Impact of variations in anonymous record linkage on weight distribution and classification

Daniel Nasseh, Jürgen Stausberg

IBE, Ludwig-Maximilians-Universität München, Germany

Abstract

Anonymous or privacy preserving record linkage is the term for systems allowing the linkage of data from different sources while maintaining an individual's anonymity. This work displays the impact of variations in the process of generating weights in a probabilistic record linkage system on different datasets, the resulting set of weights of candidate pairs and consequently on the final classification process. Furthermore, the results give insight into general problems of current unsupervised classification methods.

Given problem

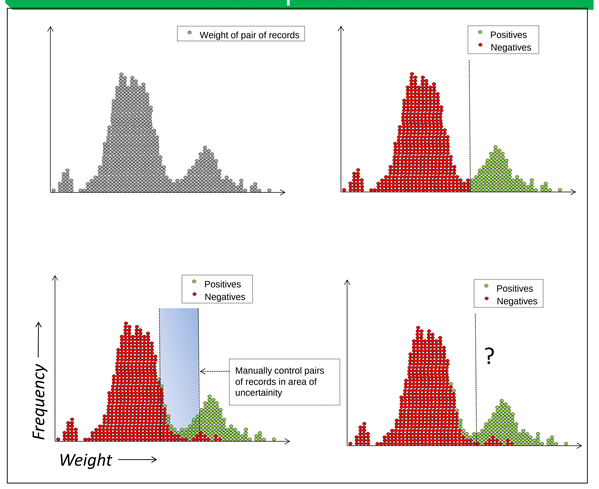

Figure 1: Comparing records during the record linkage process results in specific weights for each comparision. These weights can be rounded and displayed as histograms (a). There is a need to determine a border splitting the record pairs into groups of true and false matches (b). In most cases there are also some false positives and negatives. In case of non-anonymous record linkage it is therefore possible to manually classifiy record pairs in an area of uncertainity (c). In case of anonymous record linkage manual control is not allowed, thus, the predicted border has to be as close as possible to the real optimal border (d).

Context

This work has been a preliminary investigation of classification systems in context of the record linkage process used in a study concerning family-based- cancer of colon in Germany.

Methods (I)

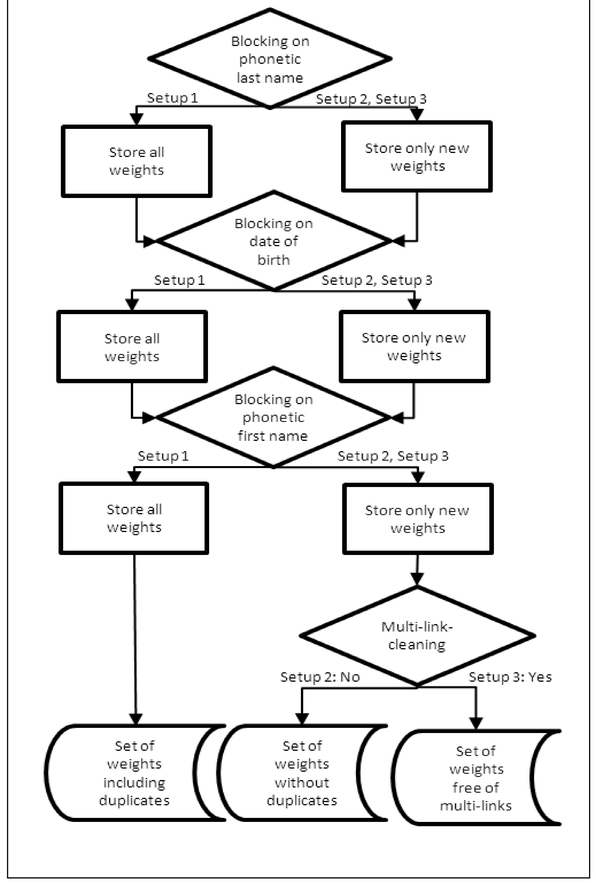

Figure 2: Different configuration setups for performing record linkage resulting in a total of six different sets of weights.

- **System:** Probabilistic Record Linkage System based on the widely used algorithm of Felligi and Sunter. Three different configurations were applied on two different datasets resulting in six different sets of weights.
- **Artificial Dataset:** Artificially created dataset based on attribute occurrences in different publicly available German datasets.
- **Census Dataset:** As a publicly available test set the relatively small 'Census' dataset has been chosen.

Results

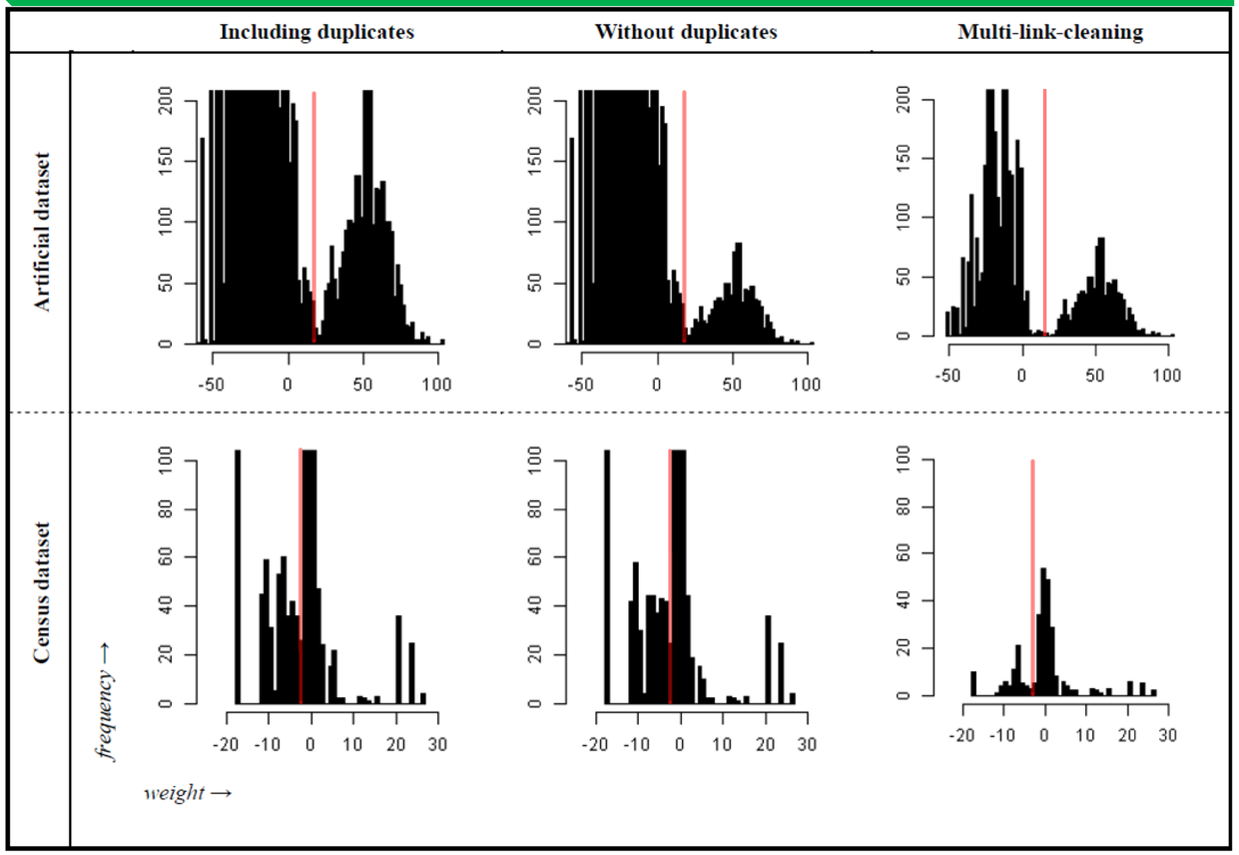

Figure 3: Histogram of weight distribution rounded into Integer values for the results of six different experimental setups. The red line indicates the optimal classifier based on the maximum F-measure. MLC in case of the artificial classifier does not unambiguously indicate the correct classifier, while the other two setups for the same dataset give better indication of the optimal classifier showcasing the danger of only reliing on one result. For the cencus dataset non of the results seem to help predicting a satisfiing classifier.

Methods (II)

- **Blocking:** Blocking is a way to limit the calculation of weights, and therefore decreases computational resources to record pairs which exclusively agree in specific blocking variables. The two different ways of blocking used here differentiate in storing calculated weights for the different blocking variables uniquely or not.
- **MLC (Multi-Link-Cleaning):** MLC is meant to remove all links which include a record which has already been part of another link with a higher weight.

Conclusion

- Using different setups for Record Linkage can help determining the right classifier and consequently improving the quality of the Record Linkage process.
- Reliing on only one setup can lead to deviating classifiers.
- Using unsupervised classification in general can lead to strong misclassifications.
- One should consider using supervised classification.
- There is a need of further research.