R 机器学习

第 04 讲　数据清洗

张敬信

2022 年 10 月 23 日

哈尔滨商业大学

一. 什么是整洁数据?

- 采用 Hadley 的表述，脏的/不整洁的数据往往具有如下特点：
  - 首行（列名）是值，不是变量名
  - 多个变量放在一列
  - 变量既放在行也放在列
  - 多种类型的观测单元在同一个单元格
  - 一个观测单元放在多个表
- 而整洁数据具有如下特点：
  - 每个**变量**构成一列
  - 每个**观测**构成一行
  - 每个观测的每个变量**值**构成一个单元格

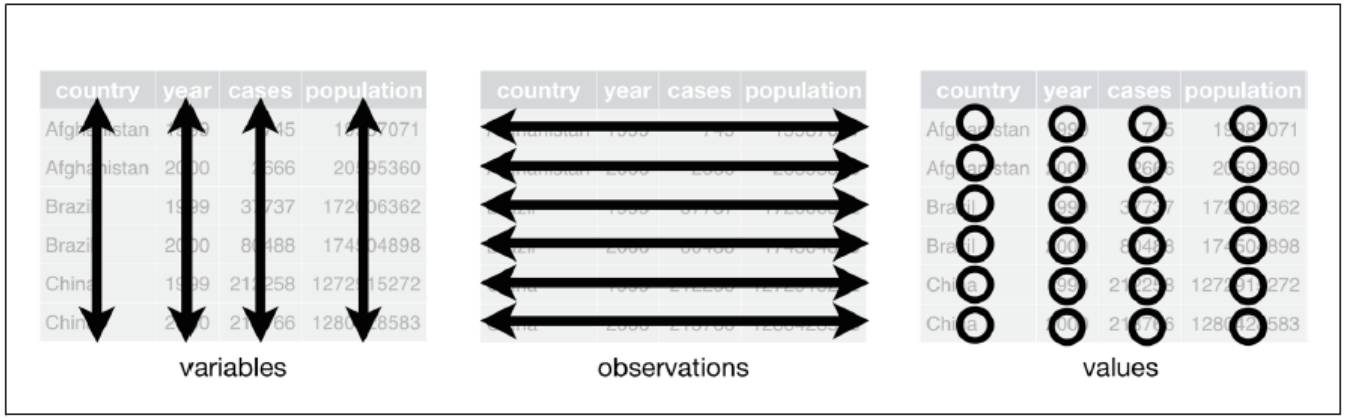

图 1: 整洁数据的 3 个特点

- `tidyverse` 系列包中的函数操作的都是这种整洁数据框，而不整洁数据，首先需要变成整洁数据，这就是**数据重塑**
- 数据重塑主要包括长宽表转化、拆分/合并列、方形化，用 `tidyr` 包实现。

- 先看一个不整洁的数据：

observation	A_count	B_count	A_dbh	B_dbh
Richmond(Sam)	7	2	100	110
Windsor(Ash)	10	5	80	87
Bilpin(Jules)	5	8	95	90

其不整洁表现在：

- `observation` 列有两个变量数据
- 列名中的 `A/B` 应是分类变量 `species` 的两个水平值
- 测量值列 `count` 和 `dbh` 应各占 1 列，而不是 2 列

- 用 tidyr 包重塑为整洁数据:

```
tidy_dt = dt %>%
  pivot_longer(-observation,
               names_to = c("speices", ".value"),
               names_sep = "_") %>%
  separate(observation, into = c("site", "surveyor"))
```

site	surveyor	speices	count	dbh
Richmond	Sam	A	7	100
Richmond	Sam	B	2	110
Windsor	Ash	A	10	80
Windsor	Ash	B	5	87
Bilpin	Jules	A	5	95
Bilpin	Jules	B	8	90

注：这里的关键是，要学会区分哪些是**变量**、**观测**、**值**。

二. 宽表变长表

- 宽表的特点是：表比较宽，本来该是”值”的，却出现在”变量（名）“中。这就需要给它变到”值”中，新起个列名存为一列，即宽表变长表：

```
pivot_longer(data, cols, names_to, values_to,
             values_drop_na, ...)
```

- data: 要重塑的数据框
- cols: 用选择列语法选择要变形的列
- names_to: 为存放变形列的列名中的”值”，指定新列名
- values_to: 为存放变形列中的”值”，指定新列名
- values_drop_na: 是否忽略变形列中的 NA

注：若变形列的列名除了”值”外，还包含前缀、变量名 + 分隔符、正则表达式分组捕获模式，则可以借助参数 names_prefix, names_sep, names_pattern 来提取出”值”。

1. 值列中只包含一个变量的值

- 以分省年度 GDP 数据为例，要变形的值列中只包含一个变量 GDP 的值

```
df = read_csv("data/分省年度 GDP.csv")
df
#> # A tibble: 4 x 4
#>   地区   `2019 年` `2018 年` `2017 年`
#>   <chr>      <dbl>    <dbl>    <dbl>
#> 1 北京市    35371.   33106.   28015.
#> 2 天津市    14104.   13363.   18549.
#> 3 河北省    35105.   32495.   34016.
#> # ... with 1 more row
```

- 要变形的列是除了地区列之外的列
- 变量（名）中的 2019 年、2018 年等是年份的值，需要作为 1 列”值”来存放，新起一个列名年份
- 2019 年、2018 年等列中的值，属于同一个变量 GDP，新起一个列名 GDP 来存放：

```
df %>%
  pivot_longer(-地区, names_to = " 年份", values_to = "GDP")
#> # A tibble: 12 x 3
#>    地区   年份      GDP
#>    <chr>  <chr>   <dbl>
#> 1 北京市 2019 年 35371.
#> 2 北京市 2018 年 33106.
#> 3 北京市 2017 年 28015.
#> # ... with 9 more rows
```

2. 值列中包含多个变量的值

- 以 family 数据集为例，要变形的值列中包行两个变量的值：dob 和 gender

```
load("data/family.rda")
knitr::kable(family, align = "c")
```

family	dob_child1	dob_child2	gender_child1	gender_child2
1	1998-11-26	2000-01-29	1	2
2	1996-06-22	NA	2	NA
3	2002-07-11	2004-04-05	2	2
4	2004-10-10	2009-08-27	1	1
5	2000-12-05	2005-02-28	2	1

- 要变形的列是除了 `family` 列之外的列;
- 变形列的列名以 “_” 分割为两部分，用 `names_to` 指定这两部分的用途: “.value” 指定第一部分将继续留作列名用来存放值，而第二部分，即包含”child1”、“child2”，作为新变量 `child` 的” 值”
- 忽略变形列中的缺失值

```
family %>%
  pivot_longer(-family,
               names_sep = "_",
               names_to = c(".value", "child"),
               values_drop_na = TRUE)
#> # A tibble: 9 x 4
#>   family child  dob         gender
#>    <int> <chr>  <date>       <int>
#> 1      1 child1 1998-11-26       1
#> 2      1 child2 2000-01-29       2
#> 3      2 child1 1996-06-22       2
#> # ... with 6 more rows
```

- 学生报名信息：每一行有 3 个观测，关于 3 名队员的信息，变成每一行只有 1 名队员的信息。用到 `names_pattern` 参数和正则表达式分组捕获。

```
df = read_csv("data/参赛队信息.csv")
knitr::kable(df, align = "c")
```

队员 1 姓名	队员 1 专业	队员 2 姓名	队员 2 专业	队员 3 姓名	队员 3 专业
张三	数学	李四	英语	王五	统计学
赵六	经济学	钱七	数学	孙八	计算机

```r
df %>%
  pivot_longer(everything(),
    names_pattern = "(.*\\d)(.*)",
    names_to = c(" 队员", ".value"))
#> # A tibble: 6 x 3
#>   队员  姓名  专业
#>   <chr> <chr> <chr>
#> 1 队员 1 张三  数学
#> 2 队员 2 李四  英语
#> 3 队员 3 王五  统计学
#> # ... with 3 more rows
```

```
dat = read_csv("data/demo_t.test.csv")
dat
#> # A tibble: 38 x 7
#>   compoundID case_1 case_2 case_3 control_1 control_2 cont
#>   <chr>       <dbl>  <dbl>  <dbl>     <dbl>     <dbl>
#> 1 com_001       485    154    268       350       432
#> 2 com_002       208    372    219       457       324
#> 3 com_003       219    125    345       473       480
#> # ... with 35 more rows
```

问题：数据共 38 行，每行是一组，包括 3 个实验样本，3 个控制样本；需求是批量地对每行，按实验组和控制组做 t 检验。

- 第 1 次宽变长

```
dat = dat %>%
  pivot_longer(-1, names_pattern = "(.*)_",
               names_to = ".value")
dat
#> # A tibble: 114 x 3
#>   compoundID  case control
#>   <chr>      <dbl>   <dbl>
#> 1 com_001      485     350
#> 2 com_001      154     432
#> 3 com_001      268     425
#> # ... with 111 more rows
```

- 第 2 次宽变长

```
dat = dat %>%
  pivot_longer(-1, names_to = "grp", values_to = "val")
dat
#> # A tibble: 228 x 3
#>   compoundID grp       val
#>   <chr>      <chr>   <dbl>
#> 1 com_001    case      485
#> 2 com_001    control   350
#> 3 com_001    case      154
#> # ... with 225 more rows
```

- group_by 分组 + t 检验

```
library(rstatix)   # 整洁统计
dat %>%
  group_by(compoundID) %>%
  t_test(val ~ grp)
#> # A tibble: 38 x 9
#>   compoundID .y.   group1 group2     n1    n2 statistic    
#> * <chr>      <chr> <chr>  <chr>   <int> <int>     <dbl> <d
#> 1 com_001    val   case   control     3     3    -0.994  2
#> 2 com_002    val   case   control     3     3    -1.91   3
#> 3 com_003    val   case   control     3     3    -3.25   2
#> # ... with 35 more rows
```

三. 长表变宽表

- 长表的特点是：表比较长。有时候需要将分类变量的若干水平值，变成变量（列名）。这就是长表变宽表[1]：

```
pivot_wider(data, id_cols, names_from, values_from,
            values_fill, ...)
```

- data: 要重塑的数据框
- id_cols: 唯一识别观测的列，默认是除了 names_from 和 values_from 指定列之外的列
- names_from: 指定列名来自哪个变量列
- values_from: 指定列”值”来自哪个变量列
- values_fill: 若变宽后单元格值缺失，设置用何值填充

[1]它与宽表变长表正好相反（二者互逆）.

- 只有一个列名列和一个值列，比如 `animals` 数据集：

```
load("data/animals.rda")
animals
#> # A tibble: 228 x 3
#>   Type    Year  Heads
#>   <chr>  <int>  <dbl>
#> 1 Sheep   2015 24943.
#> 2 Cattle  1972  2189.
#> 3 Camel   1985   559
#> # ... with 225 more rows
```

- 用 names_from 指定列名来自哪个变量；values_from 指定”值”来自哪个变量：

```
animals %>%
  pivot_wider(names_from = Type, values_from = Heads,
              values_fill = 0)
#> # A tibble: 48 x 6
#>    Year  Sheep Cattle Camel   Goat Horse
#>   <int>  <dbl>  <dbl> <dbl>  <dbl> <dbl>
#> 1  2015 24943.  3780.  368. 23593. 3295.
#> 2  1972 13716.  2189.  625.  4338. 2239.
#> 3  1985 13249.  2408.  559   4299. 1971
#> # ... with 45 more rows
```

- 多个列名列或多个值列，比如 us_rent_income 数据集有两个值列：

```
us_rent_income
#> # A tibble: 104 x 5
#>    GEOID NAME    variable estimate   moe
#>    <chr> <chr>   <chr>       <dbl> <dbl>
#> 1 01    Alabama income      24476   136
#> 2 01    Alabama rent          747     3
#> 3 02    Alaska  income      32940   508
#> # ... with 101 more rows
```

```
us_rent_income %>%
  pivot_wider(names_from = variable,
              values_from = c(estimate, moe))
#> # A tibble: 52 x 6
#>   GEOID NAME    estimate_income estimate_rent moe_income m
#>   <chr> <chr>             <dbl>         <dbl>      <dbl>
#> 1 01    Alabama           24476           747        136
#> 2 02    Alaska            32940          1200        508
#> 3 04    Arizona           27517           972        148
#> # ... with 49 more rows
```

长变宽时，经常会遇到两个问题：

- 长变宽正常会压缩行，为什么行数没变呢？
- 值不能被唯一识别，输出将包含列表列

```
df = tibble(
  x = 1:6, y = c("A","A","B","B","C","C"),
  z = c(2.13,3.65,1.88,2.30,6.55,4.21))
df
#> # A tibble: 6 x 3
#>       x y         z
#>   <int> <chr> <dbl>
#> 1     1 A      2.13
#> 2     2 A      3.65
#> 3     3 B      1.88
#> # ... with 3 more rows
```

- 想让 y 列提供变量名，z 列提供值，做长变宽，但是

```
df %>%
  pivot_wider(names_from = y, values_from = z)
#> # A tibble: 6 x 4
#>       x     A     B     C
#>   <int> <dbl> <dbl> <dbl>
#> 1     1  2.13 NA       NA
#> 2     2  3.65 NA       NA
#> 3     3 NA     1.88    NA
#> # ... with 3 more rows
```

这就是前面说到的第一个问题，本来该压缩成 2 行，但是由于 x 列的存在，无法压缩，只能填充 NA，这不是想要的效果。所以，在长变宽时要注意，是不能带着类似 x 列这种唯一识别各行的 ID 列的。

- 那去掉 x 列，重新做长变宽，但是又遇到了前面说的第二个问题：

```
df = df[-1]
df %>%
  pivot_wider(names_from = y, values_from = z)
#> # A tibble: 1 x 3
#>   A          B          C
#>   <list>     <list>     <list>
#> 1 <dbl [2]> <dbl [2]> <dbl [2]>
```

值不能唯一识别[2]，结果变成了列表列，同样不是想要的结果。

[2]值唯一识别，是指各分组（A 组 B 组 C 组）组内元素必须要能唯一识别，否则不能区分行的先后，只能打包到列表. 此时可以用参数 values_fn 指定一个汇总函数，比如 mean, 直接计算每组均值.

- 增加一个各组的唯一识别列：

```
df = df %>%
  group_by(y) %>%
  mutate(n = row_number())
df
#> # A tibble: 6 x 3
#> # Groups:   y [3]
#>   y         z     n
#>   <chr> <dbl> <int>
#> 1 A      2.13     1
#> 2 A      3.65     2
#> 3 B      1.88     1
#> # ... with 3 more rows
```

- 这才是能够长变宽的标准数据，再来做长变宽：

```
df %>%
  pivot_wider(names_from = y, values_from = z)
#> # A tibble: 2 x 4
#>       n     A     B     C
#>   <int> <dbl> <dbl> <dbl>
#> 1     1  2.13  1.88  6.55
#> 2     2  3.65  2.3   4.21
```

这回是想要的结果，新增加的列 n 若不想要，删除列即可。

- 整理电话号码

```
df = tibble(
  ID = c("A","B","B","C","D","D"),
  Tel = sample(13900000000:14000000000, 6)
)
df
#> # A tibble: 6 x 2
#>   ID            Tel
#>   <chr>       <dbl>
#> 1 A     13974972005
#> 2 B     13901022219
#> 3 B     13984830954
#> # ... with 3 more rows
```

```r
df %>%
  group_by(ID) %>%
  mutate(n = row_number()) %>%
  pivot_wider(names_from = n, values_from = Tel,
              names_prefix = "Tel")
#> # A tibble: 4 x 3
#> # Groups:   ID [4]
#>   ID           Tel1        Tel2
#>   <chr>       <dbl>       <dbl>
#> 1 A     13974972005          NA
#> 2 B     13901022219 13984830954
#> 3 C     13953386338          NA
#> # ... with 1 more row
```

- 解法 2

```
df %>%
  group_by(ID) %>%
  summarise(Tel = list(Tel)) %>%
  unnest_wider(Tel, names_sep = "")
#> # A tibble: 4 x 3
#>   ID           Tel1        Tel2
#>   <chr>       <dbl>       <dbl>
#> 1 A     13974972005          NA
#> 2 B     13901022219 13984830954
#> 3 C     13953386338          NA
#> # ... with 1 more row
```

四. 拆分列与合并列

- 拆分列与合并列也是正好相反（二者互逆）。
- `separate(data, col, into, sep, ...)`: 按分隔符 `sep` 将一列拆分为多列

```
table3
#> # A tibble: 6 x 3
#>   country      year rate
#> * <chr>       <int> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> # ... with 3 more rows
```

```r
table3 %>%    # 同时转化为数值型
  separate(rate, into = c("cases", "population"),
           sep = "/", convert = TRUE)
#> # A tibble: 6 x 4
#>   country      year cases population
#>   <chr>       <int> <int>      <int>
#> 1 Afghanistan  1999   745   19987071
#> 2 Afghanistan  2000  2666   20595360
#> 3 Brazil       1999 37737  172006362
#> # ... with 3 more rows
```

- separate_rows(): 可对不定长的列进行分列，并按行堆叠放置

```
df = tibble(Class = c("1 班", "2 班"),
            Name = c(" 张三，李四，王五", " 赵六，钱七"))
df
#> # A tibble: 2 x 2
#>   Class Name
#>   <chr> <chr>
#> 1 1 班   张三，李四，王五
#> 2 2 班   赵六，钱七
```

```r
df1 = df %>%
  separate_rows(Name, sep = ", ")
df1
#> # A tibble: 5 x 2
#>   Class Name
#>   <chr> <chr>
#> 1 1 班  张三
#> 2 1 班  李四
#> 3 1 班  王五
#> # ... with 2 more rows
```

- 若要逆操作还原回去：

```r
df1 %>%
  group_by(Class) %>%
  summarise(Name = str_c(Name, collapse = ", "))
```

- extract( ): 利用正则表达式的分组捕获，直接从一列中，提取出多组信息，生成多个列。

```
dt
#> # A tibble: 3 x 5
#>   observation   A_count B_count A_dbh B_dbh
#>   <chr>           <dbl>   <dbl> <dbl> <dbl>
#> 1 Richmond(Sam)       7       2   100   110
#> 2 Windsor(Ash)       10       5    80    87
#> 3 Bilpin(Jules)       5       8    95    90
```

```
dt %>%
  extract(observation, into = c("site", "surveyor"),
          regex = "(.*)\\((.*)\\)")
#> # A tibble: 3 x 6
#>   site     surveyor A_count B_count A_dbh B_dbh
#>   <chr>    <chr>      <dbl>   <dbl> <dbl> <dbl>
#> 1 Richmond Sam            7       2   100   110
#> 2 Windsor  Ash           10       5    80    87
#> 3 Bilpin   Jules          5       8    95    90
```

合并列

- unite(data, col, sep, ...): 用分隔符 sep 将多列合并为一列

```
table5
#> # A tibble: 6 x 4
#>   country     century year  rate
#> * <chr>       <chr>   <chr> <chr>
#> 1 Afghanistan 19      99    745/19987071
#> 2 Afghanistan 20      00    2666/20595360
#> 3 Brazil      19      99    37737/172006362
#> # ... with 3 more rows
```

```
table5 %>%
  unite(new, century, year, sep = "")
#> # A tibble: 6 x 3
#>   country     new   rate
#>   <chr>       <chr> <chr>
#> 1 Afghanistan 1999  745/19987071
#> 2 Afghanistan 2000  2666/20595360
#> 3 Brazil      1999  37737/172006362
#> # ... with 3 more rows
```

五. 综合案例：瓜子二手车数据清洗

```
library(lubridate)
df = read_csv("data/瓜子二手车汽车信息采集.csv",
              col_types = "c")
df
#> # A tibble: 6,000 x 13
#>   标题      城市  车源号 车主报价 新车指导价 上牌时间 里程
#>   <chr>     <chr> <chr>  <chr>    <lgl>      <chr>    <ch
#> 1 宝马 X3 20~ <NA>  HC-94~ 金融专 ~  NA         18-Mar   5
#> 2 奔驰 GLA 级 ~ <NA>  HC-94~ 金融专 ~  NA         18-Mar
#> 3 宝马 X1 20~ <NA>  HC-94~ 金融专 ~  NA         12-Jul   8
#> # ... with 5,997 more rows, and 3 more variables: 车主 <ch
#> #   当前采集时间 <time>, 标签 <chr>
```

- 删除缺失较多的列和行

```
df = df %>%
  select(where(~ mean(is.na(.x)) < 0.6)) %>%
  filter(pmap_lgl(., ~ mean(is.na(c(...))) < 0.5))
```

```r
df = df %>%
  mutate(上牌地 = ifelse(str_detect(排量, "\\d"),
                         排量, 上牌地),
         排量 = ifelse(str_detect(排量, "\\d"), NA, 排量),
         上牌时间 = ymd(str_c(上牌时间, "-15")),
         车龄 = (上牌时间 %--% today() / dyears(1))
                |> round(2),
         across(c(车主报价, 上牌地, 里程), parse_number),
         年款 = str_extract(标题, "\\d{4} 款"),
         型版 = str_extract(标题,
                            "[\u4e00-\u9fa5]+[型 | 版]"),
         标题 = str_replace(标题, "([a-zA-Z0-9])", " \\1"),
         temp = str_extract(标题, "^.*?(?= \\d{4})")) %>%
  separate(temp, into = c(" 品牌", " 型号"), sep = " ")
```

```
df %>%
  select(车主报价:排量, 车龄:型号)
#> # A tibble: 5,982 x 10
#>   车主报价 上牌时间    里程 上牌地 排量   车龄 年款   型版
#>      <dbl> <date>     <dbl>  <dbl> <chr> <dbl> <chr>  <chr
#> 1     30.7  2018-03-15  5.74     2   自动   4.61 2016 款 运动
#> 2     18.6  2018-03-15  2.2      1.6 自动   4.61 2017 款 动感
#> 3      9.87 2012-07-15  8.34     2   自动  10.3  2012 款 时尚
#> # ... with 5,979 more rows
```

本篇主要参阅 (张敬信, 2022), (Hadley Wickham, 2017), (Desi Quintans, 2019), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

Desi Quintans, J. P. (2019). *Working in the Tidyverse*. HIE Advanced R workshop.

Hadley Wickham, G. G. (2017). *R for Data Science*. O' Reilly, 1 edition. ISBN 978-1491910399.

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

黄湘云 (2021). *Github: R-Markdown-Template*.