Data Quality
Object Identification - Approximate Joins

IDA 2007 - LJUBLJANA - SLOVENIJA
6-8 September

Hans - J. Lenz, Freie Universität Berlin
September 2007
Agenda

- Introduction
- Data Quality
- Classification
- Pre-selection
- Evaluation
- Summary

Source: M. Neiling, H.-J. Lenz, Lecture on „Methods for Administrative Record Census“, Deutschen Statistische Gesellschaft Münster, May 20th, 2005
Object Identification in Databases

Problem:
Which elements in given databases refer to the same real-world entities\(^1\), respectively?

Example:
Administrative Record Census

1) if global, consistent identifiers are missing
State of the Art

- Two independent development directions:
  - Record Linkage (since the 50ies)
    - Well-proven, robust method
    - Statistical efficiency dominating
    - Mostly used for personal data, e.g. patient information
  - Duplicate-Detection in Databases (since the 80ies)
    - Several methods, e.g. Sorted Neighborhood Method (SNM), Machine Learning,...
    - Performance dominating (SNM, Blocking, Clustering)
  - Trend towards convergence since about five years
Algorithms

Find Partitioning & Compare pairs only within adjacent partition w.r.t. similarity or distance measures, cf. classification

Sorted Neighbourhood Method (SNM) and variants based on sorting, similarity and merging, cf. Hernandez and Stolfo (1998)

Trade-off: Accuracy and Computing Cost
Historical Development

- Newcombe et al.
- Fellegi & Sunter
- Meng & Rubin
- Churches & Christen
- Jaro
- Winkler
- Yancey & Winkler

Record Linkage
(statistical efficiency dominating)

Historical Development

Record Linkage
(statistical efficiency dominating)

Newcombe et al.  Fellegi & Sunter

Jaro  Winkler

Meng & Rubin  Churches & Christen

Workshops

Duplicate-Detection in Databases
(computational performance dominating)

Bitton & DeWitt  Wang et al.

Hernandez & Stolfo  Galhardas et al.

Lim, Srivastava et al.  Bilenko & Mooney

Administrative Record Census

- **ARC**
- **Multiple Databases:**
  Civil Register + Housing Register + Bureau of Labour File +...
  - No global primary key stored in *German* registers
  - errors: misprints, obsolete records, duplicates,...
  - null values: missing values
  - missing entities e.g. illegal residents
Housing and Building Survey

Civil Registers of municipalities

Extraction & Merging for Household Generation

Census Records
ARC in Germany – NSI view

Extraction & Merging & Duplicates Elimination

Census Records (StaLAs)

other Administrative Records (labour force,...)

suppl. Survey (education,...)

Census Test Law and Census Law!
Data Quality – Example

<table>
<thead>
<tr>
<th>No.</th>
<th>ISBN</th>
<th>Title</th>
<th>Name</th>
<th>Year</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-201-54329-X</td>
<td>An introduction to database systems</td>
<td>Date</td>
<td>1997</td>
<td>839</td>
</tr>
<tr>
<td>2</td>
<td>0-201-54329-X</td>
<td>An introduction to database systems</td>
<td>Date</td>
<td>1995</td>
<td>839</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>An introduction to spatial database systems</td>
<td>Gueting</td>
<td>1994</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0-210-14456-5</td>
<td>An introduction to database systems</td>
<td>Date</td>
<td>1977</td>
<td>536</td>
</tr>
<tr>
<td>27</td>
<td>0-201-14456-5</td>
<td>An Introduction to Database Systems</td>
<td>Date</td>
<td>1977</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>An introduction to database systems</td>
<td>Date</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Quality Metrics & Characteristics

- Provision of constraints on data quality
  - Stated by domain experts, or
  - Estimated from samples

- Analysis should be the starting point!
  - Basis for feature selection
  - Main information for pre-selection
Data Quality Metrics & Characteristics

Provision of constraints on data quality

\[ C_k(A_i, Y) \diamond v, \diamond \in \{<, \leq, =, \geq, >\} \]

- Stated by domain experts, or
- Estimated from samples

Analysis should be the starting point!
- Basis for feature selection
- Main information for pre-selection
Data Quality: Semantic Constraints

Proportion of records with missing values:

\[ \text{nulls}(A_i, Y) := P(a \in A_i \subseteq A \mid \exists Y_i \in Y: Y_i(a) = \text{NULL}), \]

e.g. \( \text{nulls}(\text{Addresses}, \text{BirthDate}) = 0.07 \)

Number of duplicates between \( A_i, A_k \subseteq A \):

\[ \text{duplicates}(A_i, A_k) := \left| \{ a \in A_i \mid \exists b \in A_k : a \equiv b \land a.\text{ID} < b.\text{ID} \} \right| \]

\( A \): database table

\( A_i \subseteq A \): Selection from \( A \)

\( Y \): Attribut set

\( |...| \): Set cardinality

\( v \): value

*) estimated from a sample of 250,000 addresses
Characteristics: Semantic Keys

- $Y$ is a **semantic key**, iff
  \[ Y(a) = Y(b) \iff a \equiv b. \]

- $Y$ is a **semantic diff-key**, iff
  \[ \text{dist}(Y(a), Y(b)) \geq \Delta \implies a \not= b \]

- $Y$ is a **$p$-approximate key**, iff
  \[
  \text{accuracy}(Y) := P(Y(a) = Y(b) \mid a \equiv b) \geq p \quad \text{and} \\
  \text{confidence}(Y) := P(a \equiv b \mid Y(a) = Y(b)) \geq p
  \]
Three-Step Identification Procedure

1. Conversion
   - Derive identifying information shared by the sources

2. Comparison
   - Apply sophisticated comparison to pairs of records

3. Classification
   - Use decision rule (e.g. induced from samples)

⇒ Apply the whole process to a pre-selection of pairs only!
Object Identification

Identificator

Classification

Comparison

Conversion

IN

OUT

identical = Yes/No

a pair of records

comparison vector

a pair of values of derived attributes
Comparison – Separability Problem

Comparison shall separate duplicates from non-duplicates

- Example: Separation of book records by Author and Title

Note: the z-axes are logarithmic scaled
Object Identification: Classification I

- Decision Tree Induction
  - Divide and cover the comparison space

- Association Rule Mining
  - Induction of a classification from a set of possibly contradicting association rules
Classification I: Decision Tree

- Input: sampled data (here: 24 pairs)

- Output: Partially ordered set of rules,

  \[\text{if } (\text{year}_1 = \text{year}_2) \text{ and } (\text{name}_1 = \text{name}_2) \text{ then } \text{identical} (\text{book}_1, \text{book}_2)\]

Breiman et al. (1984), Quinlan (1986)
Classification I: Association Rules

- Input: sampled data (here: 24 pairs)
- Output: set of association rules, e.g.

(R1) if both Pages and Year agree
then records refer to the same book
with confidence 100% and support 58%

Agrawal et al. (1993)
Classification II

- Proper Bayesian Classifier
  - Estimation of posterior distribution (AutoClass, cf. Cheeseman et al. (1988))

- Record Linkage
  - Based on Likelihood Ratio Test
  - Estimation of a log-linear model (using EM-algorithm)

\[
\lambda(a,b) = \frac{P\{f(a,b) | (a,b) \text{ is matched}\}}{P\{f(a,b) | (a,b) \text{ is not matched}\}}
\]
Classification II: Proper Bayesian Classifier

C finite set of classes \( c \)

\( \pi(c) \) prior distribution on \( C \)

\( L_x(c) \) = Likelihood for \( c \) given data \( x \)

\( P(c \mid x) \propto L_x(c) \pi(c) \) posterior distribution on \( C \)

select class \( c^* = \arg \max P(c \mid x) \)

* AutoClass Cheeseman et al. (1988)
Record Linkage (+ independence assumption)

<table>
<thead>
<tr>
<th>comparison values</th>
<th>ratio $\lambda_i$ for the $i$-th attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ISBN</td>
</tr>
<tr>
<td>0</td>
<td>9/2</td>
</tr>
<tr>
<td>1</td>
<td>1/2</td>
</tr>
<tr>
<td>2</td>
<td>2/8</td>
</tr>
</tbody>
</table>

\( \lambda \) can be calculated for each comparison case, e.g.

If only ISBN and Name disagree, we get
\[
\lambda(2, 0, 1, 0, 0) = \frac{2}{8} \cdot \frac{10}{6} \cdot \frac{0}{6} \cdot \frac{12}{2} \cdot \frac{7}{4} = 0 < 1,
\]
class 'not identical'.

Conditional independence assumption
Classification III: Neuro-Fuzzy Classifier

• 3-Layer Fuzzy Perceptron ($X,R,C$) with
  - input layer $X$,
  - hidden (linguistic rules) layer $R$,
  - output (class) layer $C$

![Diagram of a 3-layer fuzzy perceptron](image)
Pre-selection of Pairs

- Avoid to build all pairs of records – $O(n^2)$!

- Goal: Improve Performance
  - Main cost: loading of records
  - Scalable solution necessary
Pre-selection of Pairs: Example

- Three relational selectors
- Different combinations \( s = s_i \cup s_j \) or \( s = s_i \cap s_j \)
- Intersection (Example from apartment data):

```
SELECT A.*, B.*
FROM A AS A, A AS B
WHERE A.District = B.District
  AND A.Size >= B.Size - 1 AND A.Size <= B.Size + 1
  AND A.Rooms >= B.Rooms - 0.5 AND A.Rooms <= B.Rooms + 0.5
```

- Efficient processing by means of "Blocking"
Pre-selection of Pairs

- Pre-selection: Intersection/union of selectors,
  - AND/OR for relational selectors
  - Maximum-metric for metric selectors
  - Sort-Merge in general (e.g. for inverted lists)
  ⇒ Parallel execution possible

- Optimization Point of View
  - The more selective the more undetectable duplicates (\(\alpha\)-error increases)
  - The less selective the more processing costs
  ⇒ Greedy optimization (branch & bound)
Sampling of Pairs

- Samples have to be chosen from pre-selection
- Stratified sampling necessary
  - Strata for selectors
  - Stratum for additional duplicates
  - Possibly, an extra stratum for random pairs
- Selectivity of pre-selection determines
  - Proportion of duplicates in the pre-selection sample
  - Offset $\alpha$-error rate (undetectable duplicates)
**Stratified Sampling of Pairs**

A: record table,  N: Sample Size

SAME \(\subseteq A \times A\): contains the duplicate pairs in A

\(N_0\): Number of duplicates  \(N_0 < N\), \(N_0 \leq |SAME|\)

s( . ): Pre-selection \(s(A \times A) \subseteq A \times A\)

Select  \(a \in A\) randomly

IF \(s(\{a\} \times A) \neq \emptyset\) THEN

Randomly select \((a,b) \in s(\{a\} \times A)\)

IF \((a,b) \not\in P\) THEN \(P := P \cup \{(a,b)\}\)

UNTIL size\((P \setminus SAME) = N - N_0\) OR

WHILE size\((P) < N\)

Randomly select \((a,b) \in SAME\)

IF \((a,b) \not\in P\) THEN \(P := P \cup \{(a,b)\}\)

\(*\)* strata for selectors are used implicitly
Correctness Assessment for Classifiers

Precision=true pos/accepted; recall=true pos/matched;
Harmonic mean: $F = \frac{2 \text{ prec} \times \text{ rec}}{\text{ prec} + \text{ rec}}$
Evaluation of Classification Methods

3+2 methods tested on a three datasets benchmark

- Address data, apartment ads, and bibliographic data
- Different parameters were set for classification models
- Three sample sizes chosen (12 samples of pairs for each)
- Samples were split into Learn- & Test-samples

Address data:

- 250,000 records, name, address, birth date information
- Pre-selection chosen: Matching of one Phonetic Code (according to the 'Köln Phonetik Code')
- 12 attributes derived, up to 17 comparison functions considered, e.g. Minimum-Edit- and Bigram-Distances
### Address Data – Comparing Attributes

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Attributname</th>
<th>Vergleichsfunktion</th>
<th>#Werte</th>
<th>Skala</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FNameEdit</td>
<td>fctDiscreteMinEditDistance</td>
<td>7</td>
<td>ordinal</td>
</tr>
<tr>
<td>2</td>
<td>LNameEdit</td>
<td>fctDiscreteMinEditDistance</td>
<td>7</td>
<td>ordinal</td>
</tr>
<tr>
<td>3</td>
<td>ZIP</td>
<td>fctYesNo</td>
<td>2</td>
<td>nominal</td>
</tr>
<tr>
<td>4</td>
<td>LNameBigrams</td>
<td>fctPercentageOfMatchingBiGrams</td>
<td>4</td>
<td>ordinal</td>
</tr>
<tr>
<td>5</td>
<td>FNameBigrams</td>
<td>1−fctPercentageOfMatchingBiGrams</td>
<td>4</td>
<td>ordinal</td>
</tr>
<tr>
<td>6</td>
<td>LNameKoeln</td>
<td>fctHasEqualPhoneticCode</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>7</td>
<td>FNameKoeln</td>
<td>fctHasEqualPhoneticCode</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>8</td>
<td>Town</td>
<td>fctYesNonNull</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>9</td>
<td>Street</td>
<td>fctDiscreteMinEditDistance</td>
<td>7</td>
<td>ordinal</td>
</tr>
<tr>
<td>10</td>
<td>HouseNumber</td>
<td>fctHasNonemptyIntersection</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>11</td>
<td>Year</td>
<td>fctYesNonNull</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>12</td>
<td>Month</td>
<td>fctYesNonNull</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>13</td>
<td>Day</td>
<td>fctYesNonNull</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>14</td>
<td>BirthDate</td>
<td>fctYesNonNull</td>
<td>3</td>
<td>nominal</td>
</tr>
<tr>
<td>15</td>
<td>Sex</td>
<td>fctYesNo</td>
<td>2</td>
<td>nominal</td>
</tr>
<tr>
<td>16</td>
<td>FullNameBigrams</td>
<td>1−fctPercentageOfMatchingBiGrams</td>
<td>5</td>
<td>ordinal</td>
</tr>
<tr>
<td>17</td>
<td>FullNameEdit</td>
<td>fctDiscreteMinEditDistance</td>
<td>7</td>
<td>ordinal</td>
</tr>
</tbody>
</table>
Correctness Results
Address Data

<table>
<thead>
<tr>
<th>error-prob Classifier*</th>
<th>$\alpha$ %</th>
<th>$\beta$ %</th>
<th>Hmean $(\alpha,\beta)$ %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC: AssoClass</td>
<td>1,1</td>
<td>0,2</td>
<td>0,34</td>
</tr>
<tr>
<td>DT: Decision Tree</td>
<td>0,7</td>
<td>0,3</td>
<td>0,42</td>
</tr>
<tr>
<td>RL: Record Linkage</td>
<td>1,5</td>
<td>0,3</td>
<td>0,50</td>
</tr>
<tr>
<td>BC: AutoClass</td>
<td>1,6</td>
<td>0,1</td>
<td>0,19</td>
</tr>
<tr>
<td>NF: NEFClass (Neuro-Fuzzy)</td>
<td>6,6</td>
<td>4,7</td>
<td>5,49</td>
</tr>
</tbody>
</table>

*) the best classification model was selected for each method;
NF is over-fitted
Database „Appartments“

- Appartments
  - N=9.842; \( n_{\text{Doub}} \approx 2.200 \); 13 attributes
  - 12 test samples with 10,000 pairs each;
    \( n_{\text{Doub}} = 12\% \)

- Adresses
  - N=250,000; \( n_{\text{Doub}} \approx 52,000 \); 13 attributes
  - 12 test samples with 10,000 pairs each;
    \( n_{\text{Doub}} = 21\% \)

- Library
  - N= 10,000; \( n_{\text{Doub}} \approx 1.825 \)
  - 12 test samples with 10,000 pairs each
Correctness Results
Appartments

<table>
<thead>
<tr>
<th>error-prob Classifier</th>
<th>$\alpha$ %</th>
<th>$\beta$ %</th>
<th>Hmean $(\alpha, \beta)$ %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AssoClass</td>
<td>0.6</td>
<td>0.1</td>
<td>0.17</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.1</td>
<td>0.1</td>
<td>0.10</td>
</tr>
<tr>
<td>Record Linkage</td>
<td>0.5</td>
<td>0.1</td>
<td>0.17</td>
</tr>
<tr>
<td>AutoClass</td>
<td>0.4</td>
<td>0.3</td>
<td>0.34</td>
</tr>
<tr>
<td>NEFClass</td>
<td>7.2</td>
<td>4.7</td>
<td>5.69</td>
</tr>
</tbody>
</table>

the best classification model was selected for each method; NF is over-fitted
Database „Library“

- Appartments
  - $N=9.842; \ n_{Doub} \approx 2.200; \ 13 \ attributes$
  - 12 test samples with 10.000 pairs each; $n_{Doub} = 12\%$

- Adresses
  - $N=250.000; \ n_{Doub} \approx 52.000; \ 13 \ attributes$
  - 12 test samples with 10.000 pairs each; $n_{Doub} = 21\%$

- Library
  - $N=10.000; \ n_{Doub} \approx 1.825$
  - 12 test samples with 10.000 pairs each
Correctness Results

<table>
<thead>
<tr>
<th>Error-prob Classifier</th>
<th>(\alpha) %</th>
<th>(\beta) %</th>
<th>Hmean ((\alpha,\beta))%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AssoClass</td>
<td>2,5</td>
<td>1,1</td>
<td>1,53</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0,7</td>
<td>0,2</td>
<td>0,31</td>
</tr>
<tr>
<td>Record Linkage</td>
<td>3,5</td>
<td>0,5</td>
<td>0,88</td>
</tr>
<tr>
<td>AutoClass</td>
<td>3,3</td>
<td>0,03</td>
<td>0,06</td>
</tr>
<tr>
<td>NEFClass</td>
<td>12,7</td>
<td>0,6</td>
<td>1,15</td>
</tr>
</tbody>
</table>

The best classification model was selected for each method; NF is over-fitted.
Correctness Results
of all data sets

Error Rates Comparison

NEFCLASS scaled by 1/3
Record Linkage
Effects due to doubling sample size

Correctness for address data

alpha

beta

Record Linkage (10,000)
Record Linkage (20,000)
Association Rules
Effects due to doubling sampling size

Correctness for address data

- Association Rules (10,000)
- Association Rules (20,000)
Decision Tree
Effects due to doubling sampling size

Correctness for address data

- Decision Tree (10,000)
- Decision Tree (20,000)
Conclusion

- **Decision Tree** outperformed the others, is robust

- **Record Linkage** well-behaved for correct specified log-linear models and sufficient large samples

- **Association Rules** allows to control (one of) the errors

- From additional study:
  - **Proper Bayes Classifier Autoclass** works also well,
  - **Neuro-Fuzzy Classifier** possibly over parametrized
Summary – Factors

Latent Data Structure (Separability)

Classification Method

Pre-selection & Sampling

Data Quality

Preprocessing & Conversion

Comparison Space

\[ \alpha, \beta \]
Selected Publications


Selected References


Forgive me today - tomorrow I may no longer feel guilty!
Some Additional Slides
Data Quality: Semantical Constraints II

Ratio of domain size of an attribute set and the size of $A$:

\[
\text{selectivity } (A_i, Y) := \frac{|\{y \in \text{dom}(Y) \mid \exists a \in A_i : y=Y(a)\}|}{|A|},
\]

e.g. \quad \text{selectivity}(\text{Addresses}, \text{BirthDate}) = 0.11,
\quad \text{selectivity}(\text{Addresses}, \text{FullName}) = 0.72,
Approximate Keys

- Likelihood that attribute values coincide for duplicates and vice versa, based on
  - \( \text{accuracy}(Y) := P(Y(a) = Y(b) \mid a \equiv b) \),
  - \( \text{confidence}(Y) := P(a \equiv b \mid Y(a) = Y(b)) \),

**Definition.** \( Y \) is a *approximate key* with confidence \( p \), if both \( \text{accuracy}(Y) \geq p \) and \( \text{confidence}(Y) \geq p \).

- Examples (address data)
  - \( \text{accuracy}(\text{Year}) = .75 \), \( \text{confidence}(\text{Year}) = .98 \)
  - \( \text{accuracy}(\text{Street}) = .87 \), \( \text{confidence}(\text{Street}) = .98 \)
  - \( \text{accuracy}(\text{LastName}) = .93 \), \( \text{confidence}(\text{LastName}) = .99 \)
Constraints - More Examples

- $\Delta$- accuracy(Year) = 0.7613, $\Delta$- confidence(Year) = 0.9217
- $\Delta$- accuracy(Street) = 0.9579, $\Delta$- confidence(Street) = 0.9793
- anti-confidence(Street) = 0.9983
- anti-confidence(BirthDate) = 0.9990
Example: Berlin Online Apartment Advertisements database (BOA)
- 9,842 cleaned ads from Tagesspiegel & Berliner Morgenpost (from May 18th + 25th, 2002)

Semantic constraints
- Reduction of pairs by differentiating keys: 99.876%,
- Expected number of duplicates
  \[ C_k: 537 < \text{duplicates}(A) < 2.636 \] (BOA contains 2.187 duplicates)
- Other \( C_k \)'s, e.g.
Decision Tree: Example for Addresses
Pre-selection Example

- Three relational selectors
- Exhaustive search
All Results for Address Data

Correctness for address data

- Record Linkage (10,000)
- Decision Tree (10,000)
- Association Rules (10,000)
- Record Linkage (20,000)
- Decision Tree (20,000)
- Association Rules (20,000)
4. Literatur

2. Dombrowski, Erik und Lechtenböger, Jens: Evaluation objektorientierter Ansätze zur Data-Warehouse-Modellierung, Datenbank-Spektrum 15/2005
Simsalabim: Data Quality assured!