觉能力的模拟。模拟人的视觉能力就是用计算机来做图像识别和理解工作；模拟人的听觉就是用计算机来做语音（或者各种声音）识别和理解方面的工作。

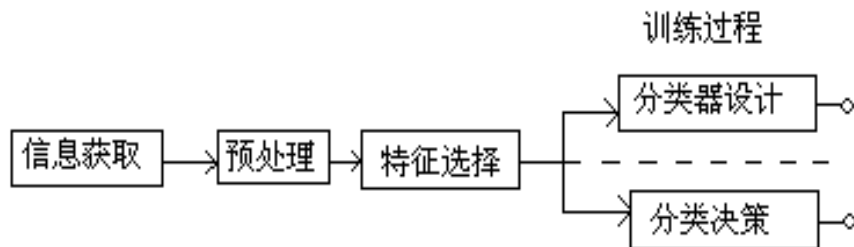
1.1 模式识别的基本概念

模式识别是一种智能活动，包括分析和判断两个过程。

分析过程：确定用于划分模式类的特征及其表达方法；

判断过程：依据待识别对象的特征，将其判属于某一个模式类。

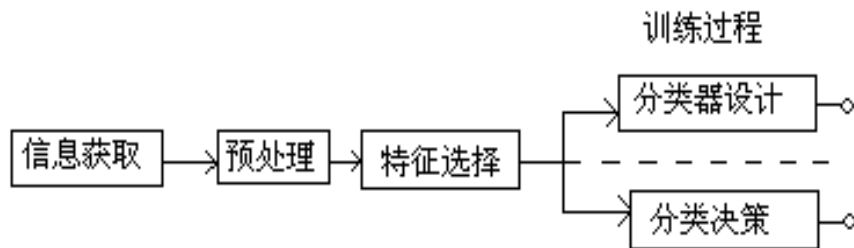
1.2 模式识别系统



统计模式识别系统组成框图

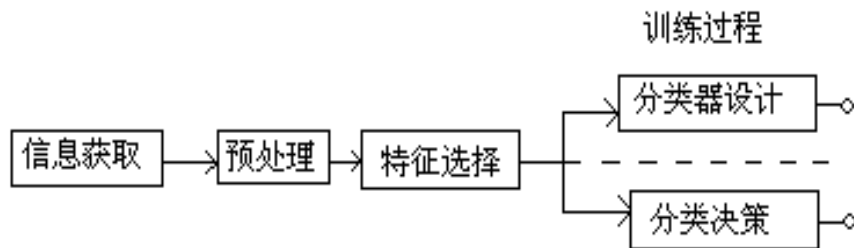
信息获取：对于人脑识别而言，人脑通过感觉器官获取模式信息。对于机器识别来说，由于计算机只能处理数字信号，计算机获取模式信息意味着实现观察对象的数字化表达。信息获取是通过传感器，将光或声音等信息转化为电信息。信息可以是二维的图像如文字、图像等；可以是一维的波形如声波、心电图、脑电图；也可以是物理量与逻辑值。

1.2 模式识别系统



预处理：在得到模式的数字化表达后,往往需要对它进行预处理，以便去除或减少噪声的影响，突出有用信息。对于图像信息，采用数字图像处理技术作为其预处理技术，主要有二值化、图像平滑、变换、增强、恢复、滤波、几何校正等。

1.2 模式识别系统



特征提取和选择：在模式识别中，需要进行特征的抽取和选择，如，一幅 64×64 的灰度图像可以得到4096个数据，这种在测量空间的原始数据通过变换获得在特征空间最能反映分类本质的特征。这就是特征提取和选择的过程。

1.2 模式识别系统

特征是用于描述模式性质(特性)的一种定量的概念,通过对模式的分析得到一组特征,称这个过程为特征形成。

特征一般有两种表达方法: (1)将特征表达为数值; (2)将特征表达为基元。

1.2 模式识别系统

(1) 当将特征表达为数值时, 一个模式的 d 个特征值就构成了一个特征向量, 记为 \mathbf{x} , 即
$$\mathbf{x} = (x_1, x_2, \dots, x_d)$$

其中, \mathbf{x} 的每个分量 $x_i(i=1, 2, \dots, d)$ 对应一个特征。

(2) 当特征表达为基元时, 一个模式表述为一个句子, 记为 x , 即

$$x = x_1 x_2 \cdots x_d$$

其中, $x_i(i=1, 2, \dots, d)$ 为基元, 反映构成模式的基本要素。

1.2 模式识别系统

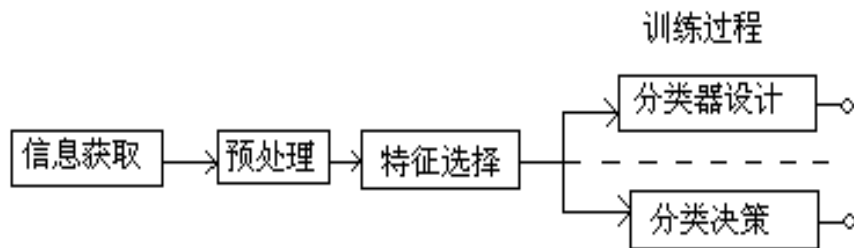
通常用于描述模式性质的特征很多, 需从一组特征中挑选出一些最有效的特征以降低特征空间维数, 即**特征选择**。

特征提取是指采用变换 (或映射) 实现由模式测量空间向特征空间的转变, 或者将特征空间的维数从高维变成低维。

1.2 模式识别系统

举例：通常遥感成像光谱仪波段数达数百个之多，如果直接用原始数据进行地物分类，会因数据量太大而导致计算复杂，且分类效果不一定好，可通过变换或映射的方法，由原始数据空间变换到特征空间，得到最能反映模式本质的特征，同时降低空间维数。

1.2 模式识别系统



分类器包括分类器设计和分类决策两部分。

分类器设计：分类器设计的主要功能是通过训练确定判决规则，使按此类判决规则分类时，错误率最低。把这些判决规则建成标准库。

分类决策：在特征空间中对被识别对象进行分类。

1.2 模式识别系统

说明:

(1)基于机器学习的模式识别系统通常由两个过程组成,即**分类器设计**(简称**设计**)和**分类判决**(简称**实现**)。一般是用一定数量的样本进行分类器设计,这些样本的所属类别已知,称为**训练样本**。实现是用所设计的分类器对待识别模式进行分类判决 (或分类决策)。

1.2 模式识别系统

说明：

(2)模式类是指具有相似特性的模式的集合，模式和模式类的关系就是元素和集合的关系。模式的分类过程，事实上就是判定表征观察对象的元素和指定集合的从属关系的过程。当元素只和某个集合具有从属关系时，就将该对象判属于该集合对应的类；当元素和多个集合具有从属关系时，既可以任选一类进行判决，也可以拒绝判决；当元素和任何一个集合都不具有从属关系时，不作分类判决，即**拒绝判决**。

1.3 模式识别的基本方法

1.模板匹配法

- (1)首先对每个类别建立一个或多个模板；
- (2)输入样本和数据库中每个类别的模板进行比较，求相关或距离；
- (3)根据相关性或距离大小进行决策。

优点：直接、简单

缺点：适应性差

1.3 模式识别的基本方法

2.统计模式识别法

统计模式识别把观察对象表达为一个随机向量(即特征向量),将模式类表达为由有穷或无穷个具有相似数值特性的模式组成的集合。

识别是从模式中提取一组特性的度量,构成特征向量来表示模式,然后通过划分特征空间的方式进行分类。

统计模式识别系统构成:主要由信息获取、预处理、特征提取和选择以及分类器4部分组成;其中,分类器包括分类器设计和分类决策。

优点:理论较成熟,适用于用较少特征就能描述观察对象的场合,能考虑干扰、噪声等的影响

缺点:对于结构复杂模式的特征提取较为困难,不能反映模式的结构特征

1.3 模式识别的基本方法

3.句法模式识别法

- (1)许多复杂的模式可以分解为简单的子模式，这些子模式组成所谓“基元”
- (2)每个模式都可以由基元根据一定的关系来组成
- (3)基元可以认为是语言中的字母，每个模式都可以认为是一个句子，关系可以认为是语法或句法
- (4)模式的相似性由句子的相似性来决定
- (5)用已知类别的训练样本进行学习，产生该类或至少是这些样本的方法，该学习和训练过程称为文法推断。

优点：适合结构性强的模式

缺点：抗噪声能力差，计算复杂度高

1.3 模式识别的基本方法

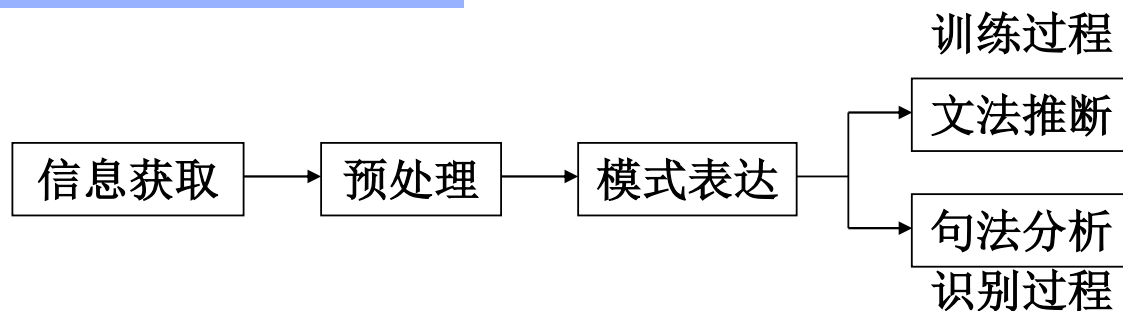


图 句法模式识别系统组成

其中，**模式表达**包括两部分：**模式分割**和**基元及关系的识别**。对于一个模式，经过预处理并对模式分解提取基元后，得到表征模式的句子，然后进行句法分析，判断它是否能被代表某个模式类的文法所接受，最终给出模式结构描述和识别结果。

1.3 模式识别的基本方法

4.神经网络模式识别

(1)神经网络模式识别主要利用人工神经网络的学习、记忆和归纳功能,先根据训练样本训练分类器,再利用分类器对待识别对象进行分类决策

(2)大规模并行计算

(3)学习、推广、自适应、容错、分布表达和计算

优点：可以有效的解决一些复杂的非线性问题

缺点：模型还在不断完善之中，目前能识别的模式类还不够多

1.3 模式识别的基本方法

表 几种基本模式识别方法的比较

方法	表达	识别函数	主要理论支撑
模板匹配	样本、像元、曲线	相关、距离度量	几何学
统计方法	特征	决策函数	概率论与数理统计
句法方法	基元	规则、语法	形式语言、自动机技术
神经网络	样本、像元、特征	网络函数	神经生理学、心理学

1.4 模式识别的应用

1. 字符识别：包括印刷体字符的识别；手写体字符的识别（脱机），各种OCR设备例如信函分拣、文件处理、卡片输入、支票查对、自动排板、期刊阅读、稿件输入；在线手写字符的识别（联机），各种书写输入板。
2. 医疗诊断：心电图，脑电图，染色体，癌细胞识别，疾病诊断。
3. 遥感：资源卫星照片，气象卫星照片处理，数字地球，图像分辨率可以达到0.3米。

1.4 模式识别的应用

4. 指纹识别，人脸识别
5. 检测污染分析，大气，水源，环境监测。
6. 自动检测：产品质量自动检测
7. 语声识别，机器翻译，电话号码自动查询，侦听，机器故障判断。
8. 军事应用
9. 二维码识别及应用(目前的热点应用，如微信中的扫一扫、商品跟踪识别等)

○ ○ ○ ○ ○ ○

国内外顶级的机器学习与模式识别公司有Google、百度等(大家课后浏览百度深度学习研究院网址<http://idl.baidu.com/>，可知百度目前在研究开发什么创新产品，也知道如何结合自己的兴趣进行创新型学习与研究)。说明：**深度学习(Deep Learning)**是机器学习目前最热门的一个研究分支，主要是采用深度神经网络方法并结合大数据进行机器学习算法及应用的研究，著名深度学习方法有**CNN(Convolution al Neural Networks**，卷积神经网络)、**DBN(Deep Belief Networks**，深度信念网络)等。

百度大脑则进行了应用研究开发：<http://ai.baidu.com/>

● 模式识别与机器学习在移动可穿戴式计算机系统中的应用举例

(借助百度云开源智能手环)



百度智能手环硬件是全部开源硬件(目前是国内外硬件开源程度最高的), 借助机器学习算法做健康生理参数智能监测与诊断, 具体包括:

- 电路原理图;
- BOM List (元器件清单);
- 硬件设计文档;
- 智能手环ROM源代码以及文档;
- 智能手环与手机的蓝牙通信协议;
- 基于机器学习的运动检测和计步算法(借助百度云运行机器学习程序, 将运动检测和计步等分析决策结果返回至用户手机客户端)。

1.5 模式识别的基本问题

一.模式(样本)表示方法

1.向量表示：假设一个样本有n个变量(特征)

$$X = (X_1, X_2, \dots, X_n)^T$$

2. 矩阵表示： N个样本， n个变量(特征)

变量 样本	x_1	x_2	\dots	x_n
X_1	X_{11}	X_{12}	\dots	X_{1n}
X_2	X_{21}	X_{22}	\dots	X_{2n}
\dots	\dots	\dots	\dots	\dots
X_N	X_{N1}	X_{N2}	\dots	X_{Nn}

1.5 模式识别的基本问题

3. 几何表示

1D表示

$$X_1=0.5 \quad X_2=3$$

2D表示

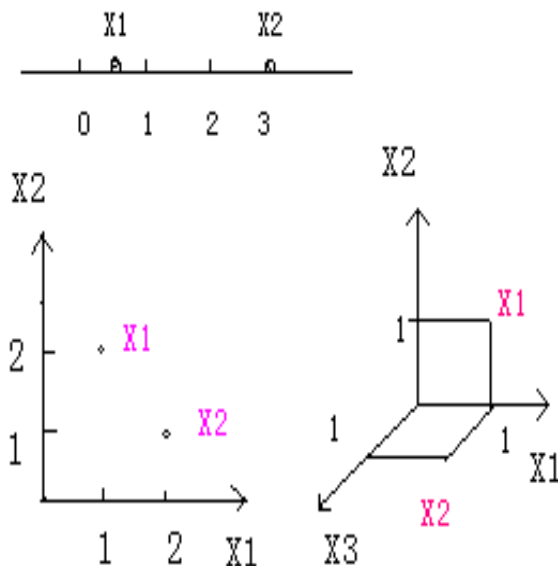
$$X_1=(x_1, x_2)^T=(1, 2)^T$$

$$X_2=(x_1, x_2)^T=(2, 1)^T$$

3D表示

$$X_1=(x_1, x_2, x_3)^T=(1, 1, 0)^T$$

$$X_2=(x_1, x_2, x_3)^T=(1, 0, 1)^T$$



1.5 模式识别的基本问题

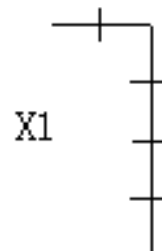
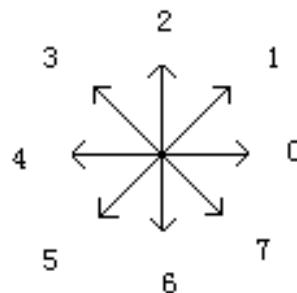
4.基元（链码）表示：

在右侧的图中八个基元分别表示0, 1, 2, 3, 4, 5, 6, 7, 八个方向和基元线段长度。

则右侧样本可以表示为

$$X_1 = 006666$$

该方法在句法模式识别中用到。

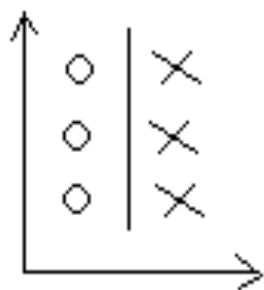


1.5 模式识别的基本问题

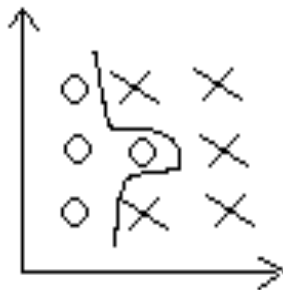
二.模式类的紧致性

1. 紧致集

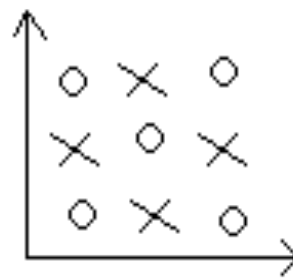
同一类模式类样本的分布比较集中，没有或临界样本很少，这样的模式类称**紧致集**。



无临界点



有临界点



临界点太多，无法分类

1.5 模式识别的基本问题

2. 临界点(样本): 在多类样本中, 某些样本的值有微小变化时就变成另一类样本称为**临界样本** (点)。

3. 紧致集的性质

① 要求临界点很少

② 集合内的任意两点的连线, 在线上的点属于同一集合

③ 集合内的每一个点都有足够大的邻域, 在邻域内只包含同一集合的点

4. **模式识别的要求**: 满足紧致集, 才能很好的分类; 如果不满足紧致集, 就要采取变换的方法, 以满足紧致集。

1.5 模式识别的基本问题

三.相似与分类

1.两个样本 x_i ， x_j 之间的相似度量满足以下要求：

- (1)应为非负值
- (2)样本本身相似性度量应最大
- (3)度量应满足对称性
- (4)在满足紧致性的条件下，相似性应该是点间距离的单调函数

2. 用各种距离表示相似性：

(1)绝对值距离

已知两个样本 $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{in})^T$
 $x_j = (x_{j1}, x_{j2}, x_{j3}, \dots, x_{jn})^T$

定义：

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|$$

1.5 模式识别的基本问题

(2) 欧几里德距离

$$d_{ij} = \sqrt{\sum_{k=1}^n (X_{ik} - X_{jk})^2}$$

(3) 明考夫斯基距离

$$d_{ij}(q) = \left(\sum_{k=1}^n |X_{ik} - X_{jk}|^q \right)^{1/q}$$

$k=2, q=1$ 时, 为街区距离

其中当 $q=1$ 时为绝对值距离, 当 $q=2$ 时为欧氏距离

1.5 模式识别的基本问题

(4)切比雪夫距离

$$d_{ij}(\infty) = \max_{1 \leq k \leq n} |X_{ik} - X_{jk}|$$

明氏距离, q 趋向无穷大时的极限情况

(5)马哈拉诺比斯距离(Mahalanobis)

$$d_{ij}(M) = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)}$$

其中 \mathbf{X}_i 、 \mathbf{X}_j 为特征向量集 $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$ 中的两个 n 维特征向量, Σ 为模式总体的协方差矩阵。马氏距离的使用的条件是样本符合正态分布。

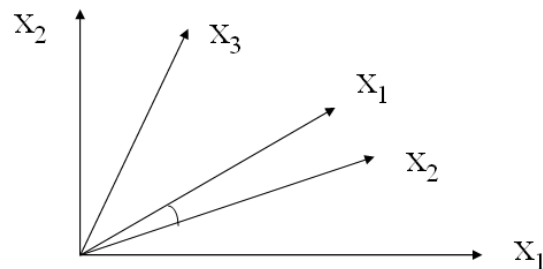
注: 实际使用中, 马氏距离中的根号有时可去掉。

1.5 模式识别的基本问题

(6) 夹角余弦
$$C_{ij} = \cos \theta = \frac{X_i^T \cdot X_j}{|X_i| \cdot |X_j|} = \frac{\sum_{k=1}^n X_{ik} X_{jk}}{\sqrt{\left(\sum_{k=1}^n X_{ik}^2\right) \left(\sum_{k=1}^n X_{jk}^2\right)}}$$

即两样本间夹角小的为一类，具有相似性。

例： x_1, x_2, x_3 的夹角如右图。



因为 x_1, x_2 的夹角小，所以 x_1, x_2 最相似。

1.5 模式识别的基本问题

(7) 相关系数

$$r_{ij} = \frac{\sum_{k=1}^n (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\sum_{k=1}^n (X_{ki} - \bar{X}_i)^2 \sum_{k=1}^n (X_{kj} - \bar{X}_j)^2}}$$

\bar{X}_i, \bar{X}_j 为 x_i, x_j 的均值

注意：在求相关系数之前，要将数据标准化。

3. 分类的主观性和客观性

(1) 分类带有主观性：目的不同，分类不同。如：鲸鱼、牛、马从生物学的角度来讲都属于哺乳类，但是从产业角度来讲鲸鱼属于水产业，牛和马属于畜牧业。

(2) 分类的客观性：科学性

判断分类必须有客观标准，因此分类是追求客观性的，但主观性也很难避免，这就是分类的复杂性。

1.5 模式识别的基本问题

四.特征的生成

1.低层特征:

(1)无序尺度: 有明确的数量和数值。

(2)有序尺度: 有先后、好坏的次序关系, 如酒分为上, 中, 下三个等级。

(3)名义尺度: 无数量、无次序关系, 如有红、黄两种颜色

2. 中层特征: 经过计算、变换得到的特征

3. 高层特征: 在中层特征的基础上有目的的经过运算形成

例如: 椅子的重量=体积*比重

体积与长、宽、高有关; 比重与材料、纹理、颜色有关。这里低、中、高三层特征都具备了。

1.5 模式识别的基本问题

五.数据的标准化

1.极差标准化

一批样本中，每个特征的最大值与最小值之差，称为极差。

极差

$$R_i = \max X_{ij} - \min X_{ij}$$

极差标准化

$$X_{ij} = \frac{(X_{ij} - \bar{X}_i)}{R_i}$$

2. 方差标准化

$$X_{ij} = \frac{(X_{ij} - \bar{X}_i)}{S_i}$$

S_i 为方差

标准化的方法很多，原始数据是否应该标准化，应采用什么方法标准化，都要根据具体情况来定。