Extending Naive Bayes Classifier with Hierarchy Feature Level Information for Record Linkage

Yun Zhou

AMBN-2015, Yokohama, Japan

16/11/2015

Overview

- “Record Linkage” aka “Matching” aka “Merge”.
- Finding records in a data set that refer to the same individual/entity across different data sources.
- Record linkage is necessary when joining data sets based on entities that may or may not share a common identifier.
  - Example: merging takeaway restaurant addresses from two websites in the same city.

Overview

- Can be accomplished manually, by visually comparing records from two separate sources.
- Approach becomes time consuming, tedious, inefficient, and unpractical as the number of records in two datasets increases.
- Technological advances in computer systems and programming techniques.
  - Economically feasible to perform computerized record linkage between large files.
  - Efficient and relatively accurate.

Deterministic Matching

- Computerized comparison where **everything** needs to match **exactly**:

Index	Name (f_1)	Address (f_2)	Borough (f_3)	Type (f_4)
1	Strada	Unit 6, RFH Belvedere Rd	Southwark	Roman
2	Strada	Unit 6, RFH Belvedere Rd	Southwark	Roman

Deterministic Matching

- Often slight variations exist in the data between the two fields for the same entity:

Index	Name (f_1)	Address (f_2)	Borough (f_3)	Type (f_4)
1	Strada	Unit 6, RFH Belvedere Rd	Southwark	Roman
2	Strad	Unit 9, RFH Belvedere Rd	Southwark	Roman

- Or variables are missing from one of the files:

Index	Name (f_1)	Address (f_2)	Borough (f_3)	Type (f_4)
1	Strada	Unit 6, RFH Belvedere Rd	Southwark	Roman
2	Strada	Unit 6, RFH Belvedere Rd	Southwark	

- These variations would prevent a match from being identified.

Probabilistic Matching

- Translating intuition into formal decision rules.
- Use the concept of probability and perform probabilistic matching.
- Recommended over traditional deterministic (exact matching) methods when:
  - coding errors, reporting variations, missing data or duplicate records.
- Estimate probability/likelihood that two records are for the same entity versus not.

Related Work

- The Fellegi-Sunter [Fellegi, I.P., Sunter, A.B. 1969] probabilistic record linkage (PRL-FS) is one of the most commonly used methods.
- Winkler [Winkler, W.E. 1990] proposed an enhanced PRL-FS method (PRL-W) that using Jaro-Winkler similarity to measure the extent of fields' match.

$$S_{jaro}(a,b) = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m}\right) & otherwise \end{cases}$$

$$S_{jaro-winkler}(a,b) = S_{jaro}(a,b) + lp(1 - S_{jaro}(a,b))$$

m **:** is the number of *matching characters*.
t **:** is half the number of *transpositions*.

l**:** is the length of common prefix.
p**:** is a constant scaling factor (p=0.1).

Related Work

$$S_{jaro}(a,b) = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m}\right) & otherwise \end{cases}$$

$$S_{jaro-winkler}(a,b) = S_{jaro}(a,b) + lp(1 - S_{jaro}(a,b))$$

m **:** is the number of *matching characters*.

t **:** is half the number of *transpositions*.

l**:** is the length of common prefix.

p**:** is a constant scaling factor (p=0.1).

- Given the record a (Andy) and b (Andrew) we find:
  - $m = 3, |a| = 4, |b| = 6$
  - $t = 0$ no transpositions are needed
  - $l = 3$
- $S_{jaro}(a,b) = 0.75$
- $S_{jaro-winkler}(a,b) = 0.75 + 3 \times 0.1 \times 0.25 = 0.825$

Related Work

- Naive Bayes classifier (NBC) and tree augmented naive Bayes classifier (TAN) are also used for record linkage [Elmagarmid, A.K. et al. 2007].
- Extended version of TAN (ETAN) [de Campos, C.P. et al. 2014, 2015].
    - ETAN relaxes the assumption about independencies between features, and does not require features to be connected to the class.

PRL-W

- We assume the existence of a function:

$$cf : a_i \times b_i \to [0,1]$$

  - *cf* is as a measure of how similar two records' fields are.
  - In PRL-FS, the *cf* is exact match function.
  - In PRL-W, the *cf* is Jaro-Winkler similarity function.

- Same as previous work, rather than concern ourselves with the exact value of $cf(a_i, b_i)$ we consider a set of $l_1, \ldots, l_s$ of disjoint ascending intervals exactly covering the closed interval [0,1].

PRL-W

- Given an interval l_k and a record-pair (*a, b*) we define two values:

$$m_{k,i} \text{ is the probability that } cf(a_i, b_i) \in I_k \text{ given that } a \sim b$$

$$u_{k,i} \text{ is the probability that } cf(a_i, b_i) \in I_k \text{ given that } a \neq b$$

$$w_{k,i}(a,b) = \ln(\frac{m_{k,i}}{u_{k,i}})$$

For a matched pair, we expect $m_{k,i}$ is large when $I_k = [0.9, 1]$. (data quality)

For a unmatched pair, we expect $u_{k,i}$ is low when $I_k = [0.9, 1]$. (randomly match)

PRL-W

- In practice, the set of matched pairs, is unknown.
- Therefore, the values $m_{k,i}$ and $u_{k,i}$ are also unknown.
- To accurately estimate these parameters, we need the expectation maximization (EM) algorithm.

PRL-W

E-step: For each pair (a, b) in $A \times B$ compute

$$g(a,b)\uparrow = \frac{p\prod_{(a,b)\in A\times B}\ \prod_{k=1}^{s} m'_{k,i}(a,b)\uparrow}{p\prod_{(a,b)\in A\times B}\ \prod_{k=1}^{s} m'_{k,i}(a,b)\uparrow + (1-p)\prod_{(a,b)\in A\times B}\ \prod_{k=1}^{s} u'_{k,i}(a,b)\downarrow}$$

p is the probability that an arbitrary pair in $A \times B$ is a match

where

$$m'_{k,i}(a,b) = \begin{cases} m_{k,i} & \text{if } cf(a_i, b_i) \in I_k \\ 1 & \text{otherwise.} \end{cases}$$

and

$$u'_{k,i}(a,b) = \begin{cases} u_{k,i} & \text{if } cf(a_i, b_i) \in I_k \\ 1 & \text{otherwise.} \end{cases}$$

PRL-W

M-step: Then recompute $m_{k,i}$, $u_{k,i}$, and p as follows:

$$m_{k,i} = \frac{\sum_{(a,b)\in A\times B} g'_{k,i}(a,b)}{\sum_{(a,b)\in A\times B} g(a,b)}, \qquad u_{k,i} = \frac{\sum_{(a,b)\in A\times B} \tilde{g}'{}_{k,i}(a,b)}{\sum_{(a,b)\in A\times B} 1-g(a,b)}, \qquad p = \frac{\sum_{(a,b)\in A\times B} g(a,b)}{|A\times B|}$$

where

$$g'_{k,i}(a,b) = \begin{cases} g(a,b) & \text{if } cf(a_i,b_i) \in I_k \\ 0 & \text{otherwise.} \end{cases}$$

and

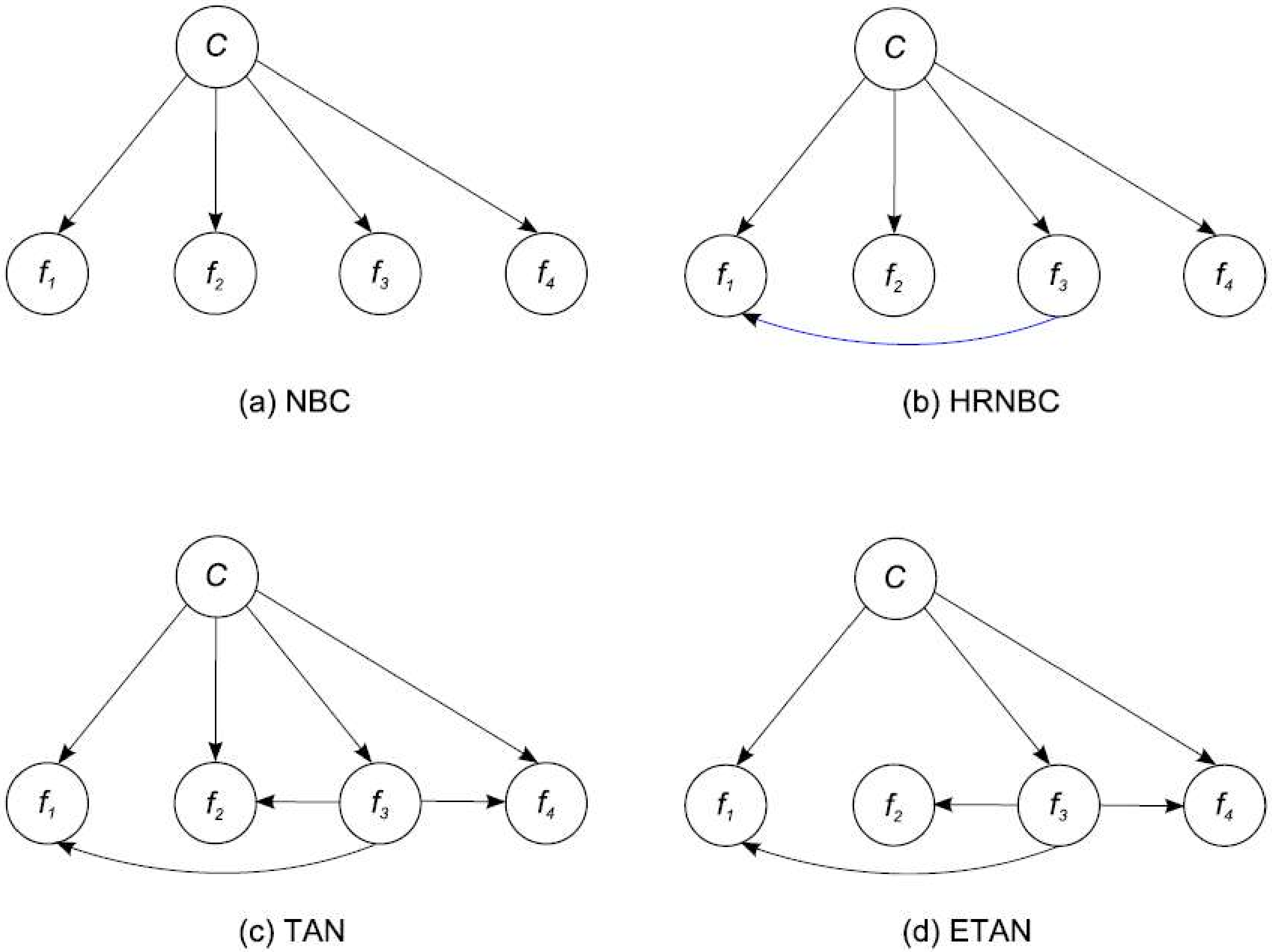
$$\tilde{g}'{}_{k,i}(a,b) = \begin{cases} 1-g(a,b) & \text{if } cf(a_i,b_i) \in I_k \\ 0 & \text{otherwise.} \end{cases}$$

Bayesian Network Classifiers

- Let record-pair feature vector $\overrightarrow{f}$ be a vector contains n features, whose values indicate the distances between two records on specific fields.
- l_k is the state/interval discretised from .
- BN classifiers can calculate the probability of C_k given the feature values (distance for each field-pair).

$$P(C_k|\overrightarrow{\boldsymbol{f}}) = P(C_k) \times \frac{P(\overrightarrow{\boldsymbol{f}}|C_k)}{P(\overrightarrow{\boldsymbol{f}})}$$

Bayesian Network Classifiers



(a) NBC

(b) HRNBC

(c) TAN

(d) ETAN

Fig. 1. The graphical representation of NBC, HR-NBC, TAN, ETAN. The blue arrow represents the dependency introduced by hierarchy feature level information.

ETAN

- Extended TAN (ETAN) is a generalization of TAN and NBC.
- ETAN allows more structure flexibility.
- ETAN's search space of structures includes that of TAN and NBC.
- The score of the optimal ETAN structure is superior or equal to that of the optimal TAN and NBC:

$$\ell(\hat{G}_{ETAN}, D) \geq \ell(\hat{G}_{TAN}, D) \quad and \quad \ell(\hat{G}_{TAN}, D) \geq \ell(\hat{G}_{NBC}, D)$$

ETAN

- In Extended TAN (ETAN) structure learning, minimum spanning tree algorithm cannot be used because any orientation of the arcs between features will not produce the same overall score (unless every feature is connected to the class node).
- Edmonds’ algorithm (finding minimum spanning arborescence)

ETAN

```
• $\hat{s} = -\infty$
• for all $s_D \in S$ do
        (arcs, classAsParent) = ArcsCreation($X, s_D$)
        EdmondsContract(arcs)
        for all root $\in X\backslash\{C\}$
                in = EdmondsExpand(root)
                $G$ = buildGraph($X$, root, in, classAsParent)
                if $s_D(G) > \hat{s}$ then
                        $\hat{G} = G$
                        $\hat{s} = s_D(G)$
• return $\hat{G}$
```

X are variables and S is a set of score functions.

(This algorithm is adapted from [de Campos, C.P. et al. 2015])

Hierarchy Restrictions Between Features

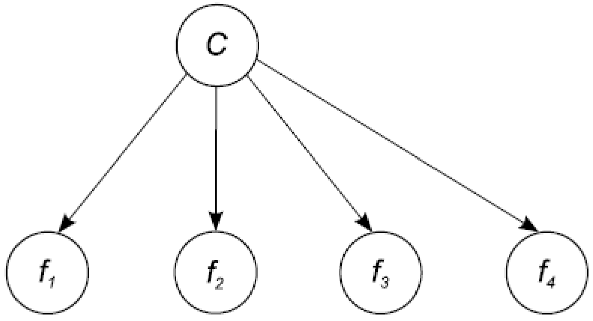
Table 1: Four restaurant records with name, address, borough/town and type information.

Index	Name (f_1)	Address (f_2)	Borough (f_3)	Type (f_4)
1	Strada	Unit 6, RFH Belvedere Rd	Southwark	Roman
2	Strada at Belvedere	Royal Festival Hall	Southwark	Italian
3	Strada	5 Lee Rd	Blackheath	Italian
4	Strada at BH	5 Lee Road	BLACKHEATH	Italian

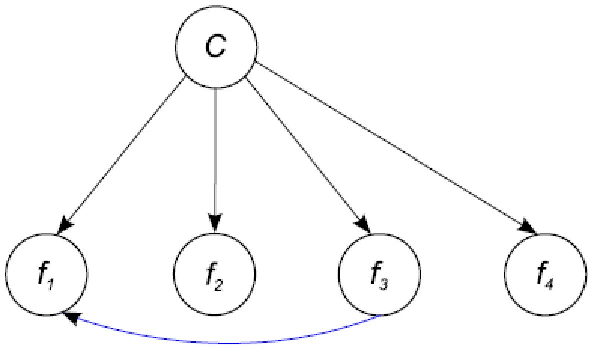
- Table 1 shows four address records, which refer to two restaurants (there are two duplicates). The correct linkage for these four records is: record 1 and 2 refer to one restaurant in Southwark, and record 3 and 4 refer to another restaurant in Blackheath.

- Even record 1 and 3 exactly match with each other in the field of restaurant name, they cannot be linked with each other because they are located in a different borough.

NBC Vs. HR-NBC



(a) NBC



(b) HRNBC

Experiments

- Testing methods
  - PRL-W, TAN, ETAN, NBC and HR-NBC.
- Datasets
  - Country, Company, Restaurant and Tungsten.

    Synthetic data

    Real data

    - Bel-Air Hotel 701 Stone Canyon Rd.
    - Hotel Bel-Air 701 Stone Canyon Road
  - Available at http://yzhou.github.io/.
- Setting
  - Affordable number (10, 50 and 100) of labelled records are used as our training data.
  - The experiments are repeated 100 times in each round.

Experiments

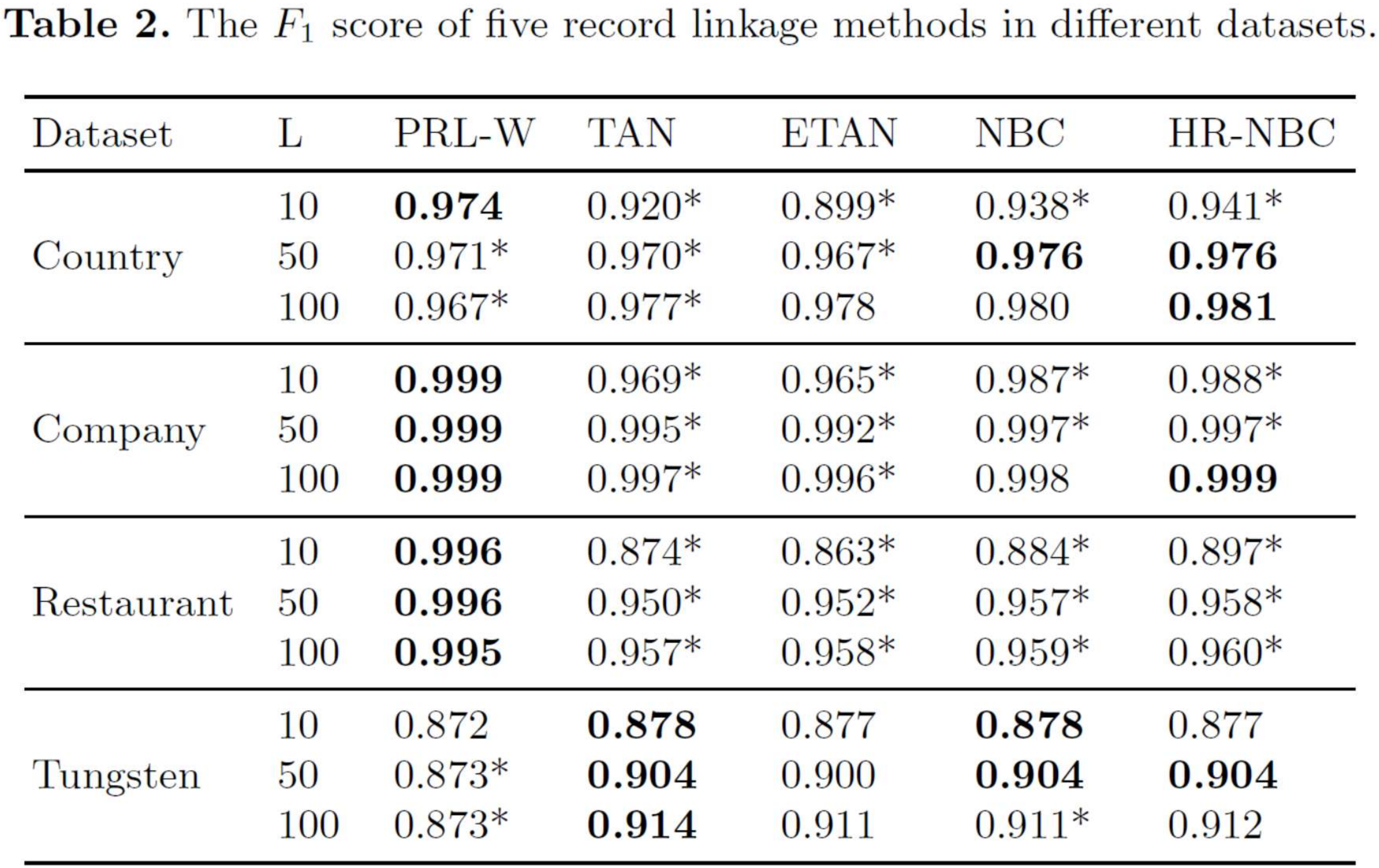


Table 2. The F_1 score of five record linkage methods in different datasets.

Dataset	L	PRL-W	TAN	ETAN	NBC	HR-NBC
Country	10	**0.974**	0.920*	0.899*	0.938*	0.941*
	50	0.971*	0.970*	0.967*	**0.976**	**0.976**
	100	0.967*	0.977*	0.978	0.980	**0.981**
Company	10	**0.999**	0.969*	0.965*	0.987*	0.988*
	50	**0.999**	0.995*	0.992*	0.997*	0.997*
	100	**0.999**	0.997*	0.996*	0.998	**0.999**
Restaurant	10	**0.996**	0.874*	0.863*	0.884*	0.897*
	50	**0.996**	0.950*	0.952*	0.957*	0.958*
	100	**0.995**	0.957*	0.958*	0.959*	0.960*
Tungsten	10	0.872	**0.878**	0.877	**0.878**	0.877
	50	0.873*	**0.904**	0.900	**0.904**	**0.904**
	100	0.873*	**0.914**	0.911	0.911*	0.912

Conclusions

- Findings
  - In settings of limited labelled data, PRL-W works well and its performance is independent of the number of labelled data.
  - TAN, NBC and HR-NBC have better performances than ETAN even the latter method provides better fit to data. (overfitting risk)
  - Compared with NBC, HR-NBC achieves equal or superior performances in all settings, which show the benefits of introducing hierarchy restrictions between features in these datasets.
- Limitations
  - The hierarchy restrictions need to be manually elicited.

Thank you!

Questions?

In Memory of Prof. Sebastian Danicic.