

第 5 章

数据的清洗与检验

学习目标

- (1) 掌握数据去重的方法
- (2) 掌握缺失值处理的方法
- (3) 掌握异常值处理的方法
- (4) 了解数据检验的作用

数据清洗是一项复杂且烦琐的工作,同时也是整个数据分析过程中最重要的环节。数据清洗的目的在于提高数据质量,将“脏”数据(“脏”数据在这里指的是对数据分析没有实际意义、格式非法、不在指定范围内的数据)清洗干净,使原数据具有完整性、唯一性、权威性、合法性、一致性等特点。常见的数据清洗操作包括重复值的处理、缺失值的处理、异常值的处理等操作,同时,为了保证数据的有效性,少不了数据校验操作。本章将针对数据清洗和校验进行详细讲解。

5.1 数据去重

数据去重又称重复数据的删除,通常指的是找出数据文件集合中重复的数据并将其删除,只保存唯一的数据单元,从而消除冗余数据。通常,数据去重方法分为两种,分别是完全去重和不完全去重。本节将针对这两种数据去重的方式进行讲解。

5.1.1 完全去重

完全去重指的是消除完全重复的数据,这里提到的完全重复数据指的是数据表记录字段值完全一样的数据。例如,现在有两个表格,分别记录不同年份的用户信息,现要求合并统计所有用户信息,发现合并后的表格存在完全重复的数据,示例如图 5-1 所示。

从图 5-1 中可以看出,合并后的数据表中存在完全相同的数据,为了便于后期更加方便地使用这些用户数据,通常情况下会对数据进行去重操作。

接下来使用 Kettle 工具演示如何消除这些完全重复的数据,具体步骤如下。

1. 数据准备

合并后的用户名单存放在 CSV 文件 merge.csv 中,具体内容如图 5-2 所示。

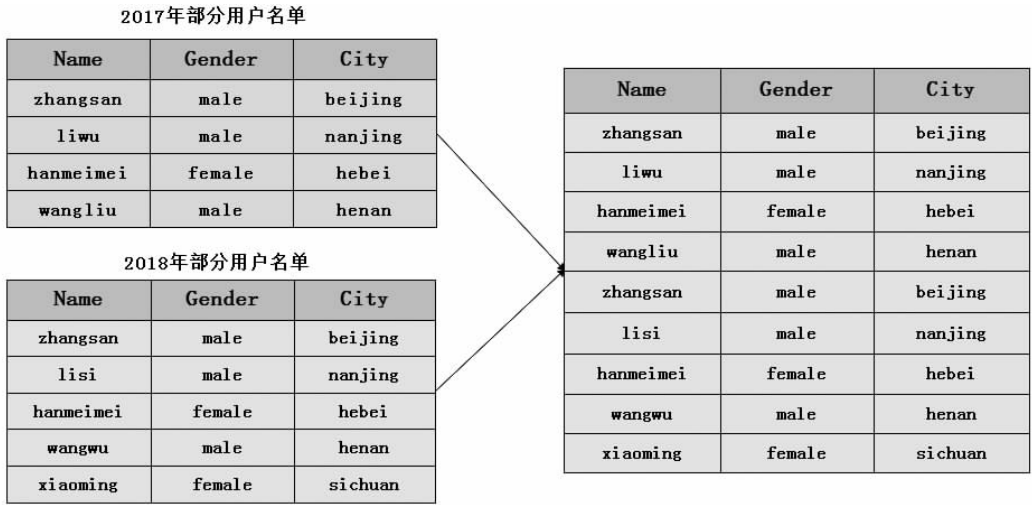


图 5-1 用户名单

	A	B	C
1	Name	Gender	City
2	zhangsan	male	beijing
3	liwu	male	nanjing
4	hanmeimei	female	hebei
5	wangliu	male	henan
6	zhangsan	male	beijing
7	lisi	male	nanjing
8	hanmeimei	female	hebei
9	wangwu	male	henan
10	xiaoming	female	sichuan

图 5-2 merge.csv 文件的内容

2. 打开 Kettle 工具,新建转换

使用 Kettle 工具创建转换 repeat_transform,并添加“CSV 文件输入”控件、“唯一行(哈希值)”控件以及 Hop 跳连接线,具体效果如图 5-3 所示。

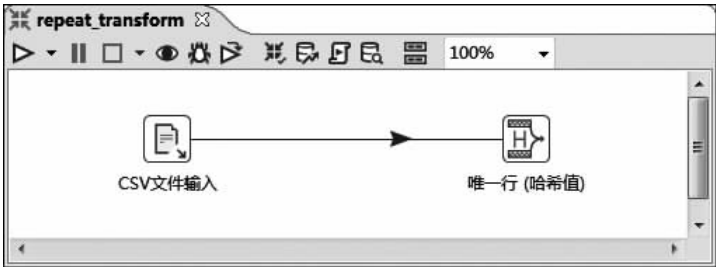


图 5-3 创建转换 repeat_transform

3. 配置“CSV 文件输入”控件

双击图 5-3 中的“CSV 文件输入”控件,进入“CSV 文件输入”界面,如图 5-4 所示。



图 5-4 “CSV 文件输入”界面

在图 5-4 中单击“浏览”按钮,选择要进行完全去重处理的 CSV 文件 merge.csv,如图 5-5 所示。



图 5-5 选择要进行完全去重处理的 CSV 文件 merge.csv

在图 5-5 中单击“获取字段”按钮,Kettle 会自动检索 CSV 文件,并对文件中的字段类型、格式、长度、精度等属性进行分析,具体如图 5-6 所示。

在图 5-6 中单击“预览”按钮,查看 CSV 文件 merge.csv 的数据是否加载到 CSV 文件输入流中,具体效果如图 5-7 所示。



图 5-6 Kettle 检索 CSV 文件



图 5-7 预览数据

从图 5-7 中可以看出, CSV 文件 merge.csv 的数据已经成功抽取到 CSV 文件输入流中, 单击“关闭”→“确定”按钮, 完成“CSV 文件输入”控件的配置。

4. 配置“唯一行(哈希值)”控件

双击图 5-3 中的“唯一行(哈希值)”控件, 进入“唯一行(哈希值)”界面, 如图 5-8 所示。

在图 5-8 中的“用来比较的字段”处添加要去重的字段, 这里可以单击“获取”按钮, 添加需要去重的字段, 具体如图 5-9 所示。

在图 5-9 中单击“确定”按钮, 完成“唯一行(哈希值)”控件的配置。



图 5-8 “唯一行(哈希值)”界面



图 5-9 添加需要去重的字段

5. 运行转换 repeat_transform

单击转换工作区顶部的  按钮,运行创建的转换 repeat_transform,实现消除 CSV 文件 merge.csv 中完全重复的数据,具体如图 5-10 所示。



图 5-10 运行转换 repeat_transform

从图 5-10 中执行结果窗口的“步骤度量”选项卡可以看出,“CSV 文件输入”控件输入 10 条数据并写入该控件 9 条数据(其中有 1 条数据为表头);“唯一行(哈希值)”控件从“CSV 文件输入”控件读取 9 条数据,写入该控件 7 条数据,拒绝 2 条数据(这 2 条数据为完全重复数据)。也就是说,CSV 文件 merge.csv 中有 2 条数据与其他数据完成重复。

选中图 5-10 中的“唯一行(哈希值)”控件,单击执行结果窗口的 Preview data 选项卡,查看是否消除 CSV 文件 merge.csv 中完全重复的数据,具体如图 5-11 所示。

#	Name	Gender	City
1	zhangsan	male	beijing
2	liwu	male	nanjing
3	hanmeimei	female	hebei
4	wangliu	male	henan
5	lisi	male	nanjing
6	wangwu	male	henan
7	xiaoming	female	sichuan

图 5-11 查看是否消除 CSV 文件 merge.csv 中完全重复的数据

从图 5-11 中可以看出,CSV 文件 merge.csv 中的数据已经没有了完全重复的值,说明通过 Kettle 工具实现了消除完全重复数据的功能。

注意: 本小节操作只是将 CSV 文件 merge.csv 的数据读取到 Kettle 中进行完全去重处理,并不会改变 CSV 文件 merge.csv 的原始数据,如需保存处理后的数据,须添加相关输出控件。

5.1.2 不完全去重

数据清洗过程中,所有字段值都相等的重复值是一定要剔除的。根据不同的业务场景,有时还需要选取其中若干字段进行去重操作。

假设现在有一份用户访问网站的数据文件 people.txt,具体内容如图 5-12 所示。

```
garcia MemberUser 15201011506 2019-03-17 11:59:12
anderson GeneralUser 13105916651 2019-04-02 12:00:32
sharma MemberUser 15803924337 2019-04-12 09:45:33
smith GeneralUser 15905741651 2019-04-22 09:55:44
johnson MemberUser 15804058842 2019-05-23 09:12:23
garcia GeneralUser 13404840224 2019-05-17 08:03:12
anderson GeneralUser 13202438054 2019-05-12 12:17:53
jones MemberUser 13606051520 2019-05-22 10:21:03
johnson GeneralUser 15000220514 2019-11-16 12:56:47
jones MemberUser 13606051520 2019-11-02 13:01:35
garcia MemberUser 15201011506 2019-08-24 15:17:57
```

图 5-12 文件 people.txt 的内容

从图 5-12 中可以看出,第 1 条记录和第 11 条记录都指向同一个用户,只是访问的时间不同。下面通过 Kettle 工具将文件 people.txt 中不完全重复的数据进行去重处理,即使用 Name(姓名)、UserLevel(用户级别)和 Phone(手机号)3 个字段作为去重处理的比较对象判

断唯一用户,具体步骤如下。

1. 打开 Kettle 工具,创建转换

使用 Kettle 工具创建转换 part_repeat_transform,并添加“文本文件输入”控件、“唯一行(哈希值)”控件以及 Hop 跳连接线,具体效果如图 5-13 所示。



图 5-13 创建转换 part_repeat_transform

2. 配置“文本文件输入”控件

双击图 5-13 中的“文本文件输入”控件,进入“文本文件输入”界面,如图 5-14 所示。



图 5-14 “文本文件输入”界面

在图 5-14 中单击“浏览”按钮,选择要去重的文件 people.txt,效果如图 5-15 所示。

在图 5-15 中单击“增加”按钮,将要去重的文件 people.txt 添加到转换 part_repeat_transform 中,具体效果如图 5-16 所示。

在图 5-16 中单击“内容”选项卡,切换到“内容”选项卡界面,如图 5-17 所示。

在图 5-17 中清除分隔符处的默认分隔符“;”,并单击 Insert TAB 按钮,在分隔符处插入一个制表符;取消勾选“头部”复选框,若不取消,则在进行数据抽取操作时会排除文件第一行的数据。“内容”选项卡的配置如图 5-18 所示。

在图 5-18 中单击“字段”选项卡,切换到“字段”选项卡界面,如图 5-19 所示。



图 5-15 选择要去重的文件 people.txt



图 5-16 添加文件 people.txt 至转换 part_repeat_transform 中

在图 5-19 中,根据文件 people.txt 的内容添加对应的字段名称,并指定数据类型,这里需要注意的是,制表符可看作是由多个空格组成,因此在“去除空字符串方式”列时,所添加的字段都应选择“不去掉空格”,否则在抽取数据操作时,会把制表符当作空格去除,而不能把制表符作为分隔符实现文本文件内容的分隔。“字段”选项卡的配置如图 5-20 所示。

在图 5-20 中单击“预览记录”按钮,查看文件 people.txt 的数据是否成功抽取到文本文件输入流中,具体效果如图 5-21 所示。

从图 5-21 中可以看出,文件 people.txt 的数据已经成功抽取到文本文件输入流中,单击“关闭”→“确定”按钮,完成“文本文件输入”控件的配置。



图 5-17 “内容”选项卡界面



图 5-18 “内容”选项卡的配置



图 5-19 “字段”选项卡界面



图 5-20 “字段”选项卡的配置



图 5-21 预览数据

3. 配置“唯一行(哈希值)”控件

双击图 5-13 中的“唯一行(哈希值)”控件,进入“唯一行(哈希值)”界面,如图 5-22 所示。



图 5-22 “唯一行(哈希值)”界面