Module 4: Deterministic Blocking

Rebecca C. Steorts

joint with Olivier Binette

Recap

Define blocking and relate it to the figure below.

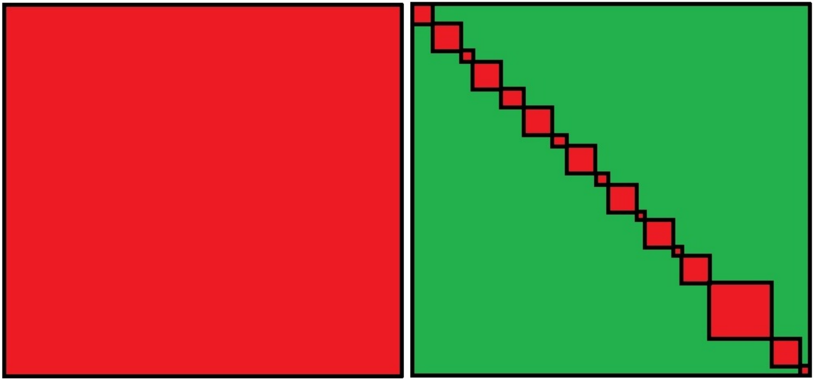


Figure 1: Left: All to all record comparison. Right: Example of resulting blocking partitions.

Traditional blocking

- What is traditional blocking?
- Consider the RLdata500 data set. What is a good blocking attribute to use? How can you validate this?
- What is probabilistic blocking?

Evaluation metrics

Review the following evaluation metrics and why we use these:

- Reduction ratio
- Precision
- Recall
- Fscore

Reading

- Binette and Steorts (2022)
- Steorts, Ventura, Sadinle, Fienberg (2014)
- Murray (2016)
- Christen (2012), Chapter 4

Agenda

- Data Cleaning Pipeline
- Blocking
- Traditional Blocking
- Probabilistic Blocking
- Evaluation Metrics
- Examples

Load R packages and data

```
knitr::opts_chunk$set(echo = TRUE,
                      fig.width=4,
                      fig.height=3,
                      fig.align="center")

if (!require("pacman")) {
  install.packages("pacman")
  library(pacman)
}
p_load(RecordLinkage, blink, italy, tidyverse, assert)
#library(RecordLinkage)
#library(blink)
#library(italy)
#library(assert)
data(italy08)
data(italy10)
data(RLdata500)
```

Data Cleaning Pipeline

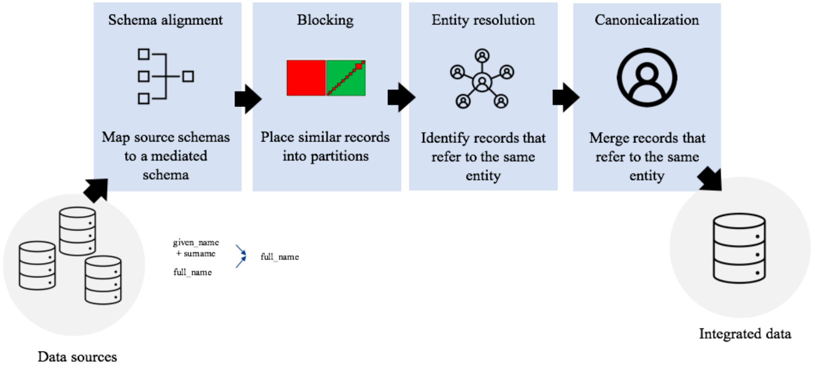


Figure 2: Data cleaning pipeline.

Blocking

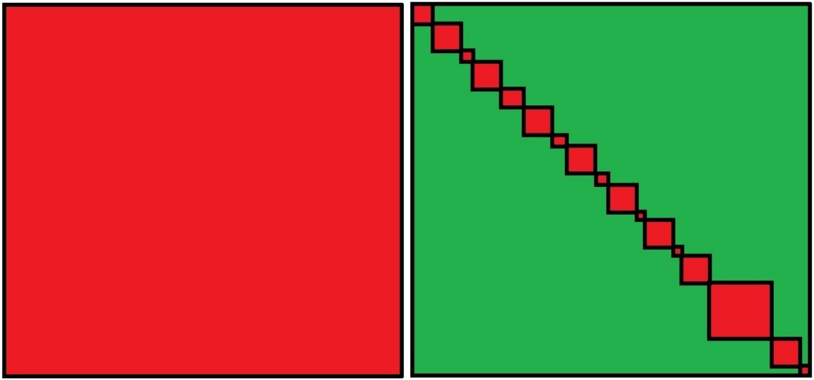


Figure 3: Left: All to all record comparison. Right: Example of resulting blocking partitions.

Blocking

- Blocking places similar records into partitions/blocks.
- ER (typically) is only performed within each block.

Traditional Blocking

- A deterministic (fixed) partition is formed based upon the data.
- A partition is created by treating certain fields that are thought to be nearly error-free as fixed.

Example: Blocking on date of birth year.

Traditional Blocking

- Benefits: simple, easy to understand, and fast to implement.
- Downsides: the blocks are treated as error free, which is not usually accurate and can lead to errors in the ER task that cannot be accounted for.

Probabilistic Blocking

- A probability model is used to cluster the data into blocks/partitions.

Example: Fellegi-Sunter (1969), or Locality Sensitive Hashing

Under both blocking approaches, record pairs that do not meet the blocking criteria are automatically classified as non-matches.

Evaluation Metrics

Evaluation metrics are important for ER as they help us evaluate our proposed methodology (as long as some notion of ground truth exists).

The three that we will focus on in this module are:

- reduction ratio
- precision
- recall
- f-measure

Reduction Ratio

The reduction ratio (RR) measures the relative reduction of the comparison space from the de-duplication or hashing technique.

See Christen (2012), Steorts, Ventura, Sadinle, Fienberg (2014) for a formal definition.

Pairwise Precision and Recall

Let's now turn to formally defining the pairwise precision and recall.

The confusion matrix

1. Pairs of data can be linked in both the handmatched training data (which we refer to as “truth") and under the estimated linked data. We refer to this situation as true positives (TP).
2. Pairs of data can be linked under the truth but not linked under the estimate, which are called false negatives (FN).
3. Pairs of data can be not linked under the truth but linked under the estimate, which are called false positives (FP).
4. Pairs of data can be not linked under the truth and also not linked under the estimate, which we refer to as true negatives (TN).

The confusion matrix

Make sure that you're able to write down the confusion matrix as we have previously defined.

Pairwise evaluation metrics

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

$$\text{Precision} = \frac{TP}{TP + FP} = 1 - \text{FDR}.$$

$$\text{F-measure} = 2 \times \frac{(precision \times recall)}{(precision + recall)}.$$

Recall

- For blocking, it is critical the recall be as close a possible to 1.
- To think about why, what does it mean if we have a blocking criterion where our recall is 0.5?

See Shrivastava and Steorts (2018) and Chen, Shrivastava, Steorts (2018) for further regarding about blocking criterion using human rights data.

Example: RLdata500

Let's return to the RLdata500 data set, where we will block by last name initial.

Our goal are the following:

- visualize the blocks
- compute the evaluation metrics introduced

Example: RLdata500

```r
head(RLdata500)
```

```
##   fname_c1 fname_c2 lname_c1 lname_c2   by bm bd
## 1  CARSTEN     <NA>    MEIER     <NA> 1949  7 22
## 2     GERD     <NA>    BAUER     <NA> 1968  7 27
## 3   ROBERT     <NA> HARTMANN     <NA> 1930  4 30
## 4   STEFAN     <NA>    WOLFF     <NA> 1957  9  2
## 5     RALF     <NA>  KRUEGER     <NA> 1966  1 13
## 6  JUERGEN     <NA>   FRANKE     <NA> 1929  7  4
```

Example: Traditional blocking

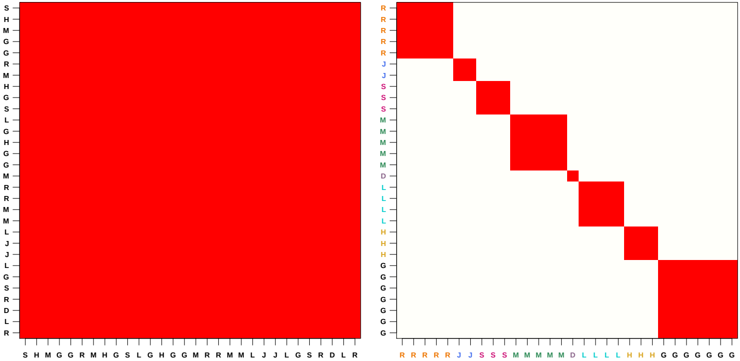


Figure 4: All-to-all record comparisons (left) versus partitioning records into blocks by lastname initial and comparing records only within each partition (right).

RLdata500 (Continued)

```r
# Total number of all to all record comparisons
choose(500,2)
```

```
## [1] 124750
```

RLdata500 (Continued)

```
# Block by last name initial
last_init <- substr(RLdata500[,"lname_c1"], 1, 1)
head(last_init)
```

```
## [1] "M" "B" "H" "W" "K" "F"
```

```
# Total number of blocks
length(unique(last_init))
```

```
## [1] 20
```

RLdata500 (Continued)

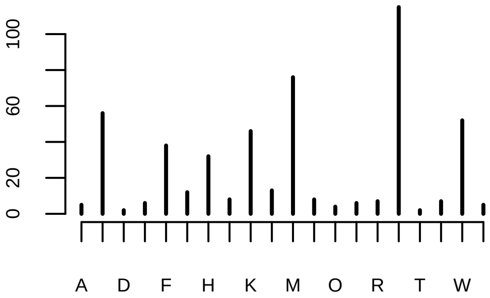
```r
# Total number of records per block
recordsPerBlock <- table(last_init)
head(recordsPerBlock)
```

```
## last_init
##  A  B  D  E  F  G
##  5 56  2  6 38 12
```

RLdata500 (Continued)

Observe that the block sizes vary.

```r
# Block sizes can vary
plot(recordsPerBlock, cex.axis=0.6, xlab="", ylab="")
```



RLdata500 (Continued)

What is the overall dimension reduction from the original space to the reduced space induced by blocking?

RLdata500 (Continued)

Recall the total number of all-to-all record comparisons made was:

```r
choose(500, 2)
```

```
## [1] 124750
```

Using blocking, we have reduced the compison space to the following:

```r
sum(choose(recordsPerBlock, 2))
```

```
## [1] 14805
```

How do we calculate the reduction ratio (RR)?

The reduction ratio is

$$\text{RR} = \% \text{ comparisons eliminated by blocking.}$$

```
(choose(500, 2) - sum(choose(recordsPerBlock, 2))) /
choose(500, 2)
```

```
## [1] 0.8813226
```

How do we calculate the RR (via a function)?

```
reduction.ratio <- function(block.labels) {
  n_all_comp = choose(length(block.labels), 2)
  n_block_comp = sum(choose(table(block.labels), 2))

  (n_all_comp - n_block_comp) / n_all_comp
}

reduction.ratio(last_init)
```

```
## [1] 0.8813226
```

Reduction Ratio

In summary, we have reduced the comparison space by roughly by 88 percent.

Evaluation metrics

Let's now code up the evaluation metrics for pairwise precision and recall.

Pairwise Precision

```r
precision <- function(block.labels, IDs) {
  ct = xtabs(~block.labels+IDs)

  # Number of true positives
  TP = sum(choose(ct, 2))

  # Number of positives = TP + FP
  P = sum(choose(rowSums(ct), 2))

  return(TP/P)
}
```

Pairwise Recall

```
recall <- function(block.labels, IDs) {
  ct = xtabs(~IDs+block.labels)

  # Number of true positives
  TP = sum(choose(ct, 2))

  # Number of true links = TP + FN
  TL = sum(choose(rowSums(ct), 2))

  return(TP/TL)
}
```

Pairwise Precision and Recall

```
precision(last_init, identity.RLdata500)
```

```
## [1] 0.003377237
```

```
recall(last_init, identity.RLdata500)
```

```
## [1] 1
```

```
precision(last_init, identity.RLdata500) ==
  recall(identity.RLdata500, last_init)
```

```
## [1] TRUE
```

Interpretation

- The recall says that 100 percent of the true matching record pairs are correctly classified as matches.
- The precision says that 0.34 percent of record pairs that are classified as matches correspond to true matches, while the rest correspond to false matches.

In summary, we would not want to use this for an entity resolution algorithm. Why is this?

Italian Survey on Household and Wealth (SHIW)

- We will now explore a case study to the Italian Survey on Household and Wealth (SHIW)
- The SHIW is a sample survey 383 households conducted by the Bank of Italy every two years (2008 and 2010).
- The data set is anonymized to remove first and last name (and other sensitive information).

SHIW

The following attribute information is available:

- PARENT (parental status)
- GENDER
- ANASC (year of birth)
- NASCREG (working status)
- CIT (employment status)
- ACOM4C (branch of activity)
- STUDIO (town size)
- Q (quality of life status)
- QUAL (whether or not Italian national)
- SETT (highest educational level obtained)
- IREG (region of italy)

Explore Data

```r
head(italy08) # first year of SHIW
```

```
##        id PARENT SEX ANASC NASCREG CIT ACOM4C STUDIO Q QUAL SETT IREG
## 1 1040021      1   2  1948      16   1      0      5 1    2    3   16
## 2 1040022     10   2  1952      16   1      0      7 1    2    3   16
## 3 1110521      1   1  1972      20   1      2      5 1    1    4   20
## 4 1110522      3   1  1935      20   1      2      2 3    6    5   20
## 5 1110523      3   2  1941      20   1      2      3 3    6    5   20
## 6  119401      1   1  1941       7   1      0      4 3    6    5    7
```

Explore Data

```r
head(italy10) # second year of SHIW
```

```
##        id PARENT SEX ANASC NASCREG CIT ACOM4C STUDIO Q QUAL SETT IREG
## 1 1040021      1   2  1948      16   1      0      5 3    6    5   16
## 2 1040022     11   2  1952      16   1      0      7 1    2    3   16
## 3 1110521      1   2  1941      20   1      2      3 3    6    5   20
## 4 1110522      2   1  1935      20   1      2      2 3    6    5   20
## 5 1110523      6   1  1972      20   1      2      5 1    1    4   20
## 6  119721      1   2  1948      16   1      2      2 2    5    4   17
```

Reformat Data

```
id08 <- italy08$id
id10 <- italy10$id
id <- c(italy08$id, italy10$id) # combine the id
italy08 <- italy08[-c(1)] # remove the id
italy10 <- italy10[-c(1)] # remove the id
italy <- rbind(italy08, italy10)
head(italy)
```

```
##   PARENT SEX ANASC NASCREG CIT ACOM4C STUDIO Q QUAL SETT IREG
## 1      1   2  1948      16   1      0      5 1    2    3   16
## 2     10   2  1952      16   1      0      7 1    2    3   16
## 3      1   1  1972      20   1      2      5 1    1    4   20
## 4      3   1  1935      20   1      2      2 3    6    5   20
## 5      3   2  1941      20   1      2      3 3    6    5   20
## 6      1   1  1941       7   1      0      4 3    6    5    7
```

Your turn (see homework 2)

- Construct a blocking criterion for the SHIW data set
- Provide code to construct the blocks
- Are your blocks well balanced?
- What is the reduction ratio?
- What is the pairwise recall and precision?
- Would you recommend your blocking criterion for an ER task? Why or why not.

Your turn solution

Let's block on gender.

```r
# block by gender
blockByGender <- italy$SEX
recordsPerBlock <- table(blockByGender)
head(recordsPerBlock)
```

```
## blockByGender
##     1     2
## 14442 12993
```

Summary

- Blocking is a method that puts similar records into blocks or bins.
- What are some examples of blocking methods. See if you can think of a few.
- Main metrics for blocking are recall and reduction ratio.
- Balance between size of blocks and number of blocks.