Contents

[<数据挖掘导论> 1](#_Toc453676332)

[数据 2](#_Toc453676333)

[数据类型 2](#_Toc453676334)

[数据质量 2](#_Toc453676335)

[预处理 3](#_Toc453676336)

[相似性和相异性的度量 4](#_Toc453676337)

[分类 6](#_Toc453676338)

[关联分析 6](#_Toc453676339)

[聚类 6](#_Toc453676340)

[异常检测 6](#_Toc453676341)

# <数据挖掘导论>

抛开理解数据的人和数据所处的领域而简单地使用数据分析技术是不可行的

数据挖掘技术可以用来支持广泛的商务智能应用，如顾客分析、定向营销、工作流管理、商店分布和欺诈检测等

数据挖掘技术用来探查大型数据库，发现先前未知的有用模式，预测未来观测结果

信息检索：使用数据库管理系统查找个别的记录，或通过因特网的搜索引擎查找特定的web页面。它们主要依赖传统的计算机科学技术和数据的明显特征来创建索引结构，从而有效地组织和检索信息

输入数据存储：可以以各种形式存储如平展文件、电子数据表或关系表，并且可以驻留在集中的数据存储库中，或分布在多个站点上。

数据预处理：融合来自多个数据源的数据，清洗数据以消除噪声或重复的观测值，选择与当前数据挖掘任务相关的记录和特征

可视化：从各种不同的视角探查数据和数据挖掘结果。在后处理阶段，还能使用统计度量或假设检验，删除虚假的数据挖掘结果

数据挖掘的难点：

可伸缩：处理海量数据集，算法必须是可伸缩的。许多数据挖掘算法使用特殊的搜索策略处理指数级搜索问题

高维性：成百上千属性的数据集

异种数据和复杂数据

数据挖掘与传统统计方法的比较？

传统的统计方法基于一种假设检验模式，即提出一种假设，设计实验来收集数据，然后针对假设分析数据。当前的数据分析任务常常需要产生和评估数千种假设，因此需要自动地产生和评估假设。

数据挖掘所分析的数据集通常不是精心设计的实验的结果，并且它们通常代表数据的时机性样本（opportunistic sample），而不是随机样本(random sample).

数据挖掘来自如下领域思想：1统计学的抽样、估计和假设检验。2人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。3最优化、进化计算、信息论、信号处理、可视化和信息检索

数据挖掘任务：

1预测任务

以说明变量函数的方式为目标变量建立模型，有两类预测建模任务：分类（用于预测离散的目标变量）和回归（用于预测连续的目标变量）。目标都是训练一个模型，使目标变量预测值与实际值之间的误差达到最小。常见的应用：检查结果判断病人是否患有某种疾病。

2描述任务（其目标是导出概括数据中潜在联系的模式（相关、趋势、聚类、轨迹和异常）。本质上，描述性数据挖掘任务通常是探查性的，并且常常需要后处理技术验证和解释结果。

## 数据

### 数据类型

决定我们应使用何种工具和技术来分析数据。

知道属性的类型是重要的，因为它告诉我们测量值的哪些性质与属性的基本性质一致，从而使得我们可以避免诸如计算雇员的平均ID这样的愚蠢行为

不同的属性类型

分类的(定性的)

标称 区分对象(=, ~=) 一对一变换

序数 确定对象的序(>, <) 值的保序变换，即新值=f(旧值)，f是单调函数

数值的（定量的）

区间 值之间的差(+, -) 新值=a\*旧值+b

比率 差和比率(\*, /) 新值=a\*旧值

数据集特性：维度，稀疏性和分辨率（常常可以在不同的分辨率下得到数据，并且在不同的分辨率下数据的性质也不同，即数据的模式也依赖于分辨率，如果分辨率太高，模式可能看不出，或者掩埋在噪声中；如果分辨率太低，模式可能不出现）

### 数据质量

如存在噪声和离群点，数据遗漏、不一致或重复，数据有偏差或者不能代表它应该描述的现象或总体情况

数据挖掘常常不能“在数据源头控制质量）。相比之下，统计学的实验设计或调查往往其数据质量都达到了一定的要求。由于无法避免数据质量问题，因此数据挖掘着眼于两个方面：1数据质量问题的检测和纠正.2使用可以容忍低质量数据的算法。

测量和数据收集问题

1. 测量误差和数据收集错误
2. 噪声和伪像

噪声通常用于包含时间或空间分量的数据。在这些情况下，常常可以使用信号或图像处理技术降低噪声，从而帮助发现可能”淹没在噪声中“的模式（信号）

伪像：数据错误可能是更确定性现象的结果，即数据的某种确定性失真。如一组照片在同一地方出现条纹

1. 精度、偏倚和准确率
2. 离群点

在某种意义上具有不同于数据集中其他大部分数据对象的特征的数据对象，或是相对于该属性的典型值来说不寻常的属性值。Outlier可以是合法的数据对象或值。因此不像噪声，离群点本身有时是人们感兴趣的对象

1. 遗漏值

处理遗漏值的策略：1删除数据对象或属性2估计遗漏值。如果属性是连续的，则可以使用最近邻的平均属性值；如果属性是分类的，则可以取最近邻中最常出现的属性3在分析时忽略遗漏值。如果某对的一个对象或两个对象都有某些属性有遗漏值，则可以仅使用没有遗漏值的属性来计算相似性

1. 不一致的值
2. 重复数据

数据集 = 数据对象的集合

数据对象 = {属性：值}

记录数据

数据表病人信息

事务数据如购物蓝信息

数据矩阵如特征空间

文档-词矩阵

序列数据 = 记录数据 + 时间信息（或空间信息）

时间自相关：分析时间数据时，如果两个测量的时间很接近，则这些测量的值通常非常相似

空间自相关：物理上靠近的对象趋向于在其他方面也相似。如地球上相互靠近的两个点通常具有相近的气温和降水量

### 预处理

原始数据必须加以处理才能适合于分析。如将连续值属性如长度转换成具有离散的分类值的属性如短中长；如数据集属性的数目常常需要减少，因为属性较少时许多技术用起来更加有效

* 聚集

将多个对象合并成单个对象，聚集是删除属性（如商品类型）的过程，或者是压缩特定属性不同值个数的过程

好处：1数据归约导致较小数据集需要较少的内存和处理时间2通过高层而不是低层数据视图，聚集起到了范围或标度转换的作用3对象或属性群的行为通常比单个对象或属性的行为更加稳定

缺点：丢失有趣的细节

* 抽样

统计学使用抽样是因为得到感兴趣的整个数据集的费用太高、太费时间；

数据挖掘使用抽样是因为处理所有的数据的费用太高、太费时间

有效抽样：样本是代表性的，前提是它近似地具有与原数据集相同的感兴趣的性质。如果数据对象的均值是感兴趣的性质，而样本具有近似于原数据集的均值，则样本就是有代表性的。我们所能做的最好的抽样方案就是选择一个确保以很高的概率得到有代表性的样本

抽样方法

无放回抽样和有放回抽样。当样本与数据集相比相对较小时，两种方法产生的样本差别不大。但是对于分析，有放回抽样较为简单，因为在抽样过程中，每个对象被选中的概率保持不变

分层抽样：从预先指定的组开始抽样。在最简单的情况下，尽管抽样的大小不同，但是从每组抽取的对象个数相同。另一种变形是从每一组抽取的对象数量正比于该组的大小。

渐进抽样（自适应抽样）：从一个小样本开始，然后增加样本容量直至得到足够容量的样本。尽管这种技术不需要在开始就确定正确的样本容量，但是需要评估样本的方法，确定它是否足够大。尽管预测模型的准确率随样本容量增加，但是在某一点准确率的增加趋于确定

* 维归约

好处：1维归约可以删除不相关的特征并降低噪声，一部分是因为维灾难。2维归约可以使模型更容易理解，因为模型可能只涉及较小的属性3数据也可通过观察属性或对三元组属性达到可视化4使用维归约降低了数据挖掘算法的时间和内存需求

维灾难：随着数据维度的增加，许多数据分析变得非常困难。特别是随着维度增加，数据在它所占据的空间中越来越稀疏。对于分类，这可能意味着没有足够的数据对象来创建模型，将所有可能的对象可靠地指派到一个类。对于聚类，点之间的密度和距离的定义失去了意义

最常用的方法：线性代数技术，将数据由高维空间投影到低维空间。如主成分分析PCA:找出新的属性（主成分），这些属性是原属性的线性组合，是相互正交的，并且捕获了数据的最大变差；如奇异值分解

* 特征子集选择

看起来这种方法可能丢失信息，但是在存在冗余或不相关特征的时候，情况并非如此。冗余特征重复了包含在一个或多个其他属性中的许多或所有信息。不相关特征包含对于手头的数据挖掘任务几乎完全没用的信息。冗余和不相关的特征可能降低分类的准确率，影响所发现的聚类的质量

* 特征创建

特征提取：对数据进行处理，获得一些较高层次的特征。最常使用的特征提取技术都是高度针对具体领域

映射数据到新的空间：使用一种完全不同的视角挖掘数据可能揭示出重要和有趣的特征。如果有大量周期模式，并且存在大量噪声，则很难检测这些模式。尽管如此，通过对该时间序列实施傅里叶变换，将它转换成频率信息明显的表示，就能检测到这些模式

特征构造：使用专家的意见构造特征

* 离散化和二元化

非监督离散化

用于分类的离散化方法之间的根本区别在于使用类信息还是不使用类信息。如果不使用类信息，则常使用一些相对简单的方法，如等宽，等频

监督离散化

如熵entropy,区间的熵是区间纯度的度量。如果一个区间只包含一个类的值（该区间非常纯），则其熵为0并且不影响总熵。如果一个区间中的值类出现的频率相等（该区间心可能不纯），则其熵最大。

* 变量变换

在统计学中，变量变换（特别是平方根、对数和倒数变换）常用来将不具有高斯（正态）分布的数据变换成具有高斯（正态）分布的数据。

使用变量变换时需要小心，因为它们改变了数据的特性。

标准化或规范化：目标是使整个值的集合具有特定的性质。统计学中术语规范化可能与使变量正态（高斯）的变换相混淆。对变量标准化是为了避免具有较大值域的变量左右计算结果。均值和标准差受离群点的影响很大，因此通常需要修改，用中位数取代均值，用绝对标准差取代标准差

### 相似性和相异性的度量

邻近度：表示相似性或相异性，由于两个对象之间的邻近度是两个对象对应属性之间的邻近度的函数

相似度：[0, 1]

相异度：>0

相关性：[-1, 1]

如果邻近度度量原来在区间[0, ]上取值，则需要使用非线性变换，d’ = d/(d+1)

任意单调减函数都可以用来将相异度转换到相似度（或相反）

#### 简单属性的相异度

标称属性：

序数属性：

, 值映射到整数0~n-1,其中n是值的个数

区间或比率的属性：

d =|x-y|

数据对象之间的相异度

Minkowski distance, 见page42

#### 记录之间的相似性度量

对称二元属性：简单匹配系数simple Matching Coefficient

SMC = 值匹配的属性个数/属性个数

SMC可以在一个仅包含是非题的测验中用来发现回答问题相似的学生

非对称二元属性：Jaccard系数，忽略0-0度量，因为稀疏性大多0

J = 匹配的个数/不涉及0-0匹配的属性个数

多元属性：余弦相似度

余弦相似度实际上是x和y之间夹角（余弦）的度量，如果余弦相似度为1,则x and y之间夹角是0.余弦相似不考虑两个数据对象的量值。

#### 记录之间的相关性度量

Person’s correlation:

连续变量：Corr(x,y) = covariance(x, y) / (standard\_deviation(x) \* standard\_deviation(y))

二元变量：

is defrieved from Pearson’s chi-squared test

如果相关度为0,则两个数据对象的属性之间不存在线性关系。然而仍然可能存在非线性关系。

#### 邻近度计算

若属性不同尺度，但属性之间独立，只需要对变量标准化，然后计算欧几里得距离

若属性不同尺度，且属性之间相关，需要计算Mahalanobis距离

数据探索

汇总统计

频率：frequency(vi) = 具有属性值vi的对象数/m

众数: 最高频率值

百分位数

位置度量：均值和中位数

均值对离群值很敏感，对于包含离群值的数据，中位数可以更稳健地提供值集中间的估计

散布度量：极差和方差 (表明属性值是否散布很宽，或是否相对集中在均值附近

极差：range(x) = max(x)-min(x)

标准差：与均值差平方和除以(m-1) = 方差的平方根

因为方差用均值计算，因此它也对离群值敏感。

对于多元数据，数据的散布用协方差矩阵表示。协方差矩阵的对角线上是属性的方差。

两个属性的协方差是两个属性一起变化并依赖于变量大小的度量。协方差的值接近

可视化

数据对象、它们的属性，以及数据对象之间的联系要转换成诸如点、线、形状和颜色等图形元素。

Stem and leaf plot

Histogram

2D historgram

Box plot

Pie chart

Scatter plot

散布图的作用：图形化显示两个属性之间的关系。当类标给出时，可以使用散布图考察两个属性将类分开的程度。

Contour plot （等高线图）

Surface plot （曲面图）

联机分析处理(OLAP)

在不同的维上或不同的属性值上聚集数据

多维数据分析的关键目标是观察聚集量，如总和或平均值

## 分类

基本概念

应用场景：根据电子邮件的标题和内容检查出垃圾邮件；根据核磁共振扫描的结果区分肿瘤是恶性的还是良性的？

分类任务的输入数据是样例的集合，用元组(X, y)表示，其中X是属性的集合，y是样例类标号。类桔必须是离散属性。

区别分类与回归？回归也是一种预测建模任务，其中目标属性是连续

分类任务是通过学习得到一个目标函数f,把每个属性集X映射到一个预先定义的类标号y. 目标函数也称为分类模型

分类模型的有两个目的：

描述性建模

分类模型可以作为解释性的工具，用于区分不同类中的对象

预测性建模

分类模型还可以用于预测未知记录的类标号

分类技术非常适合预测或描述二元或标称类型的数据集，对于序数分类（如把人分类为高收入，中等收入或低收入组），分类技术不太有效，因为分类技术不考虑隐含在目标类中的序关系。

分类法：根据输入数据集建立分类模型的系统方法。如决策树，基于规则，神经网络，支持向量机和贝叶斯。这些技术都使用学习算法确定分类模型，该模型能很好地拟合输入数据中类标号和属性集之间的联系。学习算法很到的模型不仅要很好地拟合输入数据，还要能够正确地预测未知样本的类标号。

决策树

决策树是一种由结点和有向边组成的层次结构

在决策树中，每个叶结点都赋予一个类标号。非终结点（包含根结点和内部结点）包含属性测试条件，用以分开具有不同特性的记录。

如何建立决策树？

原则上讲，对于给定的属性集，可以构造的决策树的数目达指数级。由于搜索空间是指数规模的，找出最佳决策树在计算上是不可行的。所以算法的目的是能够在合理的时间内构造出具有一定准确率的局部最优决策树

Hunt算法是许多决策树算法的基础，如CART

1. 假定分类问题的初始决策树只有一个结点，因为根结点包含超过一个类的记录，所以需要进一步的细化
2. 根据一个属性测试条件，数据集被划分为较小的子集
3. 对根结点的每个子女递归地调用Hunt算法

核心：

如何选择一个属性测试条件，作为划分数据的最优标准？也就是说，算法必须提供为不同类的属性指定测试条件的方法，并且提供评估每种测试条件的客观度量

选择最佳划分的度量通常是：根据划分后子女结点不纯性的程度。不纯的程度越低，类分布就越倾斜。均衡分布(0.5, 0.5)的结点具有最高的不纯性

其中t是给定结点t的总数

其中表示给定结点t中属性类i的记录所占的比例

例子：

如何用决策树分类法区分正常的用户访问，还是由Web爬虫产生的访问？

输入数据：Web服务器日志

Web日志记录的字段包括客户端的IP地址、请求的时间戳、请求访问的文档的网址、文档的大小、客户的身份（通过用户代理字段获得）。

Web会话是客户在一次网站访问期间发出的请求序列，每个Web会话都可以用有向图来建模，其中结点对应于网页，而有向边对应于连接网页的超链。

目标：对Web会话进行分类

特征构建：构造描述每次会话特性的特征，如遍历的深度和宽度…

训练的模型：（作为描述性模型）

* Web机器人的访问倾向于宽而浅，而正常用户访问比较集中（窄而深）
* 与正常用户不同，Web机器人很少访问与Web文档相关的图片页
* Web机器人的会话的长度趋于较长，包含了大量请求页面
* Web机器人更可能对相同的文档发出重复的请求，因为正常用户访问的网页常常会被浏览器保存

决策树归纳的特点：

* 非参数方法，也就是说，它不要求任何先验假设，不假定类和其他属性服从一定的概率分布。
* NP完全问题，一般采用启发式方法指导对假设空间的搜索
* 容易解释
* 对噪声的干扰具有相当好的鲁棒性
* 冗余属性（一个属性在数据中与另一个属性是强相关的）不会对决策树的准确率造成不利的影响。但是如果数据集中含有很多不相关的属性（即对分类任务没有用的属性），则某些不相关属性可能在树的构造过程中偶然被选中，导致决策树过于庞大。
* 大多数的决策算法采用自顶向下的递归划分方法，因此沿着树向下，记录会越来越少。在叶结点，记录可能太少，对于叶结点代表的类，不能做出具有统计意义的判决，这就是所谓的数据碎片问题。

决策树的物理解释：

可以将决策树的生长过程看成划分属性空间不相交的区域的过程，直到每个区域都只包含同一类的记录。由于测试条件只涉及单个属性，因此决策边界是直线，即平行于坐标轴，这就限制了决策树对连续属性之间复杂关系建模的表达能力。

模型存在的一般问题？

过分拟合

决策树过大容易受所谓过分拟合(overfitting)现象的影响。通过修剪初始决策树的分支，剪枝有助于提高决策树的泛化能力。模型越是复杂，出现过分拟合的几率就越高，可以采用奥卡姆剃刀原则：给定两个具有相同的泛化误差的模型，较简单的模型比较复杂的模型更可取

过分拟合会导致低训练误差，但高泛化误差。

两种可能：噪声导致的过分拟合，缺乏代表性样本导致的过分拟合

拟合不足

决策树过小会导致高训练误差和高泛化误差。因为模型尚未学习到数据的真实结构。

分类器性能的评估

K折交叉验证，留一法

## 关联分析

用来发现描述数据中强关联特征的模式。由于搜索空间是指数规模的，关联分析的目标是以有效的方式提取最有趣的模式。常见的应用：具体有相关功能的基因组、识别用户一起访问的web页面、理解地球气候系统不同元素之间的联系等

## 聚类

发现紧密相关的观测值结果。常见应用：对相关的顾客分组、找出显著影响地球气候的海洋区域以及压缩数据等。

## 异常检测

识别其特征显著不同于其他数据的观测值。常见应用：检测欺诈、网络攻击、疾病的不寻常模式、生态系统扰动等。