MiniProject_draft

Qinzhi Peng, Xinyan(Hathaway) Liu, Yujie(Johnny) Ye, Zhankai Ye

2024-09-30

Introduction

Weight-based Deduplication

Supervised Classification

1. Generating Duplication Data

```
# install.packages("RecordLinkage")
library(RecordLinkage)
```

```
## Loading required package: DBI
```

```
## Warning: package 'DBI' was built under R version 4.3.3
```

```
## Loading required package: RSQLite
```

```
## Warning: package 'RSQLite' was built under R version 4.3.3
```

```
## Loading required package: ff
```

```
## Warning: package 'ff' was built under R version 4.3.3
```

```
## Loading required package: bit
```

```
##
## Attaching package: 'bit'
```

```
## The following object is masked from 'package:base':
##
##     xor
```

```
## Attaching package ff
```

```
## - getOption("fftempdir")=="/var/folders/d5/v7l9v5kx5xqc637mjx5t_v_c0000gn/T//Rtmpq2sXBC/ff"
```

```
## - getOption("ffextension")=="ff"
```

```
## - getOption("ffdrop")==TRUE

## - getOption("fffinonexit")==TRUE

## - getOption("ffpagesize")==65536

## - getOption("ffcaching")=="mmnoflush"  -- consider "ffeachflush" if your system stalls on large wr

## - getOption("ffbatchbytes")==16777216 -- consider a different value for tuning your system

## - getOption("ffmaxbytes")==536870912 -- consider a different value for tuning your system

##
## Attaching package: 'ff'

## The following objects are masked from 'package:utils':
##
##     write.csv, write.csv2

## The following objects are masked from 'package:base':
##
##     is.factor, is.ordered

## RecordLinkage library

## [c] IMBEI Mainz

##
## Attaching package: 'RecordLinkage'

## The following object is masked from 'package:bit':
##
##     clone

## The following object is masked from 'package:base':
##
##     isFALSE
```

```r
# Load the example datasets
data(RLdata500)
data(RLdata10000)

# Generate a training set with 100 matches and 400 non-matches from RLData10000
train_pairs = compare.dedup(RLdata10000, identity = identity.RLdata10000,
                            n_match = 100, n_non_match = 400)

# Generate an evaluation set using record pairs from RLData500
eval_pairs = compare.dedup(RLdata500, identity = identity.RLdata500)
```

2. Training

```
model_rpart=trainSupv(train_pairs, method="rpart")
model_bagging=trainSupv(train_pairs, method="bagging")
model_svm=trainSupv(train_pairs, method="svm")
model_ada=trainSupv(train_pairs, method="ada")
model_nnet=trainSupv(train_pairs, method="nnet")
```

```
## # weights:  145
## initial  value 347.949220
## iter  10 value 3.901229
## iter  20 value 3.368144
## iter  30 value 3.365631
## iter  40 value 3.365130
## iter  50 value 3.365124
## iter  50 value 3.365124
## iter  50 value 3.365124
## final  value 3.365124
## converged
```

```
model_bumping=trainSupv(train_pairs, method="bumping")
```

3. Classification

```
result_rpart=classifySupv(model_rpart, eval_pairs)
result_bagging=classifySupv(model_bagging, eval_pairs)
result_svm=classifySupv(model_svm, eval_pairs)
result_ada=classifySupv(model_ada, eval_pairs)
result_nnet=classifySupv(model_nnet, eval_pairs)
result_bumping=classifySupv(model_bumping, eval_pairs)
```

4. Results

```
summary(result_ada)
```

```
##
## Deduplication Data Set
##
## 500 records
## 124750 record pairs
##
## 50 matches
## 124700 non-matches
## 0 pairs with unknown status
##
##
## 211 links detected
## 0 possible links detected
## 124539 non-links detected
##
```

```
## alpha error: 0.000000
## beta error: 0.001291
## accuracy: 0.998709
##
##
## Classification table:
##
##            classification
## true status      N      P      L
##       FALSE 124539      0    161
##       TRUE       0      0     50
```

```
summary(result_rpart)
```

```
##
## Deduplication Data Set
##
## 500 records
## 124750 record pairs
##
## 50 matches
## 124700 non-matches
## 0 pairs with unknown status
##
##
## 5317 links detected
## 0 possible links detected
## 119433 non-links detected
##
## alpha error: 0.000000
## beta error: 0.042237
## accuracy: 0.957780
##
##
## Classification table:
##
##            classification
## true status      N      P      L
##       FALSE 119433      0   5267
##       TRUE       0      0     50
```

```
summary(result_svm)
```

```
##
## Deduplication Data Set
##
## 500 records
## 124750 record pairs
##
## 50 matches
## 124700 non-matches
## 0 pairs with unknown status
##
```

```
##
## 111 links detected
## 0 possible links detected
## 124639 non-links detected
##
## alpha error: 0.000000
## beta error: 0.000489
## accuracy: 0.999511
##
##
## Classification table:
##
##            classification
## true status      N      P      L
##       FALSE 124639      0     61
##       TRUE       0      0     50
```