Record Linkage

Arwa Ashi

09 March, 2022

Contents

Table 1: Example: First Rows of 'RLdata500' Dataset

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

1 Introduction

Digital transformation after COVID 19 has increased the data collecting for public and privet sector. If the data linked in a proper way that would improve the provided service and client experience. This report represent a several data linkage methodologies. There are three stages: Pre-linkage (see data preparation section), Linkage (see methodology section), and Post-linkage. The report will be divided into data, data exploration, data preparation, methodology, result, and conclusion sections.

2 Data

Finding an available dataset for data linkage project is not easy. Based on that the 'RLdata500' and 'RLdata10000' datasets under the package 'RecordLinkage' are considered.

3 Data Exploration

The data contains the first name, last name and date of birth for individuals. Notice that the data field can be different slightly, for example two records refer to the same entity i.e. peter can have a slight change in his last name or his date of birth. The 'RLdata500' and 'RLdata10000' have 7 columns for each, and 500 and 10000 rows consequently. The first name as 'fname_c1' and last name as 'lname_c1' are separated into two columns and date of birth is separated into several columns for year as 'by', month as 'bm', and day as 'bd', See Table 1.

The dataset assumed that it marge individual information from different databases.

4 Data Preparation

First step is pre-linkage stage which is to prepare the data for evaluation by generate the features that will be used in the models. In order to do that, a 'compare.depdup' function under the 'RecordLinkage' package is used to generate the feature. The generated feature (pair) compare two ids in each row. See Table 2. The number 1 and 0 mean perfect match or no match consequently. If the number is less than 0, then it means that it is a float number for a string comparison. The final column indicates if there is a match or not.

5 Methodology

The linkage stage that has the objective of matching the records in each 'RLdata500' and 'RLdata10000' datasets with no common unique identifiers and deduplicating with a dataset. There will be two methodologies: a probabilistic method and machine learning method.

The preprocessing stage was done in data preparation section by developing link keys by using blocking as 'blockfld' function under 'compare.dedup' function.

Table 2: Example: Generated Pairs of 'RLdata500' Dataset

id1	id2	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd	is_match
1	174	1	NA	0.1428571	NA	0	0	0	NA
1	204	1	NA	0.0000000	NA	0	0	0	NA
2	7	1	NA	0.3750000	NA	0	0	0	NA
2	43	1	NA	0.8333333	NA	1	1	1	NA
2	169	1	NA	0.0000000	NA	0	0	0	NA
4	19	1	NA	0.1428571	NA	0	0	0	NA

Table 3: Example: Initial Matched. of 'RLdata500' Dataset

id	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd	Weight
313	URSULA	BIRGIT	MUELLRR	NA	1940	6	15	
457	URSULA	BIRGIT	MUELLER	NA	1940	6	15	35.608887
467	ULRIKE	NICOLE	BECKRR	NA	1982	8	4	
472	ULRIKE	NICOLE	BECKER	NA	1982	8	4	35.568245

5.1 Probabilistic Method

There are a deterministic method that is a direct match by comparing everything needs to match, and a probabilistic method that is to estimate a probability or liklihood for two records. The focusing here is by using a probabilistic matching. For the classification, a Fellegi-Sunter Model is considered. Define a cut off for string comparing at 80% by using EM algorithm as 'emWeights' function in 'RecordLinkage' package. For a summary of weights for 'RLdata500', see the following:

```
##
## Deduplication Data Set
##
## 500 records
## 1221 record pairs
##
## 0 matches
## 0 non-matches
## 1221 pairs with unknown status
##
##
## Weight distribution:
##
## [-15,-10]  (-10,-5]    (-5,0]     (0,5]    (5,10]   (10,15]   (15,20]   (20,25]
##      1006       119        34        10         0        18        12         4
##   (25,30]   (30,35]   (35,40]
##        16         0         2
```

See Table 3 for initial matched. The initial matches is used as base to determine threshold. For 'RLdata500', the threshold is 30. See table 4 for final pairs.

Table 4: Example: Final Matched. of 'RLdata500' Dataset

id	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd	Weight
48	WERNER	NA	KOERTIG	NA	1965	11	28	
238	WERNIER	NA	KOERTIG	NA	1965	11	28	29.850470
68	PETEVR	NA	FUCHS	NA	1972	9	12	
190	PETER	NA	FUCHS	NA	1972	9	12	29.850470
85	THORSKTEN	NA	MARTIN	NA	1995	11	15	
187	THORSTEN	NA	MARTIN	NA	1995	11	15	29.850470
158	PETER	NA	BECKER	NA	1960	9	5	
229	PETERS	NA	BECKER	NA	1960	9	5	29.850470
177	JOHANNNES	NA	SCHULZ	NA	1974	1	17	
207	JOHANNES	NA	SCHULZ	NA	1974	1	17	29.850470
209	ROLBF	NA	NEUMANN	NA	1967	3	29	
227	ROLF	NA	NEUMANN	NA	1967	3	29	29.850470
265	MARIANNFE	NA	MOELLER	NA	1961	9	17	
456	MARIANNE	NA	MOELLER	NA	1961	9	17	29.850470
266	KARIN	NA	HORN	NA	2002	6	4	
437	KARINW	NA	HORN	NA	2002	6	4	29.850470
298	SONJA	NA	FISCHER	NA	1989	7	17	
464	SONJAD	NA	FISCHER	NA	1989	7	17	29.850470
310	MONIKA	NA	SCHNEIDER	NA	1937	6	2	
432	MONIYKA	NA	SCHNEIDER	NA	1937	6	2	29.850470
377	SABAINE	NA	OTTO	NA	1940	7	23	
448	SABINE	NA	OTTO	NA	1940	7	23	29.850470
391	GABRIELE	NA	BECKER	NA	1990	3	27	
496	GABRIHELE	NA	BECKER	NA	1990	3	27	29.850470
395	GISOELA	NA	BECK	NA	2003	4	16	
404	GISELA	NA	BECK	NA	2003	4	16	29.850470
402	CHRISTA	NA	SCHWARZ	NA	1965	7	13	
462	CHRISTAH	NA	SCHWARZ	NA	1965	7	13	29.850470
37	HARTMHUT	NA	HOFFMSNN	NA	1929	12	29	
72	HARTMUT	NA	HOFFMANN	NA	1929	12	29	29.657824
290	HELGA	ELFRIEDE	BERGER	NA	1989	1	18	
466	HELGA	ELFRIEDE	BERGER	NA	1989	1	28	28.768566

5.2 Machine Learning Method

For the machine learning approach ((logistic regression)), a 'reclin2' packages is considered for preparing the data for the algorithm. First, creating pairs by blocking fields by using 'pair_blocking' function in 'reclin2' package. Second, comparing the pairs to get comparing score for each feature by using 'compare_pairs' function in 'reclin2' package. Third, preparing the binary parameters 'TRUE' and 'FALS' by using 'compare_vars' function in 'reclin2' package. Fourth, using 'glm' function with a family = binomial() for the logistic regression. Fifth, predict the matching probability. Sixth, selecting a matching probability that is greater than 50%. Seventh, generating a FALSE TRUE table for the result evaluation. Finally, generate the Final matching pairs. See table 5.

```
##
##          TRUE
##    TRUE   600
```

Table 5: Example: Final Matched. of 'RLdata500' Dataset

.y	.x	fname_c1.x	fname_c2.x	lname_c1.x	lname_c2.x	by.x	bm.x	bd.x	id.x	fname_c1.y	fname_c2.y	lname_c1.y	lname_c2.y	by.y	bm.y	bd.y	id.y
2	43	GERD	NA	BAUERH	NA	1968	7	27	51	GERD	NA	BAUER	NA	1968	7	27	51
25	107	MATTHIAS	NA	HAAS	NA	1955	8	8	33	MATTHIAS	NA	HAAS	NA	1955	7	8	33
34	111	HEINZ	NA	BOEHMR	NA	1938	12	20	27	HEINZ	NA	BOEHM	NA	1938	12	20	27
37	72	HARTMUT	NA	HOFFMANN	NA	1929	12	29	139	HARTMHUT	NA	HOFFMSNN	NA	1929	12	29	139
43	2	GERD	NA	BAUER	NA	1968	7	27	51	GERD	NA	BAUERH	NA	1968	7	27	51
48	238	WERNIER	NA	KOERTIG	NA	1965	11	28	212	WERNER	NA	KOERTIG	NA	1965	11	28	212
50	234	STEFAN	NA	MUELLER	NA	1957	6	1	158	STEFAN	NA	MUELLER	NA	1957	6	7	158
58	148	FRANK	NA	MUELLER	NA	1978	5	20	174	FRANK	NA	MUELLDR	NA	1978	5	20	174
68	190	PETER	NA	FUCHS	NA	1972	9	12	136	PETEVR	NA	FUCHS	NA	1972	9	12	136
71	205	CHRISTIAN	NA	GROSS	NA	2008	4	7	151	CHRISTIAN	NA	GROSS	NA	1959	4	7	151
72	37	HARTMHUT	NA	HOFFMSNN	NA	1929	12	29	139	HARTMUT	NA	HOFFMANN	NA	1929	12	29	139
78	133	STEFAN	NA	BRAUN	NA	1947	12	30	28	STEFAN	NA	BRAUN	NA	1997	12	30	28
85	187	THORSTEN	NA	MARTIN	NA	1995	11	15	138	THORSKTEN	NA	MARTIN	NA	1995	11	15	138
87	117	HANS	NA	SCHULZE	NA	1972	11	28	112	HANS	NA	SCHULZE	NA	1972	11	27	112
106	175	ANDRE	NA	MUELLER	NA	1976	1	25	188	ANDRE	NA	MUELLER	NA	1976	2	25	188
107	25	MATTHIAS	NA	HAAS	NA	1955	7	8	33	MATTHIAS	NA	HAAS	NA	1955	8	8	33
108	203	GERHARD	NA	FRIEDRICH	NA	1957	2	10	100	GERHARD	NA	FRIEDRICH	NA	1987	2	10	100
111	34	HEINZ	NA	BOEHM	NA	1938	12	20	27	HEINZ	NA	BOEHMR	NA	1938	12	20	27
112	116	GERHARD	NA	ERNST	NA	1980	12	16	124	GERHARD	NA	ERNSR	NA	1980	12	16	124
116	112	GERHARD	NA	ERNSR	NA	1980	12	16	124	GERHARD	NA	ERNST	NA	1980	12	16	124
117	87	HANS	NA	SCHULZE	NA	1972	11	27	112	HANS	NA	SCHULZE	NA	1972	11	28	112
120	165	FRANK	NA	BERGKANN	NA	1998	11	8	114	FRANK	NA	BERGMANN	NA	1998	11	8	114
125	193	CHRISTIAN	NA	MUELLER	NA	1974	8	9	193	CHRISTIAN	NA	MUELLEPR	NA	1974	8	9	193
127	142	KARL	NA	KLEIBN	NA	2002	6	29	95	KARL	NA	KLEIN	NA	2002	6	20	95
130	147	MICHAEL	NA	MYER	NA	1988	1	31	11	MICHAEL	NA	MEYER	NA	1988	1	31	11
133	78	STEFAN	NA	BRAUN	NA	1997	12	30	28	STEFAN	NA	BRAUN	NA	1947	12	30	28

6 Result

Table 4 and 5 have the matching pairs' results by using a probabilistic methodology and logistic regression consequently.

7 Conclusion

The probabilistic and machine learning approaches for records linkages are working and matched the records. For the future work, a big dataset would be considered that have more features to have the ability to evaluate the performance of each approach and consider more approaches and assumptions.