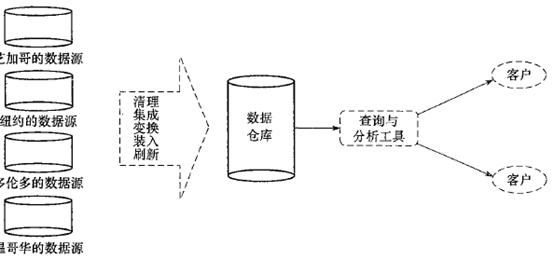
# 引论

## 数据挖掘(概念+步骤KDD)

1. **数据挖掘**：数据挖掘是从大量数据中挖掘有趣模式和知识的过程。
2. **数据挖掘**的另一个流行术语**数据中的知识发现**（KDD），以下为知识发现的**过程**。
   1. **数据清理**（消除噪声和删除不一致的数据）
   2. **数据集成**（多种数据源集合在一起）
   3. **数据选择**（从数据库中提取和分析任务相关的数据）
   4. **数据变换**（通过汇总或聚集操作，把数据变换和**统一**成适合挖掘的**形式**）
   5. **数据挖掘**（基本步骤，使用智能方法**提取数据模式**）
   6. **模式评估**（根据某种兴趣度**度量**，识别代表知识的8真正有趣的模式）
   7. **知识表示**（使用**可视化和知识表示技术**，向用户提供挖掘的知识）

**3.数据仓库**：数据仓库**是一个从多个数据源收集的信息存储库**，**存放在一致的模式下**，并且通常驻留在单个站点上。数据仓库通过数据清理、数据变换、数据集成、数据装入和定期刷新来构造。

典型框架：



## 可以挖掘什么数据:

**数据库数据**

由一组内部相关的数据和一组管理和存取数据的软件程序组成。

关系数据库是表的汇集（属性—>字段或列，元祖—>记录或行）。对象被唯一关键字标识，被一组属性值描述。

**数据仓库**

从多个数据源收集的信息储存库（例如分部遍布全世界的公司的数据库）

数据仓库举例 书P7

**事务数据**

每个记录代表一个事务。

包含一个唯一的事务标识号，以及一个组成事务的项。

例如商场的**第50条交易记录**，用户购买了A，D，F这三件物品。

**其他类型的数据**

空间数据，超文本和多媒体数据……等等

## 可挖掘什么类型的模式

**数据挖掘的模式:**

* 1. **特征**化与区分
  2. 频繁模式、关联性和相关性挖掘
  3. 分类与回归
  4. 聚类分析
  5. 离群点分析

或者:**可挖掘什么类型的模式**:

描述性（刻画目标数据中数据的一般性质）

预测性（在当前数据上做出归纳，以便进行预测）

**类/概念描述：特征化与区分**

数据特征化：目标类数据的一般特性或特征的汇总

数据区分：将目标类，与一个或多个可比较类进行比较。

例如：定期购买电脑产品的客户和不购买电脑产品的客户进行比较。

**挖掘频繁模式、关联和相关性**

频繁模式：在数据中频繁出现的模式。

频繁项集：频繁在事务数据集中一起出现（顾客在小卖部总是一起买牛奶和面包）

频繁子序列：顾客先买数码相机，再买内存卡。

频繁子结构

关联分析（例如分析，哪些商品总是一起被购买）

“computer”=>“software”[1% , 50%] 表示所有事务的1%显示计算机和软件被同时购买。购买了计算机的人，有50%的可能性会选择购买软件。

相关性（相关联的属性-值对之间的统计相关性）

**用于预测分析的分类和回归**

分类：找出描述和区分数据类或概念的模型（或函数），预测类别标号。

回归：建立连续值函数模型，预测缺失的或难以获得的数值数据值。

相关分析可能需要在分类和回归之前进行，它试图识别与分类和回归过程显著相关的属性。

**聚类分析**

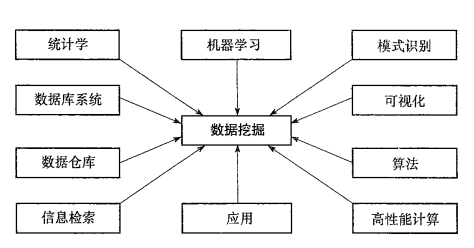
聚类分析：分析数据对象，而不考虑类标号。可以使用聚类产生数据组群的类标号。

—>“最大化类内相似性，最小化类间相似性”

**离群点分析**

异常挖掘。有时看做噪声而丢弃，但是在比如欺诈检测等应用中，罕见事件的出现，更令人感兴趣。

## 数据挖掘使用什么技术



### 机器学习(概念+经典学习问题)：

机器学习**考察计算机如何基于数据学习（或提高它们的性能）**，主要**指计算机程序基于数据自动的学习识别复杂的模式，并做出智能的决断**。

**6.经典机器学习问题：**

**a. 监督学习**：**基本指分类**，学习中的**监督**来自数据集中**标记的实例** （训练数据有类标记）

**b. 无监督学习：本质是聚类** ，输入**数据集中没有类标记**，典型的我们可以**使用聚类发现数据中的类**。

**c. 半监督学习：**是一类机器学习技术，在学习模型时，使用标记的和未标记的实例。在一种方法中，**标记的实例用来学习类模型，未标记的实例用来进一步改进类边界**。

**d. 主动学习：**是一种机器学习方法，**让用户扮演主动角色**。它要求用户对一个可能来自未标记的实例集或由学习程序合成的**实例进行标记**。

信息检索+统计学

# 第二章 认识数据

## 属性概括

属性(attribute)是一个数据字段，表示数据对象的一个特征。

标称属性(nominal attribute)意味着‘与名称相关’，它的值是一些符号或事物的名称。**每个值代表某种类别，编码或状态**，因此标称属性又被看作是分类的(categorical).这些值不必具有有意义的序。在标称属性上，数学运算是没有意义的。(name)

二元属性(binary attribute）是一种标称属性，只有0，1两种状态。通常0表示不出现，1表示出现。

序数属性(ordinal attribute)，具有先后顺序。

以上属性都是**定性的**。即它们**描述对象的特征**，而不给出实际大小或数量。

数值属性(numeric attribute)是定量的，即它是可度量的量。数值属性可以是**区间标度(属于哪个区间)的或比率标度(或者所占的百分比)**的。

区间标度属性(interval-scaled attribute)用相等的单位尺度度量。区间属性的值。

比率标度(ratiao-scaled)属性具有固定零点的数值属性。可以进行数值运算。

***标称属性、二元属性(性别)、序数属性(编号)*、数值属性(年龄)（特殊：区间标称属性和比率标度属性）、离散属性与连续属性**

**标称属性：一些符号和事物的名称，如头发颜色和学历是描述人的属性。**

**二元属性：只有0、1两种状态**

**序数属性： 可能的值具有有意义的序，如教师职称**

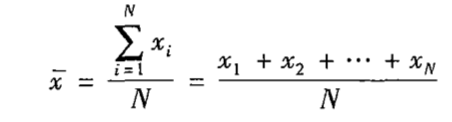
**数值属性： 定量的，用实数值表示**

## 数据的基本统计方法（应用）

### a.中心趋势度量：

**均值：**

普通均值：



加权算数均值（或加权平均）

**中位数:先排序之后确定个数(奇数和偶数个的计算方法) (https://jingyan.baidu.com/article/425e69e69f07fbbe15fc161f.html)**

**众数:是数列中出现次数最多的数，要是没有出现最多的数，就没有众数，或者全都是众数。**

**中列数:**

中列数，midrange，在统计中指的是**数据集里最大值和最小值的算术平均**。

如以下集合： 1，3，7，9，0，3，5

它的中列数即为（0+9）/2 = 4.5

### b.度量数据散布(离散程度)：

**极差、四分位数、方差、标准差和四分位数极差 (五数概括和盒图)**

**①极差：一组数据中的最大数据与最小数据的差叫做这组数据的极差.例如：12 12 13 14 16 21**

**这组数的极差就是：21-12=9**

**②四分位数:**

**四分位数有三**个**，第一个四分位数就是通常所说的四分位数，称为下四分位数，第二个四分位数就是中位数，第三个四分位数称为上四分位数，分别用Q1、Q2、Q3表示。**

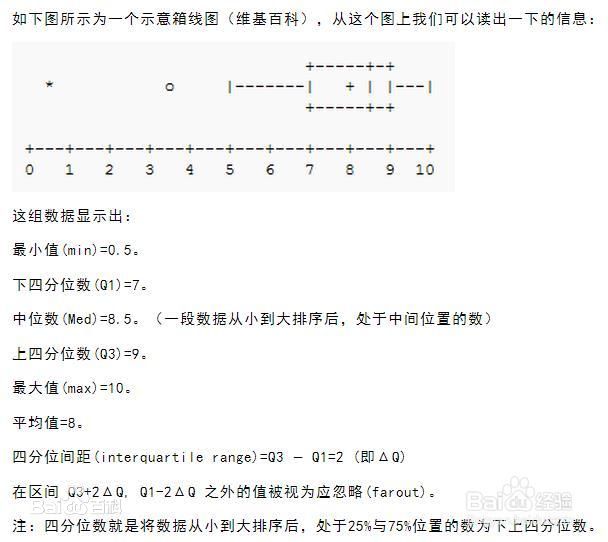
**第一四分位数 (Q1)，又称“较小四分位数”，等于该样本中所有数值由小到大排列后第25%的数字。**

**第二四分位数 (Q2)，又称“中位数”，等于该样本中所有数值由小到大排列后第50%的数字。**

**第三四分位数 (Q3)，又称“较大四分位数”，等于该样本中所有数值由小到大排列后第75%的数字。**

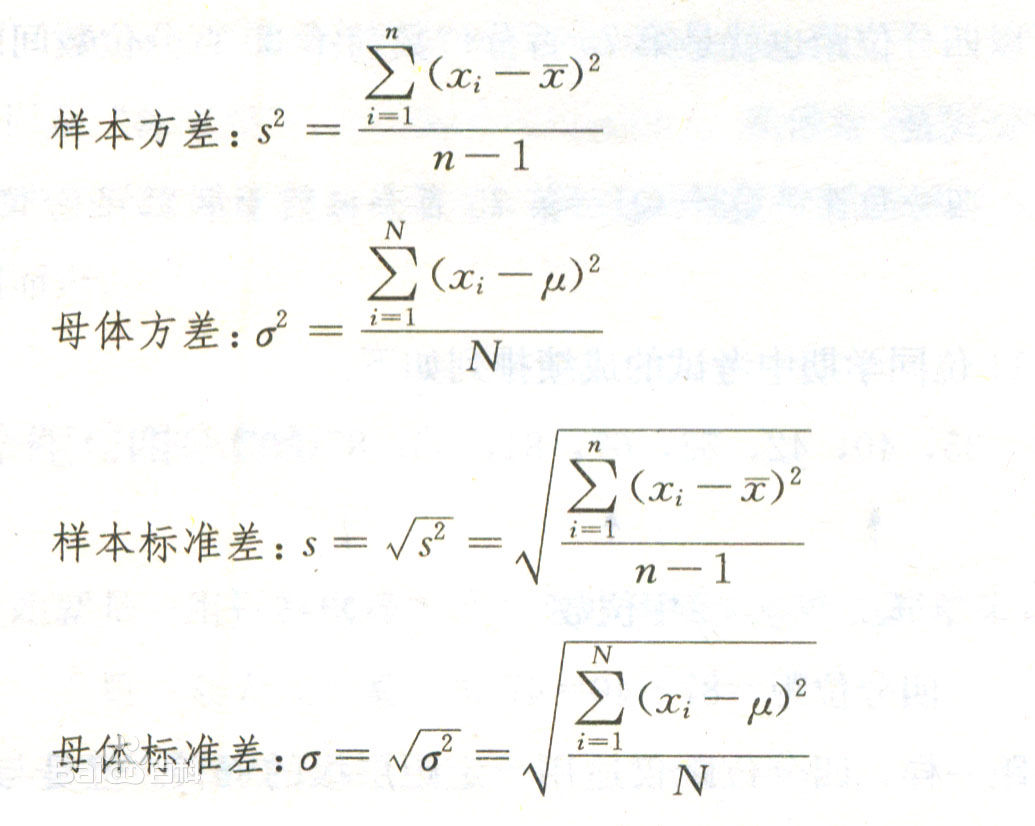
**第三四分位数与第一四分位数的差距又称四分位距/差（InterQuartile Range,IQR）**

|  |
| --- |
| **例1：由7人组成的旅游小团队年龄分别为：17、19、22、24、25、28、34，求其年龄的四分位差。计算步骤为：**   1. **计算Q1，与Q3的位置。确定4分位数的位置(先排序)**   **Q1的位置= （n + 1) / 4 = （7 + 1) / 4 = 2**  **Q3的位置= 3\*（n + 1) / 4 = 3\*（7 + 1) / 4 = 6**  **即Q1与Q3的位置分别为第2位和第6位。**   1. **确定Q1与Q3的数值。确定相应的数值🡪也就是四分位数**   **Q1=19(岁)**  **Q3=28(岁)**  **即第2位和第6位对应年龄分别为19岁和28岁。**  **③计算**四分位差/矩**。**  **Q．D．=Q3 − Q1=28-19=9(岁)**  **④含义。*说明该旅游小团队有50%的人年龄集中在19～28岁之间，最大差异为9岁。*** |



**箱图(盒图)从左往右看:最小值.下四分位(看清位置),平均值(+号),中位数(中四分位数),下四分位数,最大值**

**方差/标准差公式:**



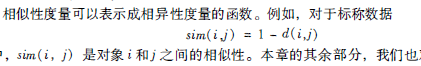
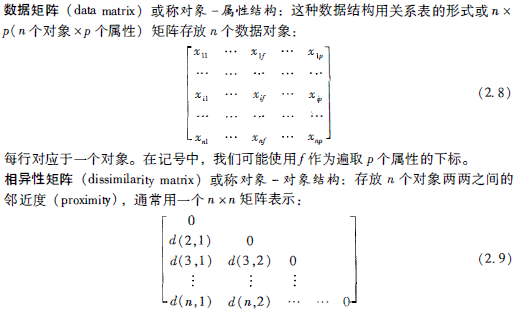
**c.数据基本图形显示：分位数图、直方图、散点图**

### c.度量数据的相似性和相异性，距离的计算。

相似性和相异性都称**邻近性,** 提供通常用于上述应用的两种数据结构：**数据矩阵（用于存放数据对象）**和**相异性矩阵（用于存放数据对象对的相异性值）,来处理由多个属性刻画的对象**

#### 数据矩阵和相异性矩阵

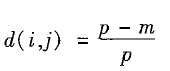
* + 1. 数据矩阵：对象-属性结构
    2. 相异性矩阵(相异度矩阵)：对象-对象结构，**存放n个对象两两之间的邻近度d(i,j)**，一般非负。数值越小，对象之间的相似度越接近（d(i,j)=0时**，即一个对象与自己的差别为0**）。

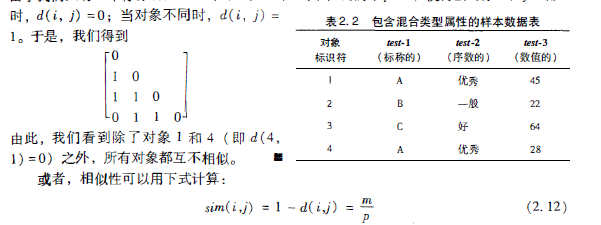
 

#### 非数值属性度量

##### 标称属性的邻近性度量

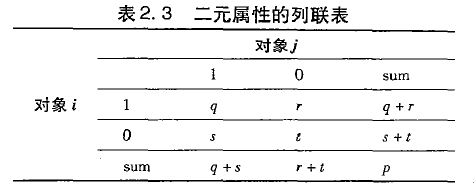
标称属性可以取两个或多个状态的状态的属性



公式：其中，**m是**匹配的数目**（i和j取值相同状态的属性数）**，**p是刻画对象的属性总数**。

##### 二元属性的邻近性度量

二元属性只有两种状态：0或1，其中0表示该属性不出现，1表示它出现 (2. l. 3节）



* + 1. **对称的**二元相异性(上面)( 就是形成的矩阵是对称的)

图片

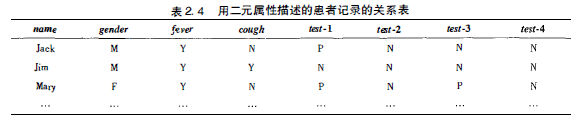
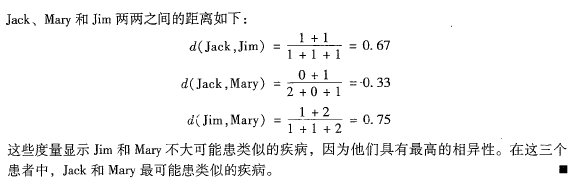
* + 1. **非对称**的二元相异性

对于非对称的二元属性()， 两个状态不是同等重要的；如病理化验的阳性（1 ）和阴性 (O）结果。 给定两个非对称的二元属性，两个都取值 1 的情况（正匹配）被认为比两个都取值。的情况（负匹配）更有意义。 因此， 这样的二元属性经常被认为是 “一元的”（只有 一种状态）。 基于这种属性的相异性被称为非对称的二元相异性， 其中负匹配数t被认为是不重要的， 因此在计算时被忽略， 如下所示：

图片

* + 1. 非对称的**二元相似性**系数sim(i,j)称为Jaccard系数(杰卡德))。

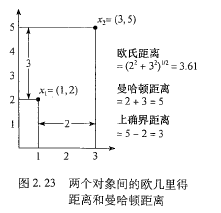
互补地， 我们可以**基于相似性(看中相似性,因为相异性t的结果未必就是相似性)**而不是基于相异性来度量两个二元属性的差别。 例如， 对象i和j之间的非对称的二元相似性可以用下式计算：

图片

##### 序列属性的度量

#### 数值属性度量

常用距离度量:欧几里得距离、闵可夫斯基距离,曼哈顿距离、、上确界距离



##### 闵可夫斯基距离（Minkowski distance ）

是欧几里得距离和曼哈顿距离的推广,闵可夫斯基距离(因为是可计算), 这种距离又称lp范数（norm），其中p就是我们的h。我们保留p作为属性数，以便于本章的其余部分一致。）当 p = 1 时，它表示曼哈顿距离（即，Li 范数）；当 p =2 表示欧几里得距离（即，L 范数）。

图片

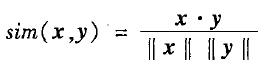
其中，h是实数，h≥1。

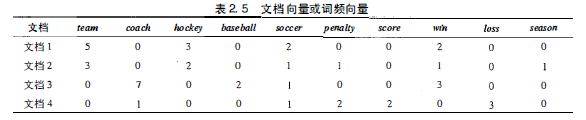
##### 上确界距离

**特征有不同的权重**

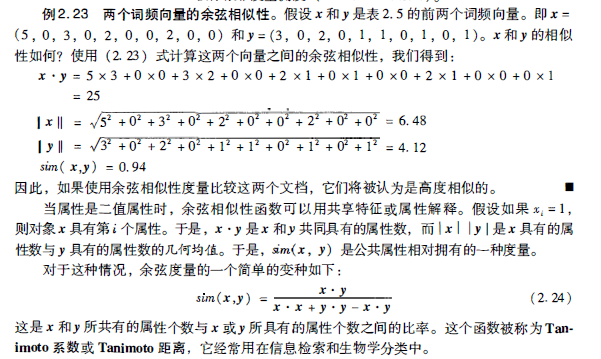
#### 相异性与相似性评估

**余弦相似性—相似度评估/词频向量**



文档用数以千计的属性表示，每个记录文档中一个特定词（如关键词） 或短语的频度。这样，每个文档都被一个所谓的词频向量（term-frequency vector） 表示。例如，在表2.5中，我们看到丈档l包含词 team 的 5个实例，而 hockey 出现 3次。正如计数值。所 示 coαch 在整个文档中未出现。这种数据可能是高度非对称的。

词频向量通常很长，并且是稀疏的（即，它们有许多 0 值）(就是说很多的词(特征)在不同的文档之中是没有的)。使用这种结构的应用包括信息检索、文本文档聚类、生物学分类和基因特征映射。对于这类稀疏的数值数据， 本章我们研究过的传统的距离度量效果并不好。例如，**两个词频向量可能有很多公共 0值，意味对应的文档许多词是不共有的，而这使得它们不相似。**我们需要一种度量，它关注两个文档确实共有的词，以及这种词出现的频率。换言之，我们需要忽略0匹配的数值数据度量。

**余弦相似性是一种度量，它可以用来比较文挡，或针对给定的查询词向量对文档排序。例子如下:**

**类似 公共属性的度量🡪相似性的度量**

# 第三章 数据预处理

## 预处理的主要任务/与KDD知识发现区别开来

**数据清理、数据集成、数据归约、数据变换（属性构造**,**规范化，数据离散化，概念分层）**

现实世界的**数据一般是脏的、不完整的和不一致的**。**数据预处理技术可以改进数据的质量**，从而有助于提高其后挖掘过程的准确率和效率。由于高质量的决策必然依赖于高质量的数据，因此数据预处理是知识发现过程的重要步骤。检测数据异常，尽早的调整数据，并归约待分析的数据，将为决策带来高回报。

|  |
| --- |
| 1). 数据清理：**填补*缺失值*、光滑噪声、识别离群点、纠正数据不一致性**，通常是一个两步迭代过程，包括偏差检测盒数据变换  **处理缺失值**：有忽略元组、人工填写、全局常量填充、  均值或中位数填充、同类均值或中位数填充、最可能值填充六种方法。  **噪声数据处理**：分箱（考察近邻数据值，有箱均值、箱中位数、箱边界光滑方法）  **回归、离群点分析:**  2). 数据集成： 将来自多个数据源的数据整合成一致的数据存储  **实体识别问题**：模式集成和对象匹配。如在一个系统中discount用于订单，而在另一个系统中用于商品，集成不正确导致商品不正确打折  **冗余和相关分析**：冗余指一个属性能由另一个或另一组属性导出，则这个属性是冗余的，可用相关分析检测到。  **标称数据**使用卡方检验，数值属性用相关系数和协方差  3). 数据规约：简化数据集的表示，包括维规约和数值规约  **维规约**：数据压缩技术（如小波变换和主成分分析）， 属性子集选择，属性构造  **数值规约**： 参数模型（如回归），非参数模型（聚类、抽样、直方图）  4). 数据变换：将数据变换成适于挖掘的形式  **变换策略**包括：  光滑(去掉数据中的噪声。这类技术包括分箱、 回归和聚类。)  属性构造: 可以由给定的属性构造新的属性并添加到属性集中  聚集: 对数据进行汇总或聚集。例如， 可以聚集日销售数据，计算月和年销售量。  规范化: 把属性数据按比例缩放  离散化(用youth代替年龄段0~10,10~20🡪就会形成概念分层): ：数值属性（例如， 年龄）的原始值用区间标签（例如， 0-10, 11 ～ 20  由标称数据产生概念分层: |

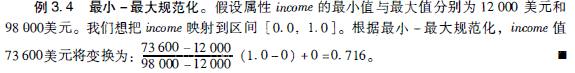
## 规范化的方法（应用计算，要会手算）

### 原因

消除量纲,减除单位不同导致数值不同带来的影响

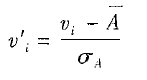
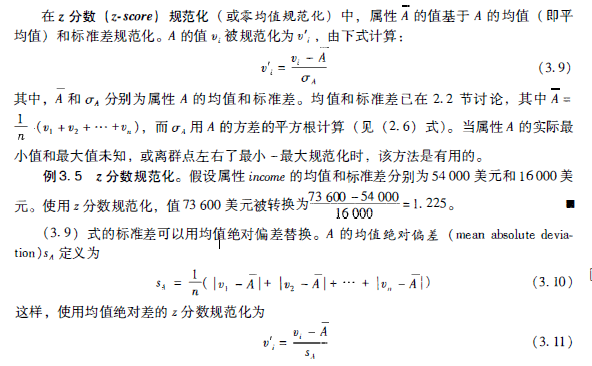
规范化数据试图赋予所有属性相等的权重。对于涉及神经网络的分类算法或基于距离度量的分类（如最近邻分类） 和聚类，规范化特别有用。如果使用神经网络后向传播算法进行分类挖掘（第9章），对训练元组中每个属性的输入值规范化将有助于加快学习阶段的速度 对于 基于距离的 方法，规范化可以帮助防止具有较大初始值域的属性（如 income ） 与具有较小初始值域的属性（如二元属性） 相比权重过大 。在没有数据 的先验知识时，规范 化也是有用的。

### 图片最小-最大规范化



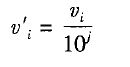
映射在某个区间中.一般是(0,1)🡪所以区间长度是1

### z分数规范化

当属性A**的实际最小值和最大值未知**，或**离群点(也就是说离群点可能是最大值/最小值)左右了最小一最大规范化**时，该方法是有用的。

对于离群点， **均值绝对偏差 SA** 比标准差更加鲁棒。 因为在计算均值绝对偏差时，点到均值的偏差 （即|xi-均x|） 取平方， 因此离群点的影响多少有点降低。

### 小数定标规范化

小数定标规范化通过移动属性A的值的小数点位置进行规范化。 小数点的移动位数依赖于A的最大绝对值。A的值vi被规范化为vi’,由下式计算：

例子:假设A的取值由－986 到917。A的最大绝对值为986。 因此， 为使用小数定标规范化， 我们用1000（即 j =3）除每个值。 因此， －986被规范化为－ 0.986， 而917被规范化为 0.917。

### 总结

# 第四章 数据仓库与联机分析处理

## 数据仓库的概念、特征及数据仓库与数据库的区别；

**数据仓库：**是一种**语义上一致的数据存储**，它**充当决策支持数据模型的物理实现**，并存放企业战略决策所需要的信息。

**特征：面向主题的、集成的、时变的、非易失的**。

**区别：**

数据仓库（Data Warehouse）是一个**面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）**的数据集合，用于支持管理决策。

（1） 面向主题：指数据仓库中的数据是按照一定的主题域进行组织。

（2）集成：指对原有分散的数据库数据经过系统加工, 整理得到的消除源数据中的不一致性。

（3）相对稳定：指一旦某个数据进入数据仓库以后只需要定期的加载、刷新。

（4）反映历史变化：指通过这些信息，对企业的发展历程和未来趋势做出定量分析预测。

主要区别在于：(根据特点来回答)

（1）数据库是**面向事务**的设计，数据仓库是**面向主题**设计的。(主题的)

（2）数据库一般**存储在线交易数据(实时数据)**，数据仓库存储的**一般是历史数据(**反映历史变化**)**。

（3）数据库设计是**尽量避免冗余**，数据仓库在设计是**有意引入冗余**。(特点:集成)

（4）数据库是为**查询数据而设计**，数据仓库是**为分析数据而设计**。(目的不一样)

## 数据仓库的体系构建；

底层是**仓库数据库服务器**，中间层是**OLAP服务器**，顶层是**前端客户层**

数据仓库：一种多层体系结构（三层）

（1）底层是仓库数据库服务器，它几乎总是一个关系数据库系统。

（2）中间层是OLAP服务器，其典型的实现使用（i）关系OLAP（ROLAP）模型（即扩充的关系DBMS，将多维数据上的操作映射为标准的关系操作），或者使用（ii）多维OLAP（MOLAP）模型（即专门的服务器，直接实现多维数据和操作）。

（3）顶层是前端客户层，包括**查询和报告工具**、分析工具和数据挖掘工具（预测，分析）

## 数据仓库模式的构建；（应用）--就是画模型图

**数据仓库的模型（从结构角度看，有三种模型）**

企业仓库用于收集跨越**整个企业的各个主题的所有信息**，它提供整个企业范围的数据集成。

而数据集市包含对特定的用户有用的、企业范围数据的一个子集，其范围限于**所选的特定定的主题**。

虚拟仓库是操作型数据库上**视图的集合**。**为了便于可视化查询**

## olap操作（应用）

**上卷：**通过沿一个维的概念分层向上攀升或者通过维归约在数据立方体上聚集。 **(维度上升)**

**下钻：上卷的逆操作**，它由不太详细的数据到更详细的数据。  **(维度下降)**

**切片：**在给定的立方体的**一个维上**进行选择，导致一个子立方体。 **选择单个维度**

**切块:** 在给定的立方体的**两个或者多个维上**进行选择，导致一个子立方体 **选择多个维度**

**转轴：**是一种目视操作，它转动数据的视角，提供数据的替代表示

**其他：钻过，执行涉及多个事实表的查询； (常常select )**

钻透，使用关系SQL机制，钻透到数据立方体的底层，到后端关系表。

**何谓OLTP和OLAP？它们的主要异同有哪些？**

OLTP即联机事务处理，是以传统数据库(1)为基础、面向操作人员和低层管理人员(2)、对基本数据进行查询和增、删、改等的日常事务处理。OLAP即联机分析处理，是在OLTP基础上发展起来的、以数据仓库基础上的、面向**高层管理人员和专业分析人员**、为企业决策支持服务。

OLTP和OLAP的主要区别如下表：

|  |  |
| --- | --- |
| OLTP | OLAP |
| 数据库数据 | 数据库或数据仓库数据 |
| 细节性数据 | 综合性数据 |
| 当前/实时数据(3) | **历史数据** |
| 经常更新 | 不更新，但周期性刷新 |
| 一次性处理的数据量小 | 一次处理的数据量大 |
| 对响应时间要求高 | 响应时间合理 |
| 用户数量大 | 用户数据相对较少 |
| 面向操作人员，支持日常操作 | 面向决策人员，支持管理需要 |
| 面向应用，事务驱动 | 面向分析，分析驱动 |

# 第五章 数据立方体技术

## 数据立方体的概念；

数据立方体：一种多维数据模型

（1）数据立方体：允许以多维对数据建模和观察，它由维和事实定义。

1）维是一个单位想要记录的实体：商品的销售，设计维time、item、location；每个维也可以有一个与之相关的表，称为维表，它将进一步描述维。例如，item的维表可以包含属性item\_name、iterm\_type等。

**多维数据模型的模式:**

星形: **星型模式的核心是一个大的中心表（事实表），一组小的附属表（维表）**

雪花: 星型模式的扩展，其中某些维表被规范化，进一步分解到附加表（维表）中

星座：包含多个事实表，而维表是公共的，可以共享，这种模式可以看做星型模式的汇集. 是数据仓库最长使用的数据模式, 为了避免冗余和数据复用，套用现成的模式

## 数据立方体的计算方法有哪些；

* 完全立方体计算的**多路数组聚集**—计算完全方体

* BUC：**从顶点方体向下**计算冰山立方体--计算稀疏冰山立方体
* Star-Cubing：使用**动态星树结构**计算冰山立方体
* 为快速高维OLAP预计算壳片段

## *数据立方体有效计算的一般优化技术有哪些？*

**优化技术1**：先**排序、和分组**

**优化技术2**：同时**聚集和缓存中间结果**

**优化技术3**：当存在多个子女方体时，由**最小的子女聚集**

**优化技术4**：可以使用**先验剪枝**方法有效的计算冰山立方体

# 第六章 挖掘频繁模式、关联性和相关性：基本概念和方法

## 频繁模式和相关规则的相关概念；

**频繁模式：**是指**频繁的出现**在数**据集中的模式**（如项集、子序列或子结构）-🡪**支持度衡量**

**关联规则：挖掘发现大量数据中项集之间的有趣关联联系。置信度衡量**

规则的**支持度**和**置信度**是**规则兴趣度的两种度量**。典型情况下，**若满足最小支持度阈值和最小置信度阈值，则关联规则被认为是有趣的**

## （应用）频繁项集的挖掘方法；

**Apriori算法：通过限制候选产生发现频繁项集**

**先验性质：**频繁项集的所有非空子集也一定是频繁的

（1）连接步

（2）剪枝步

**由频繁项集产生关联规则**

**提高Apriori算法效率**

## (应用)挖掘频繁项集的模式增长方法（FPgrowth，）---

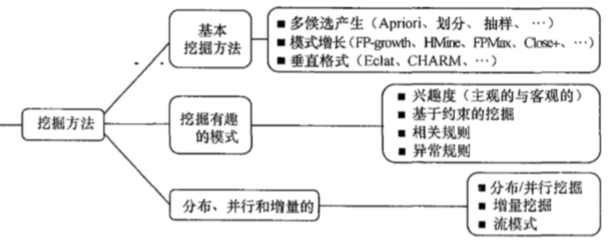
采用分治策略：首先，将代表频繁项集的数据库压缩到一颗**频繁模式树**，该树仍保留项集的关联信息。然后，把这种**压缩后的数据库划分成一组条件数据库**，每个数据库**关联一个频繁项或“模式段”**，并分别挖掘每个条件数据库。对于每个“模式片段”，只需要考察与它相关联数据集。

# 第七章 高级模式挖掘

## 高级模式挖掘的方法

模式挖掘是比频繁模式挖掘更一般的术语，前者还涵盖了**稀有模式和负模式**。

下边这个图不知道对不对？



模式挖掘：一个路线图

**模式挖掘的类型:**

**多层、多维空间中的模式挖掘**

**基于约束的频繁模式挖掘**

**挖掘高维数据和巨型模式**

**挖掘压缩或近似模式**

模式探索与应用

# 第八章 分类：基本概念

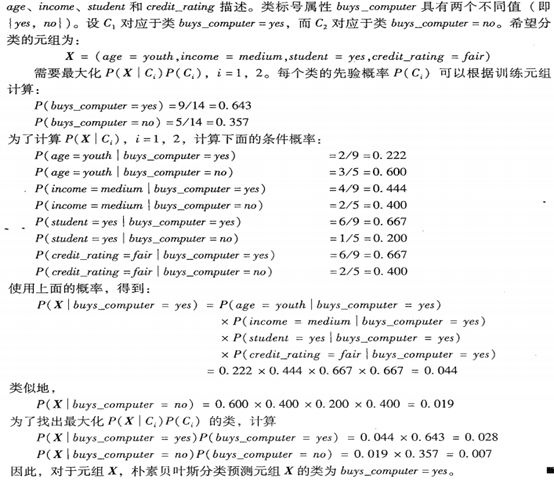
## 分类的基本概念

**数据分类是一个两阶段过程**，**包括学习阶段（构建分类模型）**和**分类阶段（使用模型预测给定数据的类标号）**。

## 决策树（应用）

8.2.1, 8.2.2, 8.2.3

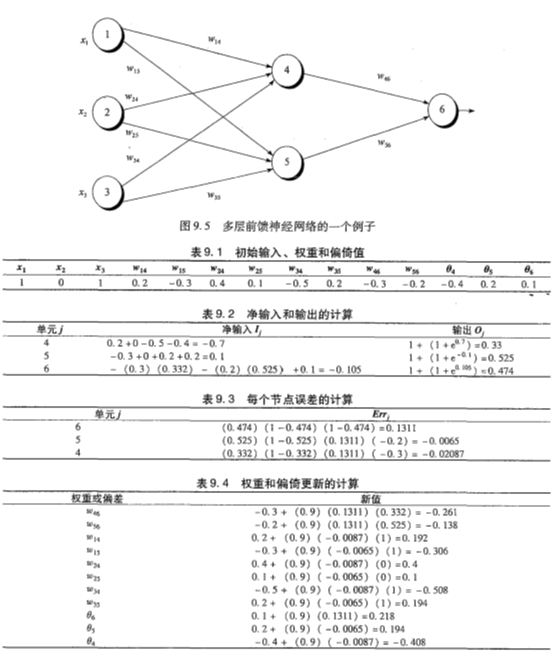
## (应用)朴素贝叶斯



第九章 分类：高级方法

## (应用)神经网络

神经网络是一组连接的输入/输出单元，其中每个连接都与一个权重相关联。在学习阶段，通过调整这些权重，使得它能够预测输入元组的正确类标号来学习。由于单元之间的连接，神经网络学习又称**连接者学习**。



**这个是一一对应的w4,6 A4的值乘以单元6的误差,一一对应**

## 支持向量机：

**一种对线性和非线形数据进行分类的方法，它是一种算法**，它使用一种非线形映射，把**原训练数据映射到较高的维上**。在新的维上，**它搜索最佳分离超平面“决策边界”**，使用到足够高维上的、合适的非线形映射，两个累的数据总可以被超平面分开。**使用支持向量（“基本”训练元组）和边缘（由支持向量定义）发现该超平面**。

## 惰性学习：

或指从**近邻学习**，当**给定一个训练元组**时，惰性学习法**先简单地存储它**，并且一致等待，**知道给定一个检验元组。仅当看到检验元组时，它才进行泛化**，以便**根据与存储的训练元组的相似性**对该元组进行**分类**。由于惰性学习法存储训练元或“实例”，它们也称基于实例的学习法。在做分类或数值预测时，惰性学习法的计算开销可能很大(因为需要存储的东西较多)。

## 遗传算法：

试图利用自然进化的思想。一般而言，遗传学习开始如下：**创建一个由随机产生的规则组成的初始群体**。每个规则可以用一个二进位串表示。**根据适者生存的原则(根据适应度/相似度)，形成新的群体**，它由**当前群体中最适合的规则以及这些规则的后代组成**。通常，规则的拟合度用它在训练样本集上的分类准确率评估**。后代通过使用诸如交叉和变异等遗传操作来创建**。在交叉操作中，来自规则对的**子串交换**，形成新的规则对。在变异操作中，**规则串中随机选择的位被反转**。

## 粗糙集：

粗糙集理论可以用于分类**，发现不准确数据或*噪声*数据内 的结构联系**。用于离散值属性，因此，连续值属性必须在使用前离散化。粗糙集理论基于给定训练数据内部的**等价类**的建立。形成一个等价类的所有数据元组是不加区分的；也就是说，对于描述数据的属性，这些样本是等价的。给定现实世界数据，通常有些类不能被可用的属性区分。粗糙集可以用来近似的或“粗略的”定义这些类。

## 模糊集：

也称可能性理论 ，做为传统二值逻辑和概率论的一种替代。它允许我们处理高层抽象，并且提供了一种处理数据的不精确测量的手段。最重要的是，模糊集理论允许我们处理模糊或不精确的事实。

区别:

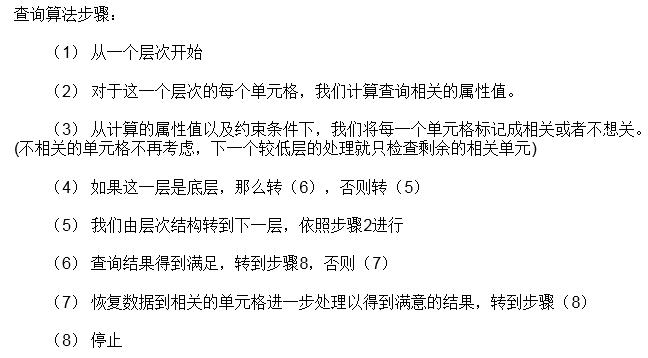
|  |  |  |
| --- | --- | --- |
|  | 粗糙集 | 模糊集 |
| 粗糙集与模糊集都能处理不完备（imperfect) 数据 | | |
| 方法不同 | 粗糙集则强调数据的不可辨别（indiscernibility），**不精确**(但是对)（imprecision) 和模棱两可（ambiguity). | 模糊集注重描述信息的**含糊（vagueness) 程度** |
| 例子 | 使用图像处理中的语言来作比喻，当论述图像的清晰程度时，粗糙集强调组成图像象素的大小， | 而模糊集则强调象素存在不同的灰度. |
| 侧重点 | **粗糙集研究的是不同类中的对象组成的集合之间的关系，重在分类。** |  |

第十章 聚类分析：基本概念和方法

## 1.聚类分析的基本概念

**聚类是一个把数据对象划分成子集的过程。每个子集是一个簇**，是的簇中**的对象彼此相似**，但与其他簇中的对象不相似。**由聚类分析产生的簇的集合称做一个聚类**。

## 2.聚类分析有哪些方法？

划分方法:层次方法，基于密度的方法，**基于网格的方法(下图)**

## 3. (应用)PAM中心点划分算法的具体步骤(看ppt)

**PAM，一种基于中心点或中心对象进行划分的k-中心点算法。**

**输入：**

* + k：结果簇的个数
  + D：包含n个对象的数据集合

**输出：**k个簇的集合

**方法：**

（1）从D中随机选择k个对象作为初始的代表对象或种子；

（2）repeat

（3） 将每个剩余的对象分配到最近的代表对象所代表的簇；

（4） 随机地选择一个非代表对象Orandom；

（5） 计算用Orandom代替代表对象Oj的总代价S；

（6） if S<0, then Orandom替换Oj，形成新的k个代表对象的集合；

（7）until不发生变化；

第十一章 高级聚类分析

## 高级聚类分析的方法及其含义

* + **基于概率模型的聚类**：某些应用中需要模糊或灵活的簇指派，每个数据对象可能属于多个簇()。**模糊簇**，**基于概率模型的聚类，期望最大化算法。**
  + **基于高维数据的聚类**：大多聚类方法**在维度不高时，少于10个属性，运行良好**。然而，**存在一些重要的高维应用，需要在在高维数据上进行聚类分析**。
  + **基于聚类图和网络数据的聚类：**在图和网络数据上的聚类分析题去有价值的知识和信息。术语“图”和“网络”可以互换地使用。
  + **具有约束的聚类**：通常，用户具有背景知识希望把它们集成到聚类分析中，可能还会有一些特定应用的要求，这些信息可以作为聚类约束来建模。

第十二章 离群点检测

## 离群点概念

假定使用一个给定的统计过程来产生数据对象集。离群点是一个数据对象，它显著不同于其他数据对象，好像它是被不同的机制产生的一样。

## 离群点的类型

全局离群点：数据对象显著的偏离数据集中的其余对象

情景离群点：如果数据对象在给定特定情景下，显著的偏离其它对象

集体离群点：数据对象的**某个子集显著偏离整个数据集离群点检测方**法

**离群点的检测方法**

**监督、半监督和无监督方法**

**统计方法、基于邻近性的方法和基于聚类的方法:**

**1.统计学方法：参数方法，非参数方法**

**对给定的数据集合假设了一个分布或概率模型(例如, 正态分布), 然后根据模型采用不一致性检验来确定离群点 .**

**检验要求的参数:数据集参数: 例如, 假设的数据分布,**

**分布参数: 例如平均值和方差和预期的离群点的数目)**

**参数的方法：假设数据服从特定分布，分布的参数通过最大似然估计得到**

**非参数方法：够造直方图，检测数据是否落入直方图的某一箱中**

**2.基于邻近性方法：基于距离的离群点检测和嵌套循环方法，基于网格的方法，基于密度的离群点检测**

**3.基于聚类的方法**