第一章 数据挖掘概论

数据挖掘概论

- 大数据?
- 数据挖掘?
- 大数据思维
- 数据挖掘应用

大数据的三个层面

理论、技术、实践

什么是数据?

> 数据是事实或观察的结果，是对客观事物的逻辑归纳，适用于表示客观事物的未经加工的原始素材。（原料）
>
> 数据是可定量分析的记录

三元空间世界

世界的演化发展经历了”一元空间“、”二元空间“，正在向”三元空间“发展:

- 人类社会诞生之前，世上只有物理空间（一元）
- 人类社会形成和发展，产生了社会空间（二元）
- 人类社会进入信息社会，正逐步形成数据空间（三元）；数据成为物质和能源之外的新型能源

人类对传统二元空间的认知，形成了自然科学、社会科学，对数据空间的认识将逐渐形成“**数据科学/计算科学**”

什么是大数据?

大数据(Big Data)是指“无法用现有的软件工具提取、存储、搜索、共享、分析和处理的海量的、复杂的数据集合。”

大数据思维

大数据——认识和改造世界的第四范式

数据的本质是生产资料和资产

数据不在是社会生产的“副产物”，而是可被二次乃至多次加工的原料。从中可以探索更大价值。它变成了生产资料。

大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产，大数据就是“未来的新石油”。

大数据的四大特性

- Volume：数据量特别大
    - 非结构化数据的超大规模和增长
    - 总数据量的80%~90%

    - 比结构化数据增长快10倍到50倍
    - 是传统数据仓库的10倍到50倍
- Variely：数据多样化
    - 大数据的异构和多样性
    - 很多不同样式（文本、图像、视频、及其数据）
    - 无模式或者模式不明显
    - 不连贯的语法或句义
- Velocity：数据处理速度快
    - 大量的不相关信息
    - 对未来趋势与模式的可预测分析
    - 深度复杂分析（机器学习、人工智能 VS 传统商务智能）
- Value：数据处理价值密度低
    - 实时分析而非能量式分析
    - 数据输入、处理与丢弃
    - 立竿见影而非事后见效

复杂性（Complexity）

大数据的数据度量?

1 Byte	**8 bit**
1 KB	**1024 Bytes**
1 MB	**1024 KB**
1 GB	**1024 MB**
1 TB	**1024 GB**
1 PB	**1024 TB**
1 ZB	**1024 PB**
1 YB	**1024 ZB**
1 BB	**1024 YB**
1 NB	**1024 BB**
1 DB	**1024 NB**

> 1 PB 相当于50%的全美学术研究图书馆藏书信息内容
>
> 5 EB 相当于至今全世界人类所讲过的话语
>
> 1 ZB 如果全世界海滩上的沙子数量总和
>
> 1 YB 相当于 7000 位人类体内的微细胞总和

大数据的构成

大数据 = 海量数据 + 复杂类型的数据

- 海量交易数据（交易）
- 海量交互数据（人的交互行为）
- 海量数据处理（物联网）

> 所谓的大数据思维，是指一种意识、认为公开的数据一旦处理得当就能为千百万人急需解决的问题提供答案。——《大数据时代》

数据挖掘是一个新兴、交叉学科领域

从大量的数据中挖掘那些令人**感兴趣**的、**有用**的、**隐含**的、**先前未知**的和**可能有用的模式或知识**。

数据挖掘的主要功能

- 特异群组分析：特异对象的数据分析
- 关联分析：寻找数据中的关联和相关性
- 分类分析：将数据分成不同群组，群组之间差异明显
- 聚类分析：将数据分成不同群组，群组之间差异明显
- 孤立点分析：检测和分析异常数据
- 演变分析：分析随时间变化研究对象的发展规律或趋势

数据+挖掘

数据->筛选->预处理->变换->数据挖掘->解释评价

问题定义->数据收集参与处理->数据挖掘算法执行->结果解释和评估

数据挖掘主要步骤

- 数据清理（消除噪音或不一致数据）
- 数据集成（多种数据源可以组合在一起）
- 数据选择（从数据库中提取与分析任务相关的数据）
- 数据变换（数据变换或统一成适合挖掘的形式；如：通过汇总或聚集操作）
- 数据挖掘（基本步骤，使用智能方法提取数据模式）
- 模式评估（根据某种兴趣度度量，识别提供知识的真正有趣的模式）
- 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）

大数据核心技术的三大模块

- **大数据处理与采集**
- **大数据分析**
- **大数据存储与管理**

未来热点应用领域：

- 网络（Web）数据挖掘：从大量的、含噪声的、无结构化的网络数据汇总提取出隐藏在背后的、有价值的知识
- 文本挖掘：抽取有效、新颖、有用、可理解的、散步在文本文件中的有价值知识，并且利用这些知识更好地组织信息的过程
- 社交网络分析：以人物为节点，以人际关系为边，将人物节点联结起来而构成的网络
- 生物信息大数据分析：最重要的任务是从海量的数据中提取**新知识**（例如：全基因组分析、表现组分析、转录组分析、单细胞RNA测序分析、蛋白质结构和特性预测……）
- 医学大数据挖掘：大数据的分析和挖掘在医学领域的应用包含很多的方向，比如临床操作的比较效果研究，临床决策支持系统，医疗数据透明度，远程病人监控……）
- 金融大数据挖掘
- 交通出行大数据挖掘
- 生活消费大数据挖掘
- 气象大数据挖掘
- 能源大数据挖掘
- 流行病大数据挖掘
- ……

第二章 数据预处理

2.1 为什么要进行数据预处理

现实世界的数据是脏的!

由于现实生产和实际生活以及科学研究的多样性、不确定性、复杂性等，导致采集到的原始数据比较散乱，它们是不符合挖掘算法进行知识获取研究所要求的规范和标准!

大数据主要具有以下特征

- 不完整性：因为一些人为因素造成的数据属性值的丢失或不确定的情况。
- 含噪声：数据中存在错误和异常属性值（偏离期望值）
- 杂乱性（不一致性）：数据源不同，在编码或命名上有差异

没有高质量的数据，就没有高质量的数据挖掘结果!

2.2 数据预处理常见方法

- 数据清洗：去噪声 ，纠正不一致，实现数据标准化
- 数据集成：合并来自多个数据源的数据集，形成一个新综合数据集
- 数据变换：把原始数据转换成统一为适合挖掘的形式
- 数据归纳：寻找依赖于目标数据的有用特征。缩减数据规模最大限度的精简数据量

一、数据清洗

1. 空缺值处理方法

- 直接忽略掉空缺值的整个属性或元组（但当某类属性空缺值所占百分比很大时，其效果非常差！）
- 人工填写缺失值（工作量大，可行性低）
- 用属性的均值填充缺失值（同组缺失数据由同一值填补，结果扭曲了目标属性的分布）
- 用同类样本的属性均值填充缺失值（适用于分类数据挖掘）
- 数学模型填补法（利用回归、贝叶斯计算公式或决策树归纳来确定缺失值）

2. 噪声处理方法

1. 分箱

   分箱是指把待处理的数据按照一定规则放进“箱子”中，采用某种方法对各个箱子中的数据进行处理。

   - 等深分箱法

     每箱具有相同的记录数，每个箱子的记录数称为箱子的深度。

   - 等宽分箱法

     在整个属性值区间上平均分布，即每个箱子的区间范围设定为一个常量，称为箱子的宽度。

2. 平滑处理

- 首先排序数据，并将他们分到等深的箱中
- 然后按可以按箱的平均值平滑，按箱中值平滑，按箱边界平滑等
  - 按箱的平均值平滑：箱中每一个值被箱中的平均值替换
  - 按箱中值平滑：箱中每一个值被箱中的中值替换
  - 按箱边界平滑：箱中的最大和最小值被视为箱边界，箱中每一个值被最近的边界值替换

3. 聚类

- 将数据集合分组为若干个簇，在簇外的值即为孤立点，这些孤立点就是噪声数据，应对这些孤立点进行删除或替换。相似或相临近的数据聚合在一起形成各个聚类集合，在这些聚类集合之外的数据即为异常数据。

4. 回归

- 通过发现两个相关的变量之同的相关关系，构造一个回归函数，使得该函数能够更大程度地满足两个变量之间的关系、使用这个函数来平滑数据。

二、数据集成

- 数据集成
  - 将多个数据源中的数据整合到一个一致的存储中
  - 这些源可以是多个数据库，数据立方体或一般文件

数据集成中需要注意的问题：模式匹配、数据冗余、数据值冲突

- 模式匹配
  - 整合不同数据源中的元数据
  - 进行实体识别
- 数据冗余

- 同一属性值不同的数据库中会有不同的字段名
- 有些冗余可以被相关分析检测到

$$r_{A,B} = \frac{\sum(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B}$$

如果 A 、B 间具有较高的相关系数，表明 A 或 B 可以作为冗余而去掉。

- 除了检查属性是否冗余外，还要检查记录行的冗余。