Identifying data set specific duplicate patient records

Scott L DuVall, BS

Department of Biomedical Informatics, University of Utah

Introduction

Probabilistic models are commonly used in the identification of duplicate records. These methods are usually more accurate than deterministic methods, but are exponentially more computationally complex[1]. Thus to make them computationally feasible, they rely on deterministic blocking strategies[2]. This project investigates how machine learning methods can be used to automatically determine an optimal blocking strategy using duplicate records already identified.

Duplicate records – Complete copies or fragmented pieces of a patient's health information spread across multiple computer records thought to belong to separate individuals.

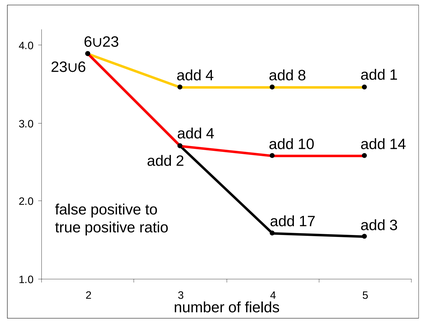
Probabilistic models – Algorithms that use statistics to determine the probability that records match.

Deterministic models – Algorithms that determine whether record pairs match based on exact agreement of a subset of fields.

Blocking strategy – An initial deterministic model run to reduce the number of record pairs into small "blocks" or sets.

Discussion

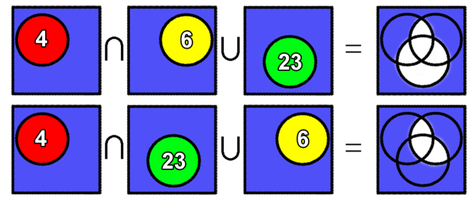
The limitation of heuristic approximation is a propensity for local minima. As demonstrated in the following figure several field combinations may achieve the same accuracy, but the look-ahead modification was necessary to ensure that when a set was added to the solution, it would be optimal for larger set combinations.



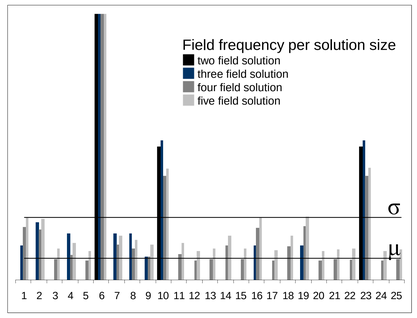
For example, set 6 was determined to be the best set initially. Adding set 23 would bring the solution to 100% sensitivity and it was added to the solution. When the next set is added, though, different specificities were achieved depending on the order in which the sets are combined.

At the second fork, adding set 4 or set 2 to the solution of 23∪6 achieves the same specificity locally, but the look-ahead modification shows us that globally set 2 is a better choice.

Shown graphically below, distribution laws must be taken into account when set unions and intersections are used in combination.



One of the reasons why the heuristic approximation may have preformed well is because the data set had an "early signal." This means that fields that appeared in solutions for a small number of fields also ended up in the final solution. In the following figure, the frequency a field occurred in solutions of different sizes is shown.



Conclusion

Supervised learning techniques such as the modified set covering approximation show promise for increasing accuracy and automation in the identification of duplicate records over traditional untrained probabilistic models.

The results demonstrate that known duplicates can be used for discovering unknown duplicates. Future work will further explore the characteristics of known duplicates and their use in training data set specific models.

References

1 Grannis SJ, Overhage JM, McDonald CJ. Analysis of Identifier Performance using a Deterministic Linkage Algorithm. AMIA Proc. 2002;305-309.

2 Jaro MA. Advances in Record-linkage Methodology Applied to Matching the 1985 Census of Tampa, Florida. J Am Stat Assoc, 89:414-420.

3 Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms, 2/e. The MIT Press, 2001.

Methods

Records were disassembled into individual demographic fields. Matching field values were placed in sets, combined using a heuristic set covering approximation[3], and compared to the set of known duplicates. The following modifications were made:

Set intersection – Traditional set covering doesn't account for false positives and only uses set unions.

Look-ahead – The best sets for one solution size were all compared at the next larger solution size and the best overall was added.

Sensitivity first – As the goal in blocking is to include all duplicates, sensitivity was given first preference. Then specificity was maximized.

Results

The set covering approximation was able to calculate a blocking strategy in polynomial time. Although overhead exists in the preprocessing required for set covering, the reduction in computational complexity makes these methods feasible for use in large data sets. The set covering optimization was able to calculate an optimal solution in 1/10,000th the time of the gold standard - an exhaustive comparison of all possible set combinations.

fields	gold standard	set covering
1	1	8
2	14	20
3	660	79
4	37,500	93
5	1,693,260	165

time in seconds

As the modified set covering method achieves the reduction in complexity through approximation, the blocking strategy calculated cannot be guaranteed to be the globally optimal solution. Nevertheless, experimental results showed that with the modifications, an accurate descriptive model can reliably be calculated. For the first four field solutions, the optimization achieved the same accuracy as the gold standard. For the fifth field solution, the same sensitivity with 99.79% specificity was achieved.

fields	gold standard	set covering
1	*	*
2	3.888	3.888
3	2.706	2.706
4	1.587	1.587
5	1.548	1.551

false positive pairs divided by true positive pairs

* no sets achieved 100% sensitivity

Acknowledgements

This work has been supported by funding from the National Library of Medicine and Robert Wood Johnson Foundation.

Thanks to Reed M Gardner, PhD
Alun Thomas, PhD and
Lisa Cannon-Albright, PhD.

Contact Information

Scott L DuVall
scott.duvall@utah.edu
(801) 581-4080

The University of Utah
Biomedical Informatics