R for bioinformatics, data wrangler, part 2

HUST Bioinformatics course series

Wei-Hua Chen (CC BY-NC 4.0)

24 September, 2024

section 1: TOC

前情提要

pipe

- pipe

dplyr

- select()
- filter()
- mutate()
- summarise()
- arrange()
- group_by() ...

今次提要

tidyr

- pivot_longer(); ## 代替 gather
- pivot_wider(); ## 代替 spread

section 2: data wrangler - tidyr

tidyr

what is `tidyr` ?

The goal of tidyr is to help you create **tidy** data.

Figure 1: dplyr logo

more to read

- tidyr offical page at tidyverse
- R for data science

tidyr 安装

只需安装一次即可！

```
# The easiest way to get tidyr is to install the whole tidyverse:
install.packages("tidyverse")

# Alternatively, install just tidyr:
install.packages("tidyr")

# Or the development version from GitHub:
# install.packages("devtools")
devtools::install_github("tidyverse/tidyr")
```

Get the cheatsheet at here

宽数据向长数据转变

get data ready

```
library(tidyverse); ## 先装入包;
grades2 <- read_tsv(file = "data/talk06/grades2.txt");

grades2;
```

```
## # A tibble: 3 x 6
##   name         Microbiology English Chinese Bioinformatics Chemistry
##   <chr>               <dbl>   <dbl>   <dbl>          <dbl>     <dbl>
## 1 Zhi Liu               100      50      69             NA        NA
## 2 Weihua Chen            89      99      NA             99        NA
## 3 Kang Ning              NA      NA      20            100        76
```

宽数据的特点

优点：

- 自然，易理解；

缺点：

- 不易处理；
- 稀疏时问题较大；

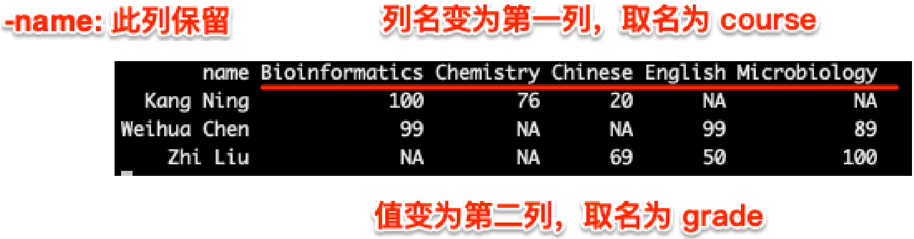
宽数据**向** 长数据**转变**

```
library(kableExtra);
grades3 <- grades2 %>% pivot_longer( - name, names_to = "course", values_to = "grade" );
kbl( grades3 );
```

name	course	grade
Zhi Liu	Microbiology	100
Zhi Liu	English	50
Zhi Liu	Chinese	69
Zhi Liu	Bioinformatics	NA
Zhi Liu	Chemistry	NA
Weihua Chen	Microbiology	89
Weihua Chen	English	99
Weihua Chen	Chinese	NA
Weihua Chen	Bioinformatics	99
Weihua Chen	Chemistry	NA
Kang Ning	Microbiology	NA
Kang Ning	English	NA
Kang Ning	Chinese	20
Kang Ning	Bioinformatics	100
Kang Ning	Chemistry	76

pivot_longer explained!

```
grades3 <- grades2 %>% pivot_longer( - name, names_to = "course", values_to = "grade" );
```



-name: 此列保留

列名变为第一列，取名为 course

```
         name Bioinformatics Chemistry Chinese English Microbiology
    Kang Ning            100        76      20      NA           NA
 Weihua Chen              99        NA      NA      99           89
      Zhi Liu             NA        NA      69      50          100
```

值变为第二列，取名为 grade

Figure 2: pivot_longer explained!

有 NA 值怎么办？

```
grades3_1 <- grades3[ !is.na(grades3$grade),  ];
grades3_2 <- grades3[ complete.cases( grades3 )   ,  ];

## -- 更好的方法 ~~
grades3_long <- grades2 %>%
  pivot_longer( - name,
                names_to = "course",
                values_to = "grade",
                values_drop_na = TRUE);
```

values_drop_na 即可消除；

有 NA 值怎么办？cont.

```
kbl( grades3_long );
```

name	course	grade
Zhi Liu	Microbiology	100
Zhi Liu	English	50
Zhi Liu	Chinese	69
Weihua Chen	Microbiology	89
Weihua Chen	English	99
Weihua Chen	Bioinformatics	99
Kang Ning	Chinese	20
Kang Ning	Bioinformatics	100
Kang Ning	Chemistry	76

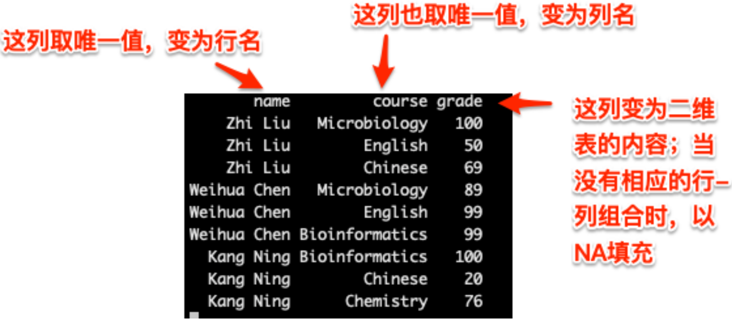
长变宽

```
grades3_wide <- grades3_long %>%
  pivot_wider( names_from = "course", values_from = "grade" );

grades3_wide;
```

```
## # A tibble: 3 x 6
##   name         Microbiology English Chinese Bioinformatics Chemistry
##   <chr>               <dbl>   <dbl>   <dbl>          <dbl>     <dbl>
## 1 Zhi Liu               100      50      69             NA        NA
## 2 Weihua Chen            89      99      NA             99        NA
## 3 Kang Ning              NA      NA      20            100        76
```

pivot_wider 怎么用？



这列取唯一值，变为行名

这列也取唯一值，变为列名

```
         name        course grade
      Zhi Liu  Microbiology   100
      Zhi Liu       English    50
      Zhi Liu       Chinese    69
 Weihua Chen  Microbiology    89
 Weihua Chen       English    99
 Weihua Chen Bioinformatics    99
    Kang Ning Bioinformatics   100
    Kang Ning       Chinese    20
    Kang Ning     Chemistry    76
```

这列变为二维表的内容；当没有相应的行–列组合时，以NA填充

Figure 3: `pivot_wider` function explained

宽长数据转换练习

用 pivot_wider 和 pivot_longer 对下面的数据 mini_iris 进行宽长转换:

```
mini_iris <- iris[ c(1, 51, 101),  ];
kbl( mini_iris);
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
101	6.3	3.3	6.0	2.5	virginica

iris 是鸢尾属一些物种花瓣的量表

宽变长， cont.

```
## -- 注意：第一、二个参数可以自行命名，分别对应原始数据中的 column names 及 values ...
mini_iris.longer <- mini_iris %>%
  pivot_longer( - Species, names_to = "type", values_to = "dat" );
kbl( mini_iris.longer  );
```

Species	type	dat
setosa	Sepal.Length	5.1
setosa	Sepal.Width	3.5
setosa	Petal.Length	1.4
setosa	Petal.Width	0.2
versicolor	Sepal.Length	7.0
versicolor	Sepal.Width	3.2
versicolor	Petal.Length	4.7
versicolor	Petal.Width	1.4
virginica	Sepal.Length	6.3
virginica	Sepal.Width	3.3
virginica	Petal.Length	6.0
virginica	Petal.Width	2.5

长变宽

```r
## -- 注意：第一、二个参数可以自行命名，分别对应原始数据中的 column names 及 values ...
mini_iris.wider <- mini_iris.longer %>%
  pivot_wider( names_from =  "type", values_from = "dat" );
kbl( mini_iris.wider  );
```

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.1	3.5	1.4	0.2
versicolor	7.0	3.2	4.7	1.4
virginica	6.3	3.3	6.0	2.5

比较复杂的例子

```
grades2 <- read_delim( file = "data/talk05/grades2.txt", delim = "\t",
                       quote = "", col_names = T);
kbl( grades2 );
```

name	class	course	grade
CHEN	1	bioinformatics	90
CHEN	1	chemistry	92
CHEN	2	chinese	35
CHEN	3	german	62
LI	1	bioinformatics	44
LI	2	chinese	68
LI	3	microbiology	95
LI	3	japanese	90
WANG	1	bioinformatics	35
WANG	1	chemistry	76
WANG	1	mathmatics	82
WANG	3	german	100
WANG	3	spanish	78

这是哪种数据类型？长还是宽？？

怎么变成宽数据？

```
grades2_wide <- grades2 %>%
  pivot_wider( names_from = course, values_from =  grade );

grades2_wide;
```

```
## # A tibble: 8 x 10
##   name  class bioinformatics chemistry chinese german microbiology japanese
##   <chr> <dbl>          <dbl>     <dbl>   <dbl>  <dbl>        <dbl>    <dbl>
## 1 CHEN      1             90        92      NA     NA           NA       NA
## 2 CHEN      2             NA        NA      35     NA           NA       NA
## 3 CHEN      3             NA        NA      NA     62           NA       NA
## 4 LI        1             44        NA      NA     NA           NA       NA
## 5 LI        2             NA        NA      68     NA           NA       NA
## 6 LI        3             NA        NA      NA     NA           95       90
## 7 WANG      1             35        76      NA     NA           NA       NA
## 8 WANG      3             NA        NA      NA    100           NA       NA
## # i 2 more variables: mathmatics <dbl>, spanish <dbl>
```

再变成长数据

又怎么把它变回来？

```
a <-
  grades2_wide %>%
  pivot_longer( ! c( name, class ),
                names_to = "course",
                values_to = "grade",
                values_drop_na = T
               );

kbl( a );
```

name	class	course	grade
CHEN	1	bioinformatics	90
CHEN	1	chemistry	92
CHEN	2	chinese	35
CHEN	3	german	62
LI	1	bioinformatics	44
LI	2	chinese	68
LI	3	microbiology	95
LI	3	japanese	90

另一种变法

又怎么把它变回来 ???

```
b <- grades2_wide %>%
  pivot_longer( bioinformatics:spanish, ## 选择成绩所在的列!
                names_to = "course", values_to = "grade",
                               values_drop_na = T
                               );

kbl( b );
```

name	class	course	grade
CHEN	1	bioinformatics	90
CHEN	1	chemistry	92
CHEN	2	chinese	35
CHEN	3	german	62
LI	1	bioinformatics	44
LI	2	chinese	68
LI	3	microbiology	95
LI	3	japanese	90
WANG	1	bioinformatics	35
WANG	1	chemistry	76

tidyr::separate

将一列拆成多列

```
table3 <- read_tsv(file = "data/talk06/table3.txt");
```

```
## Rows: 6 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): country, rate
## dbl (1): year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## 1. 将 rate 列拆分

table3 %>%
  separate(rate, into = c("cases", "population"), sep = "/");
```

```
## # A tibble: 6 x 4
##   country      year cases  population
##   <chr>       <dbl> <chr>  <chr>
## 1 Afghanistan  1999 745    19987071
## 2 Afghanistan  2000 2666   20595360
## 3 Brazil       1999 37737  172006362
## 4 Brazil       2000 80488  174504898
## 5 China        1999 212258 1272915272
```

tidyr::separate 同时进行格式转换

如何把分拆后的列正确识别为数字 ??

```r
table3 %>%
  separate(rate, into = c("cases", "population"), convert = TRUE)
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <dbl>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

tidyr::separate 按字符长度分割

把年拆分为世纪和年

```
table5 <- table3 %>%
  separate(year, into = c("century", "year"), sep = 2)

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
## 4 Brazil      20      00    80488/174504898
## 5 China       19      99    212258/1272915272
## 6 China       20      00    213766/1280428583
```

tidyr::unite

```
## 将上页分拆的结果进行合并

table5 %>%
  unite(new, century, year, sep = ""); ## sep 参数默认是 _
```

```
## # 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
## 4 Brazil      2000  80488/174504898
## 5 China       1999  212258/1272915272
## 6 China       2000  213766/1280428583
```

seperate 与 unit 小结

- 分割和合并后，原列会消失！可使用 `remove = FALSE` 保留原列
- 更多示例见：https://r4ds.had.co.nz/tidy-data.html

section 3：小结与作业

小结

今次提要

- tidyr (超级强大的数据处理) part 2

下次预告

- dplyr, tidyr 和 forcats 的更多功能与生信操作实例

important

- all codes are available at Github: `https://github.com/evolgeniusteam/R-for-bioinformatics`

练习 & 作业

- `Exercises and homework` 目录下 `talk06-homework.Rmd` 文件；
- 完成时间： 见钉群的要求