Example session for Weight-based deduplication

Andreas Borg, Murat Sariyar
April 9, 2020

This document shows an example session using the package `RecordLinkage`. A single data set is deduplicated using an EM algorithm for weight calculation. Conducting linkage of two data sets differs only in the step of generating record pairs.

1 Generating record pairs

The data to be deduplicated is expected to reside in a data frame or matrix, each row containing one record. Example data sets of 500 and 10000 records are included in the package as `RLData500` and `RLData10000`.

```r
data(RLdata500)
RLdata500[1:5,]
## 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
```

For deduplication, `compare.dedup` is to be used. In this example, blocking is set to return only record pairs which agree in at least two components of the subdivided date of birth, resulting in 810 pairs. The argument `identity` preserves the true matching status for later evaluation.

```r
pairs=compare.dedup(RLdata500,identity=identity.RLdata500,
blockfld=list(c(5,6),c(6,7),c(5,7)))
summary(pairs)
```

```r
##
## Deduplication Data Set
##
## 500 records
## 571 record pairs
##
## 49 matches
## 522 non-matches
## 0 pairs with unknown status
```


# Weight calculation

Weights are calculated by means of an EM algorithm. This step is computationally intensive and might take a while. The histogram shows the resulting weight distribution.

```r
pairs=emWeights(pairs)

hist(pairs$Wdata, plot=FALSE)
```

## $breaks
## [1] -15 -10 -5 0 5 10 15 20 25 30 35 40 45
##
## $counts
## [1] 352 13 0 0 5 26 42 123 9 0 0 1
##
## $density
## [1] 0.1232924694 0.0045534151 0.0000000000 0.0000000000
## [5] 0.0017513135 0.0091068301 0.0147110333 0.0430823117
## [9] 0.0031523643 0.0000000000 0.0000000000 0.0003502627
##
## $mids
## [1] -12.5 -7.5 -2.5 2.5 7.5 12.5 17.5 22.5 27.5
## [10] 32.5 37.5 42.5
##
## $xname
## [1] "pairs$Wdata"
##
## $equidist
## [1] TRUE
##
## attr("class")
## [1] "histogram"
```

# Classification

For determining thresholds, record pairs within a given range of weights can be printed using `getPairs`. In this case, 24 is set as upper and −7 as lower threshold, dividing links, possible links and non-links. The summary shows the resulting contingency table and error measures.

```r
getPairs(pairs,30,20)
```

<table>
<thead>
<tr>
<th>id</th>
<th>fname_c1</th>
<th>fname_c2</th>
<th>lname_c1</th>
<th>lname_c2</th>
<th>by</th>
<th>bm</th>
<th>bd</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>457</td>
<td>URSULA</td>
<td>BIRGIT</td>
<td>MUELLER</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1The output of `getPairs` is shortened in this document.
## Deduplication Data Set

<table>
<thead>
<tr>
<th>Record</th>
<th>ID</th>
<th>Name</th>
<th>Gender</th>
<th>Date</th>
<th>Age</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>25</td>
<td>ULRIKE</td>
<td>NICOLE</td>
<td>BECKER</td>
<td>NA</td>
<td>1982</td>
</tr>
<tr>
<td>25</td>
<td>26</td>
<td>ULRIKE</td>
<td>NICOLE</td>
<td>BECKER</td>
<td>NA</td>
<td>1982</td>
</tr>
<tr>
<td>26</td>
<td>27</td>
<td>ULRICH</td>
<td></td>
<td></td>
<td>NA</td>
<td>1962</td>
</tr>
<tr>
<td>27</td>
<td>28</td>
<td>SILKE</td>
<td></td>
<td></td>
<td>NA</td>
<td>1962</td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>MATTHIAS</td>
<td></td>
<td></td>
<td>NA</td>
<td>1955</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td>MATTHIAS</td>
<td></td>
<td></td>
<td>NA</td>
<td>1955</td>
</tr>
<tr>
<td>30</td>
<td>31</td>
<td>MATTHIAS</td>
<td></td>
<td></td>
<td>NA</td>
<td>1955</td>
</tr>
<tr>
<td>31</td>
<td>32</td>
<td>MATTHIAS</td>
<td></td>
<td></td>
<td>NA</td>
<td>1955</td>
</tr>
<tr>
<td>32</td>
<td>33</td>
<td>ANDRE</td>
<td></td>
<td></td>
<td>NA</td>
<td>1976</td>
</tr>
<tr>
<td>33</td>
<td>34</td>
<td>ANDRE</td>
<td></td>
<td></td>
<td>NA</td>
<td>1976</td>
</tr>
<tr>
<td>34</td>
<td>35</td>
<td>ANDRE</td>
<td></td>
<td></td>
<td>NA</td>
<td>1976</td>
</tr>
</tbody>
</table>

### Weight Distribution:

- [-15,-10]: 352
- [-10,-5]: 13
- [-5,0]: 0
- [0,5]: 0
- [5,10]: 5
- [10,15]: 26
- [15,20]: 42
- [20,25]: 123
- [25,30]: 9
- [30,35]: 0
- [35,40]: 0
- [40,45]: 1
Review of the record pairs denoted as possible links is facilitated by `getPairs`, which can be forced to show only possible links via argument `show`. A list with the ids of linked pairs can be extracted from the output of `getPairs` with argument `single.rows` set to `TRUE`.

```r
possibles <- getPairs(pairs, show="possible") possibles[1:6,]
```

```markdown
## Classification table:
##
<table>
<thead>
<tr>
<th>classification</th>
<th>true status</th>
<th>(N)</th>
<th>(P)</th>
<th>(L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>358</td>
<td>163</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>TRUE</td>
<td>0</td>
<td>35</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>
```

```r
links=getPairs(pairs,show="links", single.rows=TRUE) link_ids <- links[, c("id1", "id2")]
```

```markdown
## Weight
##
<table>
<thead>
<tr>
<th>id</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>21.691086</td>
</tr>
<tr>
<td>5</td>
<td>21.691086</td>
</tr>
</tbody>
</table>

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>289</td>
<td>289</td>
<td>399</td>
</tr>
<tr>
<td>297</td>
<td>297</td>
<td>388</td>
</tr>
<tr>
<td>357</td>
<td>357</td>
<td>414</td>
</tr>
<tr>
<td>313</td>
<td>313</td>
<td>457</td>
</tr>
<tr>
<td>467</td>
<td>467</td>
<td>472</td>
</tr>
<tr>
<td>183</td>
<td>183</td>
<td>444</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>107</td>
</tr>
<tr>
<td>106</td>
<td>106</td>
<td>175</td>
</tr>
<tr>
<td>370</td>
<td>370</td>
<td>478</td>
</tr>
<tr>
<td>127</td>
<td>127</td>
<td>142</td>
</tr>
</tbody>
</table>