

International Journal of Population Data Science

Journal Website: www.ijpds.org



Consistently evaluating data linkage classification results

Peter Christen^{1,2}, Sumayya Ziyad², and Charini Nanayakkara²

¹University of Edinburgh, Edinburgh, United Kingdom

²Australian National University, Canberra, Australia

Objectives

Data linkage is commonly viewed as the problem of classifying record pairs into matches and non-matches. In situations where ground truth data are available, performance measures such as precision, recall, F-measure, sensitivity, and specificity, are commonly used to evaluate the quality of matches obtained with a trained data linkage classifier.

data set perform comparatively. This will allow a better selection of the most suitable classifier for a given data linkage problem and lead to improved quality of linked data sets.

Methods

Comparing multiple classifiers using such measures can, however, lead to inconsistent evaluation because for a given measure the same numerical result can be obtained from different classification outcomes. This can cause a suboptimal classifier being selected and potentially result in linked data sets of poor quality. To overcome this problem, we propose the Consistent Record Linkage (CRL) measure, an application focused evaluation method that ensures data linkage classifiers are assessed in a fair and transparent way. The CRL-measure allows the definition of maximum acceptable error rates, and it provides information about the robustness of a classifier based on identified classification thresholds.

Results

Using both synthetic and real-world data sets, we illustrate how the CRL-measure can provide more detailed information about the performance of data linkage classification results compared to traditional performance measures. Based on user selected maximum acceptable error rates, the CRL-measure identifies the range of classification thresholds where error rates are below these maximums, thereby obtaining high linkage quality. This indicates the robustness of a classifier with regard to a varying classification threshold. Furthermore, the CRL-measure shows a user if a given data linkage classifier is actually able to achieve a certain linkage quality or not.

Conclusion

The CRL-measure provides users with consistent information about how multiple data linkage classifiers trained on the same

