

3-1-cn 数据预处理概述

1

本节我们讨论数据预处理相关知识

2 数据预处理主要包括数据清洗、数据转换和数据归约

3

为什么我们需要进行数据清洗？因为当我们从不同的数据资源中收集数据时，结果中难免会有一些脏数据。

我们来看一个例子，在这个表中，在第二行，对于记录 id=2，缺少年龄属性。

如果我们计算平均年龄，这个缺失的年龄值会严重影响平均结果。

重复第三和第四条记录。

第六条记录是冲突记录，根据年龄和出生数据。

第七个记录是根据出生数据的异常年龄 101。所有这些脏数据都会破坏分析结果。

4

有一个俗语说，只要事情有可能出错，它就一定会出错。

5

如何防止脏数据出现

1) 制定数据标准：

统一多个数据源的属性值

编码尽可能清楚地给出属性名称和属性值

2) 优化系统设计：

尽可能对键值使用选项，而不是手动填写条目。

重要属性出现在显眼位置，使用必填选项

应修改异常值

6

数据预处理可以包括数据清理；数据转换和数据缩减

1) 数据清洗包括缺失数据处理和噪声数据处理。

1.1 对于缺失数据我们可以直接忽略元组，或者记录。或者我们可以通过平均值或最可能值手动填充缺失值。

1.2 消除噪声数据的方法包括分箱法、回归法和聚类法。

2) 至于数据转换，它包括规范化、属性选择、离散化和概念层次生成。

3) 数据预处理的第三个方面是数据缩减。数据约简包括数据立方体聚合、属性子集选择、数量约简（即减少数据量）和降维。

4) 数据整合：将所有数据存储在数据库、数据仓库或文件中，形成完整的数据集。

5) 数据规格：剔除不能描述系统关键特征的数据属性，只保留能够描述关键特征的数据属性集的一部分以及其他…

7

数据清洗技术的研究最早是从纠正美国的社会保险号开始的。后来随着信息也和商业的快速发展，加速了这方面技术的研究，内容包括

1) 对数据集进行异常检测，通常采用统计方法，检测数据的数值型属性，通过计算属性值的均值和标准差等指标，在每一个属性的置信区间内识别异常的属性和记录。

2) 对数据进行去重：数据去重的过程就是重复清洗数据记录的过程。这个过程在数据仓库应用中特别重

要，因为在集成来自不同数据源的数据时，可能产生大量的重复数据记录。

3) 对缺失数据的清洗：大多采用近似值替换缺失值的方法对数据进行清洗，得到近似值的方法包括贝叶斯网络、神经网络、KNN 分类、粗糙集理论等，这些方法的核心就是判断缺失记录与其他完整记录之间的相似度。

国内这个方面的研究还处于初始阶段，数据清洗工作主要集中在银行，保险和证券等对客户数据的准确性要求很高的行业。这些行业只做针对自己客户进行数据清洗的工作，而且只开发针对具体应用的软件没有通用产品

8

对海量数据进行分析挖掘需要很长时间，为了让数据挖掘更加有效，需要对数据进行规约。数据规约的主要研究内容：

1) 高纬度数据的降维处理

该过程主要采用删除冗余数据属性的方法，且删除冗余数据属性往往需要用到某领域的业务知识，常用的降维方法包括逐步向前选择法、逐步向后删除法，判定树归纳法等

2) 减少数据量

当处理大量数据需要花费较长时间时，无法满足某些实时性要求较高的应用要求，此时，需要对数据量进行缩减。此过程采用的主要方法，包括直方图，聚类等，进而从数据集中选择较小规模的数据

3) 数据离散化技术

该技术可以将连续属性转换为离散属性值，降低属性值得个数，从而降低处理数据的运算时间

9

本节课我们概要的介绍了数据预处理，今天就学习到这里，谢谢大家