Australian National University

COMP3430 / COMP8430
Data wrangling

Lecture 18: Record pair classification (2)
(Lecturer: Peter Christen)

Lecture outline

- Cost based classification
- Rule based classification
- Machine learning based classification
- Managing transitive closure

Cost based classification (1)

- In record linkage classification we can make two types of mistakes
  (1) A record pair that is a true match (same entity) is classified as a non-match (**false negative**)
  (2) A record pair that is a true non-match (different entities) is classified as a match (**false positive**)
- Traditionally it is assumed both types of errors have the same costs
- **Question**: *In which applications / situations do these two types of errors have different costs?*

Cost based classification (2)

- If costs for mis-classification are known (or can be estimated), a cost-optimal decision can be made
- Based on the probabilistic record linkage approach (previous lecture), for record pair *r* we can calculate the overall cost *c* as:

  $c(r) = c_{U,U} * P(r \in non\text{-}match, r \in U) + c_{U,M} * P(r \in non\text{-}match, r \in M) +$

  $\qquad c_{M,U} * P(r \in match, r \in U) + c_{M,M} * P(r \in match, r \in M)$

  where the record pair is classified as a *match* or *non-match* while its true match status is *M* or *U*
- The aim is to minimise the overall costs *c* for all record pairs

Rule based classification (1)

- A different approach compared to probabilistic record linkage
- A set of rules is used to classify a record pair as a match or non-match (and possibly a potential match)
- Rules are applied on the calculated attribute similarities, where individual tests are combined using logical operations (AND, OR, NOT)
- The ordering of rules is important if different rules in a rule set classify a record pair into matches and non-matches (i.e. which rules are applied first)
  (several rules might *trigger* (be true) for a given record pair)

Rule based classification (2)

- Example rules:

  $(sim(GivenName)[r_i, r_j] \geq 0.9) \wedge (sim(Surname)[r_i, r_j] = 1.0) \wedge$
  $(sim(BMonth)[r_i, r_j] = 1.0) \wedge (sim(BYear)[r_i, r_j] = 1.0)$:
  $[r_i, r_j] \rightarrow Match$

  $(sim(GivenName)[r_i, r_j] \geq 0.7) \wedge (sim(Surname)[r_i, r_j] \geq 0.8) \wedge$
  $(sim(StrName)[r_i, r_j] \leq 0.6) \wedge (sim(Suburb)[r_i, r_j] \leq 0.6)$:
  $[r_i, r_j] \rightarrow Non\text{-}Match$

Rule based classification (3)

- Rules should have high *accuracy* and high *coverage*
  - High accuracy means they correctly classify record pairs that are covered by the rule into their correct class (of matches and non-matches)
  - High coverage means a rule covers a large number of all record pairs (not just a few)
- Rule sets can be build *manually* or they can be *learned*
  - Manually based on domain knowledge (time-consuming and expensive)
  - Learning based requires training data in the form of true matching and non-matching record pairs

Machine learning based classification (1)

- Machine learning algorithms learn patterns, classes, rules, or clusters from data
- *Supervised* techniques require training data in the form of ground truth (for record linkage: record pairs of true matches and true non-matches)
  - These are classification and regression techniques
  - Example techniques are decision trees, support vector machines, neural networks, logistic regression, Bayesian classifiers, etc.
- Unsupervised techniques do not require training data
  - They cluster similar data points, or extract frequent patterns and rules from data
  - Example techniques are clustering and association rule mining

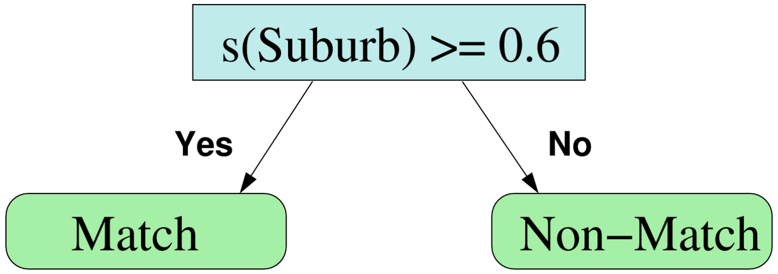
Machine learning based classification (2)

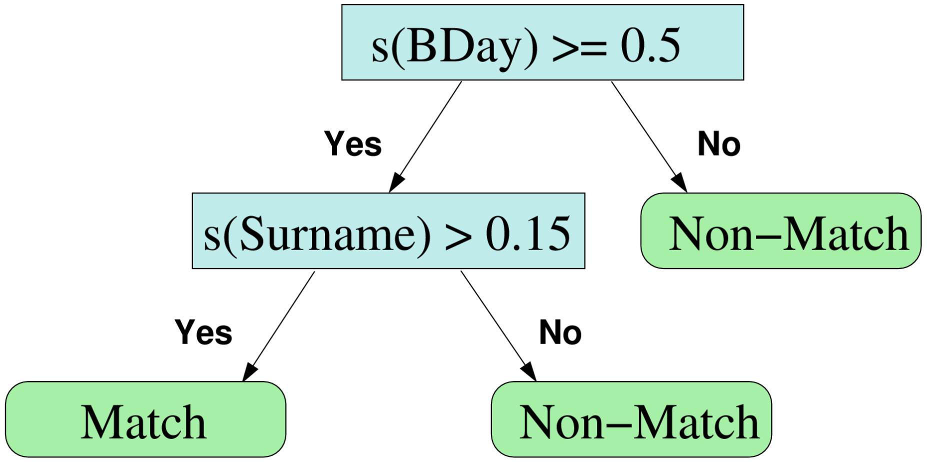
- Many machine learning techniques have been used / adapted for record linkage
- A major challenge for supervised techniques is to obtain training data of good quality and variety
  - Actual truth often not known (for example it is impossible to call all individuals that correspond to true matches)
  - Easy to get clear true matches and non-matches
  - Difficult to get borderline cases (such as same or similar name and different address)
- Another challenge is the class imbalance
  (many more non-matching record pairs compared to matching ones)

Machine learning based classification (3)

- Example: Decision trees learned using a small training data set

GName	SName	StNum	StName	Suburb	BDay	BMonth	BYear	*Class*
0.6	0.8	0.0	1.0	0.6	0.5	0.5	1.0	**M**
0.0	0.15	0.0	0.5	0.0	0.5	0.0	0.75	**U**
0.2	0.0	0.0	0.1	0.15	0.0	0.0	0.75	**U**
0.0	0.25	1.0	0.4	0.6	1.0	1.0	0.75	**M**





Managing transitive closure

- When record pairs are classified individually, the result might be inconsistent with regard to *transitivity*
- If record **a1** is classified as a match with record **a2**, and **a2** is classified as a match with record **a3**, then **a1** and **a3** must also be a match
- Special post-processing and clustering techniques need to be applied