北京大学信息管理系

《数据挖掘导论》讲义

第二章 数据准备

北京大学信息管理系

2016 年秋

目录

第二章 数据准备

2.1 数据类型

数据的区分方法有很多，按照数据基本性质划分，可分为标称（nominal）、序数（ordinal）、区间（interval）和比率（ratio）；按照可能取值的个数判断，可分为离散（discrete）和连续（continuous）。

2.1.1 离散型数据

离散型数据具有有限或无限可数个值，例如顾客的唯一标识号 customer_ID 是无限可数的，因为顾客的数量可以无限增长（虽然事实上实际值的集合是可数的）；而一个国家的邮政编码是有限个值。一般来说，离散型数据可对应数据的标称属性和序数属性。

- 标称属性：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，并且不必具有有意义的序。因此标称属性又被看做是分类的，它只提供足够的信息来区分对象。例如头发颜色 hair_color 的可能值为黑色、棕色、黄色、红色、白色；职业 occupation 的可能值为教师、医生、警察等。标称属性除了可以用文字来表示，还可以用数字，例如刚才的 hair_color 属性，我们可以用 0 代表白色，1 代表黑色；代表顾客唯一标示的 customer_ID 的取值可能也都是数字，但这些数字并没有数学上的运算意义。
- 二元属性是一种特殊的标称属性，只有两个类别或状态：一般可以用 0 或 1 表示，0 代表不具备该属性，1 表示具有该属性。一个二元属性如果是对称的，说明它的两种状态具有同等价值，即哪个结果应该用 0 或 1 并无偏好，例如代表性别 gender 的男、女属性；如果是非对称的，说明不同状态的结果不是同等重要，例如艾滋病病毒化验的阳性和阴性结果，我们一般更注重于化验结果为阳性的人群（即稀有结果的出现），因此一般用 1 来表示相对重要的结果，0 表示另一个。
- 序数属性：同标称属性一样都是将数据进行分类，但它的取值之间具有有意义的序。例如成绩 grade 的取值可能为优、良、可、差；教师职称 professional_rank 可分为教授、副教授、讲师、助教。这些值可以提供足够的信息确定对象的序，但两种序之间的差并不能定量计算。

2.1.2 连续型数据

连续型数据的经典定义是取实数值的属性，如温度、高度或重量等，通常连续型数据用浮点变量表示。但是在实践中，连续型数据通常既包括比率属性，也包括区间属性。

- 比率属性：是具有固有零点的数值属性，例如年龄、绝对温度、长度、工作年限等。对于这些数据来说，差和倍数都是有意义的。
- 区间属性：没有绝对的零点，例如摄氏温度中 0℃不表示没有温度，因此我们虽然可以计算温度值之差，但不嫩说一个温度值是另一个的倍数。类似地，日历日期也没有绝对的零点，因为 0 年并不对应于时间的开始。

2.2 数据预处理

2.2.1 数据预处理的原因

数据预处理的目的是提高数据挖掘的质量，以及降低实际挖掘所需要的时间。数据质量问题可以从应用的角度考虑，如果某数据能满足其应用要求，那么它就是高质量的。但是当今现实世界的数据库极易受噪声、缺失值和不一致数据的侵扰，这种低质量的数据将导致低

质量的挖掘结果。

数据质量涉及许多因素，包括准确性、完整性、一致性、时效性、可信性和可解释性。

例 2.1 数据质量。假设你是某超市的经理，负责分析某部门的销售数据。当你仔细研究并审查公司的数据库和数据仓库时，发现许多元组在一些属性上没有值，有的虽然有值，但不符合常理或直接使用的系统默认值（比如用户不希望填写自己真实的出生年月时，会选择系统默认的“1 月 1 日”）。这种情况说明了现实数据很难满足的三个要素——准确性、完整性和一致性。不正确的数据可能因为收集数据的设备出了问题，人在输入数据时出现错误，数据传输故障以及输入字段格式不一致（如日期的格式）等原因引起。不完整数据的出现可能有更多原因，例如顾客认为某些选项可填可不填而造成有些感兴趣的属性缺少属性值，或数据库中仅包含聚集数据等。

想象这样一种情景，你企图监控每个销售代理的月销售量，但发现有些销售代理并未在月底及时提交相关记录，从而影响了数据质量，这种情况说明了数据时效的重要性。反映用户信赖程度的可信性对于数据质量也非常重要，假设当你发现数据库中出现错误之前就已经将该数据传给了销售部门的用户，即便之后这些错误被改正，但用户也很可能不再相信该数据。此外，反映数据是否容易理解的可解释性也是数据质量的一大影响因素。假设该超市的数据中使用了很多会计专有名词，销售部门对这方面的专业知识并不了解，从而不知道如何对它们进行解释和分析，于是对于销售部门的用户来说，它仍然是一个低质量的数据。

2.2.2 数据预处理的主要步骤

数据预处理的步骤主要包括数据清理、数据集成、数据变换和数据归约，这些步骤相互之间并不完全独立。图 2-1 概括了每个步骤可能遇到的问题以及相对应的解决方法，详细内容见 2.3 节至 2.6 节。

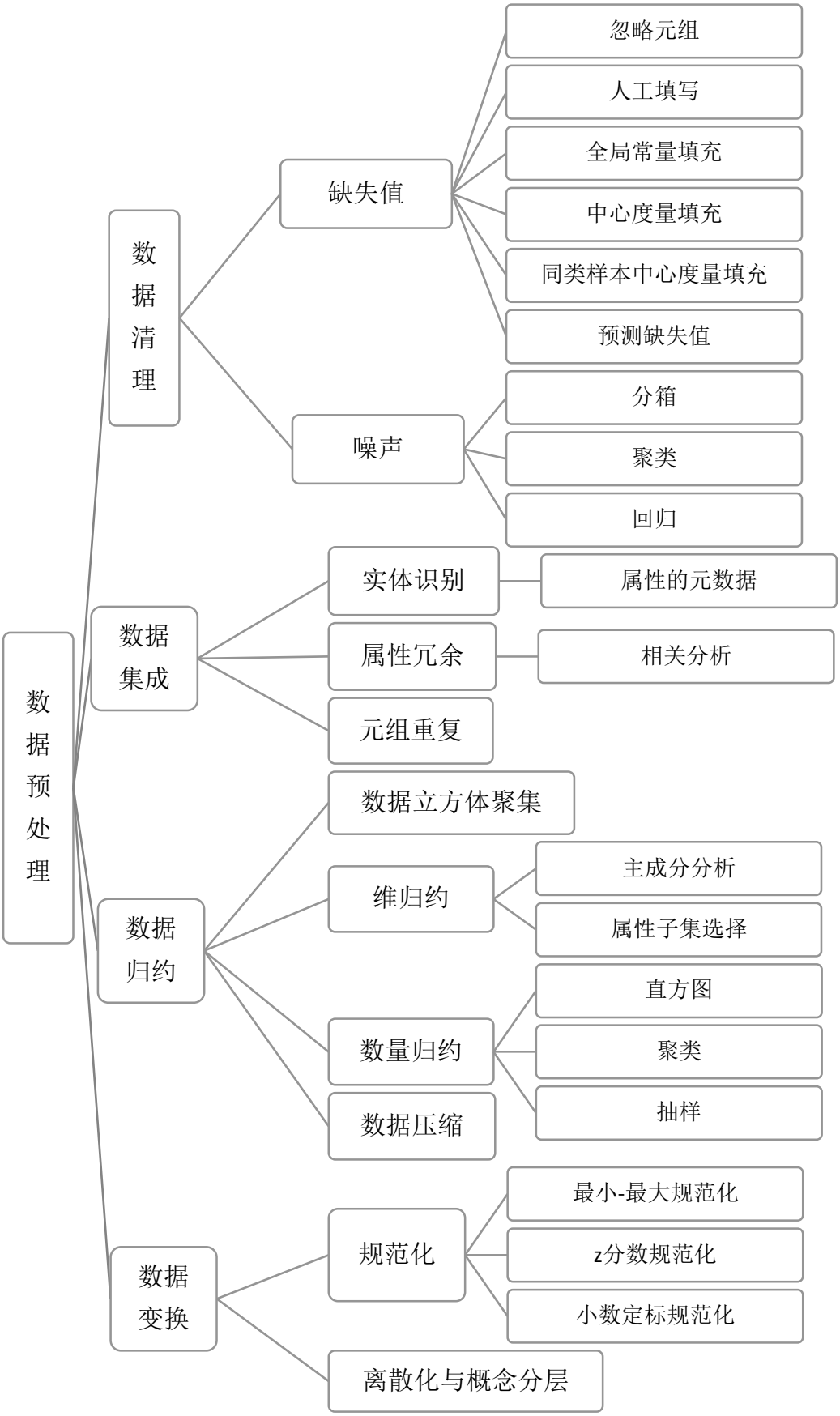

图 2-1 数据预处理的步骤、问题和方法

2.3 数据清理

现实世界的数据一般是不完整的、有噪声的和不一致的。数据清理过程试图通过填充缺失值、光滑噪声以及识别孤立点，以纠正数据中的不一致。

2.3.1 空缺值

现实数据并不总是完整的，例如某超市的数据库表中顾客收入字段往往因为多种原因并没有相应值。引起空缺值的原因有很多，包括设备异常、数据因为误解或得不到重视而没有被输入、数据与其他已有数据不一致而被删除、对数据的改变没有进行日志记载等。

如何才能为空缺的位置填上合适的值？

（1）忽略元组：当类标号缺少时通常这么做（假定挖掘任务设计分类或描述）。除非元组有多个属性缺少值，否则该方法不是很有效。当每个属性缺少值的百分比变化很大时，它的效果非常差。

（2）人工填写空缺值：工作量大，可行性低。

（3）使用一个全局变量填充空缺值：将缺失的属性值用同一个常量（比如使用 unknown 或-∞）来替换。但是如果缺失的值过多，则很有可能导致挖掘程序误认为“unknown”是一个有意义的值，因为它的出现频率过高。因此该方法虽然简单，但并不十分可靠。

（4）使用属性的中心度量值填充空缺值：对于呈对称分布的数据来说，可以使用平均值来填补缺失；对于非对称分布的数据来说，可以使用中位数。例如某超市的顾客收入数据呈对称分布，且平均收入为 5600 元/月，则使用该值替换 income 中的缺失值。

（5）使用与给定元组属同一类的所有样本的中心度量值：例如如果将顾客按照 age 分类，则用年龄相同的顾客的平均收入替换 income 中的缺失值。

（6）使用最可能的值填充空缺值：使用贝叶斯公式或判定树等基于推断的方法。例如，利用数据集中其他顾客的属性，可以构造一棵决策树，从而预测 income 的缺失值。

2.3.2 噪声数据

噪声（noise）是指一个测量变量中的随机错误或偏差。引起不正确属性值的原因包括数据收集工具问题、数据输入错误、数据传输错误、技术限制、命名规则不一致等。

噪声数据的处理方法包括分箱、回归和孤立点分析。

（1）分箱（binning）

分箱是指把待处理的数据按照一定的规则放进一些箱子中，通过考察每个箱子中的数据，采用某种方法分别对各个箱子中的数据进行处理。

分箱前需要对数据按目标属性值的大小进行排序，常用方法有等深分箱法和等宽分箱法。

- 等深分箱法，又称等频分箱法，是按照元组个数来分箱，每个箱子具有相同的元组数，每箱具有的元组数为箱的深度。
- 等宽分箱法，是在整个属性值的区间上平均分布，即每个箱的区间范围是一个常量，该区间范围为箱子的宽度。

例 2.2 分箱方法。某超市的顾客收入 income 排序后的值如下（单位：元/月）：800, 1000, 1200, 1500,1500,1800,2000,2300,2500,2800,3000,3500,4000,4500,4800,5000。若按照深度为 4 的等深分箱法对其分箱，则可分为以下 4 个箱子。

箱子 1: 800,1000,1200,1500　　箱子 2: 1500,1800,2000,2300

箱子 3: 2500,2800,3000,3500　　箱子 4: 4000,4500,4800,5000

若按照宽度为 1000 元的等宽分箱法对其分箱，则可得到不一样的结果：

箱子 1: 800, 1000, 1200,1500,1500,1800

箱子 2: 2000, 2300,2500,2800,3000