3-4-cn
1
本节我们讨论数据预处理中的数据转换相关问题
2
数据转换可以包括数据集成和数据转换。 我们先来看看数据集成。
3
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
模式匹配
数据冗余
数据值冲突
4
数据集成中第一个是模式匹配。 集成来自不同数据源的元数据。我们可以从不同的数据源识别现实世界的实体并将它们映射在一起，
例如 A.cust-id=B.customer_no。
5
第二个是数据集成 **中的**数据冗余，同一属性在不同的数据库中会有不同的字段名。
一个属性可以由另外一个属性导出。如：一个顾客数据表中的平均月收入属性，它可以根据月收入属性计算出来。
有些冗余可以被相关分析检测到
6
第三个 数据集成**中的问题是**数据值冲突
对于现实世界的实体，其来自不同数据源的属性值可能不同。 比如表示的差异，不同的尺度，或者编码的差异等.
例如:
重量属性在公制系统，使用公斤、克，在英制系统，使用 pound 磅。
不同地点的相同价格属性使用不同的货币单位，美元、英镑、人民币
7
数据转换可以包括数据集成和数据转换。 我们再来看看数据转换
8
数据转换是指，为了便于高效分析，我们需要将数据从一种形式转换为另一种形式。
比如我们可以使用像 Binning Clustering Regression 这样的平滑方法来消除噪声
或离散化连续数据，并增加粒度。通过这样做，我们可以减少进一步分析的数据量。
9
对数据进行汇总
avg(), count(), sum(), min(), max()…
例如：每天销售额（数据）可以进行合计操作以获得每月或每年的总额。
可以用来构造数据立方体
10
第三种数据转化技术-数据泛化
用更抽象（更高层次）的概念来取代低层次或数据层的数据对象
例如：街道属性，就可以泛化到更高层次的概念，诸如：城市、国家。同样对于数值型的属性，如年龄属性，就可以映射到更高层次概念，如：年轻、中年和老年。
11
第四种数据转换技术，数据规范化 Normalization
将数据按比例进行缩放，使之落入一个特定的区域，以消除数值型属性因大小不一而造成挖掘结果的偏差。如将工资收入属性值映射到[-1.0,1.0]范围内。

方法：
（1）最小–最大规范化
（2）零–均值规范化（z–score 规范化）
（3）小数定标规范化
12
第五个转变是属性构建。利用已有的属性集构造新的属性并将其添加到已有的属性集中，有助于挖掘更深层次的模式知识，提高挖掘结果的准确性。
例如：根据 width 和 height 属性，可以构造一个新的属性：area。
13
本节课我们概要的介绍了数据集成和数据转换，本节课就到这里，谢谢大家