Supplementary Material 3

Comparison metrics

Sensitivity or true-positive rate is the proportion of true matches that have been correctly identified as matches. It is also known as recall and is measured as follows:

$$\text{Sensitivity} = \frac{\text{Number of TP}}{\text{Number of TP + Number of FN}} \tag{1}$$

Specificity or true-negative rate is the proportion of true non-matches that have been correctly identified as non-matches. It is measured as follows:

$$\text{Specificity} = \frac{\text{Number of TN}}{\text{Number of TN + Number of FP}} \tag{2}$$

This metric is not recommended due to the fact that blocking enables identification of the large number of potential matches leaving a bulk of true non-matches in the comparison space and tending to dominate the calculation.

Positive predictive value (PPV) or precision refers to the proportion of all classified matches that are true matches. It is measured as follows:

$$\text{PPV} = \frac{\text{Number of TP}}{\text{Number of TP + Number of FP}} \tag{3}$$

Ideally both types of errors (FPs and FNs) need to be minimized but there is always an underlying trade-off between sensitivity and precision (when one is higher, the other is invariably lower). F-measure, which is the harmonic mean of sensitivity and PPV, is thus a way of finding the best compromise between the two metrics. It is calculated as follows:

$$\text{f-measure} = 2 * \frac{\text{PPV*Sensitivity}}{\text{PPV + Sensitivity}} \tag{4}$$

Receiver operating characteristic (ROC) curve is plotted with the true positive rate (sensitivity) on the vertical axis against the false positive rate (1-specificity) on the horizontal axis for a varying threshold. Area under ROC curve (AUC) is a single numerical measure between 0.5 and 1 as the ROC curve is always plotted in the unit square, with a random classifier having an AUC value of 0.5 and larger values indicating better classifier performance.

The AUC has the statistical property of being equivalent to the statistical Wilcoxon-Mann-Whitney U-Statistic and is also closely related to the Gini coefficient. While ROC curves are considered robust against skewed class distributions, the problem when using them in data linkage is the number of true negatives, which only appears in the false positive rate. Therefore, this rate will be calculated too low, resulting in too optimistic ROC curves.