A Supervised Learning Approach for Identifying Duplicated Text Records

Samuel L. Ventura and Rebecca Nugent

Carnegie Mellon University Department of Statistics

Introduction: Why Deduplication?

Which correspond to the **same unique person**?

Last	First	Middle	City	St	Assignee	Year
...	...	...	...	...	...	...
Millar	David	A.	Stanford	CA	Stanford University	2004
Miller	David	A.	Fair Haven	NJ	UNC	1995
Miller	David	A.B.	Stanford	CA	Stanfrod University	2002
Miller	David	Andrew	Stanford	CA	Lucent Technologies Inc.	1996
Miller	David	Andrew	Fair Haven	NJ	Lucent Technologes, Inc.	1996
Miller	David	B.	Los Angeles	CA	Agilent Technologies, Inc.	2000
Miller	David	D.	Billerca	MA	Lucent Technologies, Inc.	2002
...	...	...	...	...	...	...

Source: United States Patent & Trademark Office (USPTO)

Which **addresses** are the same?

Address
...
123 East Main Street, Pittsburgh, Pennsylvania
0123 E. Main St., Pgh., Pa.
123 Main St., East Pittsburgh, PA
123 East Street, Portland, MA
123 East Street, Portland, WA
...

Approach to Deduplication

Goal: **Determine unique entities** in set of records

(1) Compare pairs of text records
(2) Determine pairwise matching probability
(3) Identify clusters of duplicated entities
(4) Consolidate records within clusters

Notation: Record, Field, Similarity, Distance

$\boldsymbol{x_i}$: the $\boldsymbol{i^{th}}$ record, $\boldsymbol{i = 1, \ldots, n}$

$\boldsymbol{x_{im}}$: the $\boldsymbol{m^{th}}$ field of $\boldsymbol{x_i}$, $\boldsymbol{m = 1, \ldots, M}$

γ_{ijm}: the similarity of $\boldsymbol{x_{im}}$, $\boldsymbol{x_{jm}}$

Comparing Pairs of Text Records

Long Strings (e.g. last name): Jaro-Winkler[1]

$$\gamma_{ijm} \in [0, 1]$$
$$\gamma_{ijm} = 1 \text{ if } x_{im} = x_{jm}$$

Short Strings (e.g. state code): Exact matching

$$\gamma_{ijm} = 1 \text{ if } x_{im} = x_{jm}$$
$$\gamma_{ijm} = 0 \text{ if } x_{im} \neq x_{jm}$$

Lists (e.g. co-authors): Intersection / Union

$$\gamma_{ijm} = \frac{|x_{im} \cap x_{jm}|}{|x_{im} \cup x_{jm}|}$$

1 Winkler, W.E. (1990)

Labeled Records and Classification

Manual record identification for 824 inventors[2]
$\Longrightarrow$ **98,762 labeled USPTO inventor records**

(1) **Train classification models**
on pairwise comparisons of labeled records

(2) **Calculate pairwise matching probabilities**
for all record-pairs ($\boldsymbol{p_{ij}} > \mathbf{0.5} \Longrightarrow$ Match)

2 Akinsanmi et al (2013)

Deduplication Method	False Negatives (%)	False Positives (%)
Lai et al (2009)[3]	8.39	4.13
Classification Trees	2.23	2.49
Logistic Regression	1.68	1.64
Random Forests	0.62	0.74

Random Forests: low, balanced error rates

3 Lai et al (2009): weighted sums of similarity scores

Conditioning and Classifier Aggregation

Different types of record-pairs
$\Longrightarrow$ **different matching characteristics**
$\Longrightarrow$ **different classifiers**

(1) Condition on a feature of the record-pairs
(2) Train a classifier on each conditional subset
(3) Use appropriate classifier for prediction

Deduplication Method	False Negatives (%)	False Positives (%)
Lai et al (2009)	8.39	4.13
Random Forests	0.62	0.74
Conditional RF	0.61	0.64

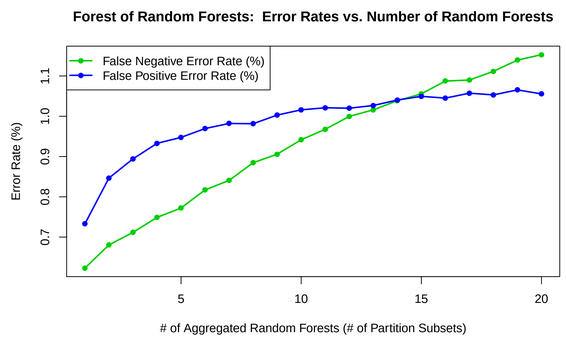
Conditioning further reduces error rates

Forest of Random Forests

"Too much" training data
$\Longrightarrow \binom{98,782}{2} >$ **1 billion** labeled record-pairs

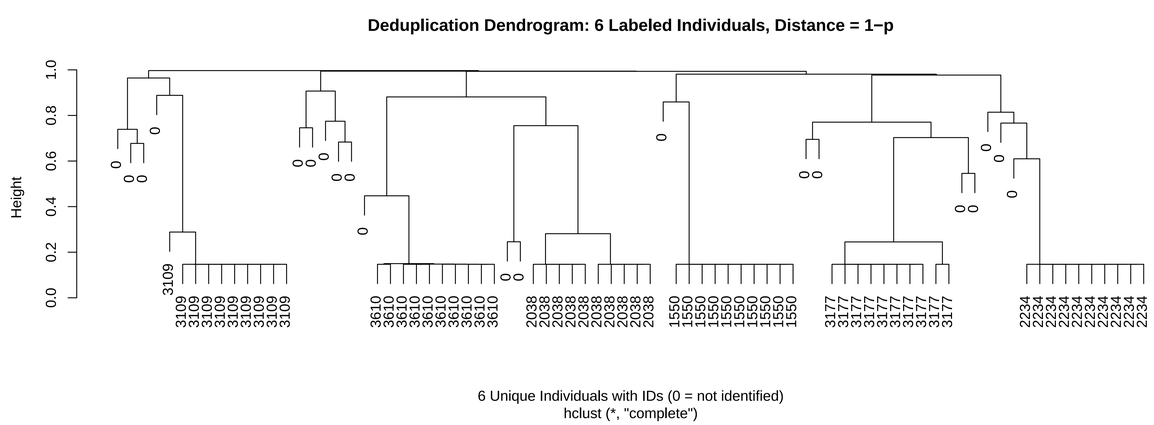
(1) Randomly partition the training data
(2) Train classifiers on each subset
(3) Aggregate (average) predictions of classifiers

Combine this approach with conditioning when any conditional subset is "too large"



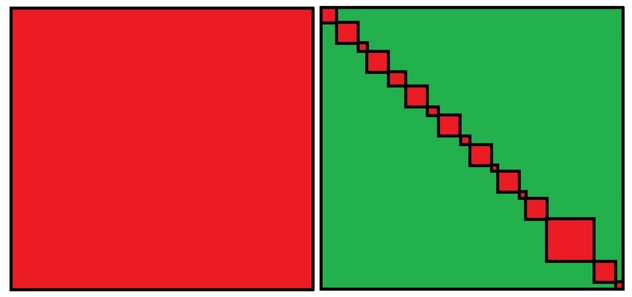
Clustering to Resolve Transitivity Violations

Cluster with $\boldsymbol{D}$, the distance matrix, where $\boldsymbol{D[i, j] = d_{ij} = h(p_{ij})}$ = the distance between $\boldsymbol{x_i}$, $\boldsymbol{x_j}$



Issues & Future Work

Reducing Deduplication Run Time



All $\binom{n}{2}$ comparisons: $\boldsymbol{O(n^2)}$ run time (left)

Solution: Blocking to limit # of comparisons (right)

Current/Future Work: Use labeled training data to find blocking schemes which reduce run time without increasing deduplication error

Calculating Similarity Scores

USPTO subset: 453,972 inventor records
$\Longrightarrow \binom{453,972}{2} \approx$ **100 billion** calculations / field

Solution: Store / re-use repeated comparisons
$\Longrightarrow \binom{20,000}{2} \approx$ **200 million** calculations / field

Automate Choice of Conditioning Scheme

Current/Future Work: Automatically determine the best conditioning scheme given input data

Extending to US Census Contexts

(1) **Labeled Census records**
SSNs are susceptible to error
SSN "labels" for classification?

(2) **Computational feasibility for Census data**
2010 Census: 300 million records
Scale classification/clustering approach?

Acknowledgments

NSF SES-1130706 – NCRN-MN: "Data Integration, Online Data Collection and Privacy Protection for Census 2020"

NSF RTG Grant 25951-1-1121631

NSF Science of Science and Innovation Policy Grant, "Quantifying the Resilience of Industry Innovation Ecosystems" (Award 0830354)

NSF Science of Science and Innovation Policy Grant, "CAREER: Rethinking National Innovation Systems – Economic Downturns, Offshoring, and the Global Evolution of Technology" (Award 1056955)

Carnegie Mellon University
Statistics Department

Corresponding Author: Samuel Ventura. Mail: sventura@stat.cmu.edu