目录

7 清洗脏数据 **1**
本章小结 3
(1) 思维导图 3
(2) Python实现 3
参考文献 3

7 清洗脏数据

实际业务中获得的数据并不都是正确的, 常常不可避免地存在着不完整、不一致、不准确和重复的数据, 这些数据统称为“脏数据(dirty data)”, 包括: 重复值、无效值、错误值、离群值和缺失值等. 脏数据可能使数据分析过程陷入混乱, 导致不可靠的结果. 因此需要仔细清理脏数据, 从而提升数据质量, 使得数据满足分析的需求.

在清洗之前, 需要识别脏数据. 常用的识别方法包括:

(1) 还原数据收集过程, 考察收集过程中可能出现的数据质量问题.

(2) 结合实际业务知识, 采用一定的逻辑规则, 借助于编程等方式搜索脏数据.

(3) 采用探索性数据分析方法, 找出错误值、无效值等.

清洗脏数据的方式可以简单归纳为两种: 第一种方法是丢弃脏数据; 第二种方法是替换修正脏数据. 无论使用何种处理方式, 都会涉及修改数据, 需要相当谨慎. 本章主要讨论重复值、无效值和错误值等三种脏数据的识别及其处理. 离群值和缺失值的处理分别在第八和第九章介绍.

重复值包括重复的变量和样本观测, 常出现在由多源数据合并后的数据集中. 例如, 同一个变量在两个数据库中被赋予不同的字段名, 在合并后的数据集中成为两个变量. 识别重复的变量可以通过考察数据来源和相关性分析等方式实现. 识别重复的样本观测可采用软件包中专门的函数快速实现. 一般, 直接删除重复值.

无效值包括无效的变量和样本观测, 一般需要结合实际业务和数据分析需求予以判断. 例如, 在数据库中附加的一些字段解释、与本次分析无关的变量和样本观测等, 都可能属于无效数据. 可以直接删除无效值.

错误值属于噪声数据, 包括各种与实际情况不相符的数值. 研究表明, 约40%的数据集中含有错误值(Fayyad et al., 2003). 相对于重复值和无效值, 错误值对数据分析的危害更大, 可能导致出现严重背离实际情况的分析结果,

表 C－1　可能出现的逻辑错误

(1) 年龄小于 16，即 1992 年后出生的；
(2) 学位年份－出生年份＜16；
(3) 教师资格证年份－出生年份＜16；
(4) 目前职称年份－出生年份＜16；
(5) 获得特级教师年份－出生年份＜16；
(6) 年龄－教龄＜16
(7) 年龄＋4＜由学历推断的年龄
(8) 课程与学校类型不符(如小学教师选择了物理课)
(9) 课程门数与具体课程的填写结果不符
(10) 课程的课时数设置与教育部规定不符

图 7.1: 数据逻辑错误示例(来源: 丁钢(2010))

因而尤其需要关注. 错误值可能有多种来源, 例如, 在问卷调查中, 由于被调查者记忆或填写错误, 导致采集了错误的数据; 在数据处理中, 工作人员可能使用了错误的公式或代码衍生特征, 等等. 某些错误值难以被识别, 例如, 由被调查者填写的部分错误数据, 因为我们往往不知道其真实值, 难以比对识别. 但也存在一部分可被识别的错误值, 例如, 与实际业务或经验相违背的取值, 或者可借助于其他变量的信息进行逻辑判断的错误值. 当错误值能被识别时, 可以考虑借助相关变量的信息予以修正.

例7.1. 丁钢*(2010)* 描述了中国中小学教师专业发展状况调查项目的情况*:* 经过多阶段随机抽样, 入样的中小学教师到达所在校的机房, 进行问卷填答. 问卷题目均为选择题, 题目间的关联跳转是自动的, 待所有题目填答完成才能成功提交. 这些操作方式较好地保证了数据质量, 避免了重复值、无效值和缺失值等问题.

但是, 错误值仍然难以避免. 数据分析人员对问卷数据中的部分变量进行逻辑错误识别, 识别规则包括如图*7.1*所示的*10*项, 如*:* 年龄过小、所教课程与学校类型不相符等. 按照上述规则编程运行, 识别到了一些错误值. 数据分析人员提出了一些替换修正错误值的方案. 例如, 对于小于*16*的年龄值, 以与该被调查者同类教师的平均年龄进行替换修正[1].

清洗脏数据时需要注意:

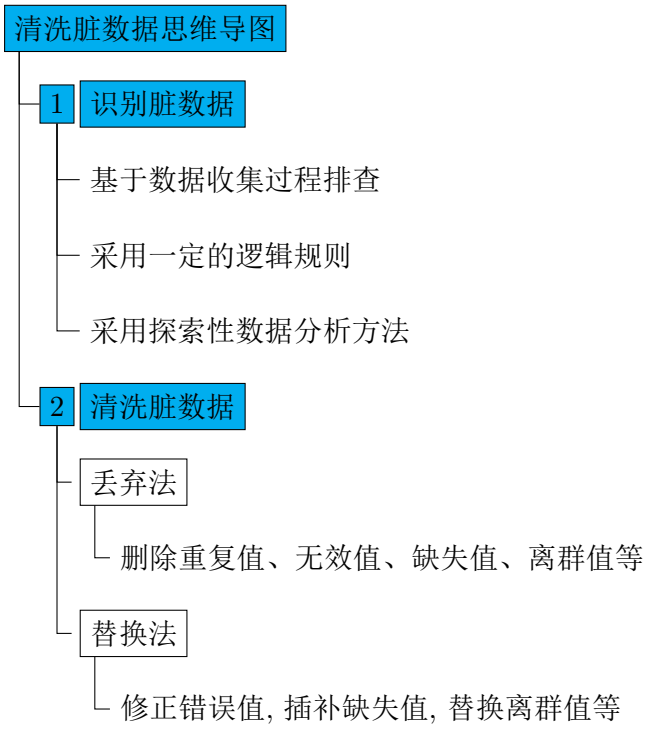
(1) 清洗过程中，应保留原始数据, 即清洗产生的变量应额外添加, 而非直接替换原始变量. 直到最终的清洗工作结束时, 从中提取所需的变量, 构建为新的数据集. 这样可以保证在数据预处理反复进行的过程中, 可以适时检查和调整.

[1]可参看丁钢(2010) 附录C关于替换修正规则的详细描述.

(2) 应反复检查清洗过程和结果, 避免出错.

本章小结

(1) 思维导图



(2) Python实现

表7.1列出了识别和处理重复样本观测的Python函数.

References

[1] 丁钢主编(2010) 中国中小学教师专业发展状况调查与政策分析报告. 上海: 华东师范大学出版社.

[2] Fayyad U.M., Piatetsky-Shapiro G., Uthurasamy R. (2003) Summary from the KDD-03 panel - data mining: the next 10 years. ACM SIGKDD Explorations Newsletter, 5(2):191-196.

表 7.1: 识别和处理重复样本观测的Python函数

内容	函数名称与所属模块	重要参数	应用提示
识别重复的样本观测	duplicated (pandas)	keep: 标记重复的样本观测. 默认值为‘first’, 将第一次出现的重复样本观测标记为False, 其余的重复观测标记为True; 若取值为‘last’, 将最后一次出现的重复样本观测标记为False, 其余的重复观测标记为True; 若取值为False, 标记所有的重复观测为True.	-
删除重复的样本观测	drop_duplicates (pandas)	(1) keep: 保留哪些重复的样本观测. 默认值为‘first’, 保留第一次出现的重复样本观测, 删除其余的重复观测; 若取值为‘last’, 保留最后一次出现的重复样本观测, 删除其余的重复观测; 若取值为False, 删除所有的重复样本观测. (2) inplace: 是否更改现有数据集. 默认值为False, 即保留现有数据集, 产生一个新的删除了重复样本观测的数据集; 若取值为True, 将更改现有数据集, 即删除其中的重复样本观测.	-