Training data selection for record linkage classification

ABSTRACT

This paper presents a new two-step approach for record linkage, focusing on the creation of high-quality training data in the first step. The approach employs the unsupervised random forest model as a similarity measure to produce a similarity score vector for record matching. Three constructions were proposed to select non-match pairs for the training data, with both balanced (symmetry) and imbalanced (asymmetry) distributions tested. The top and imbalanced construction was found to be the most effective in producing training data with 100% correct labels. Random forest and support vector machine classification algorithms were compared, and random forest with the top and imbalanced construction produced an F1 -score comparable to probabilistic record linkage using the expectation maximisation algorithm and EpiLink. On average, the proposed approach using random forests and the top and imbalanced construction improved the F1 -score by 1% and recall by 6.45% compared to existing record linkage methods. By emphasising the creation of high-quality training data, this new approach has the potential to improve the accuracy and efficiency of record linkage for a wide range of applications.