Abstract Number - 124

Applying machine learning to improve the accuracy of probabilistic linkage

Robespierre Pita¹, Samila Sena¹, Rosemeire Fiaccone¹, Leila Amorim¹, Mauricio Barreto², Marcos Barreto¹,³, Spiros Denaxas³

¹Universidade Federal da Bahia; ²Oswaldo Cruz Foundation (FIOCRUZ); ³University College London

Record linkage refers to the process of comparing data from different sources and decide if they match (refer to the same entity) or not. This process is widely used in several domains to generate aggregated data to be used for different purposes, such as decision making, monitoring, and assessment studies.

Deterministic or probabilistic methods can be used for record linkage depending on the existence of key attributes common to all data sources involved. Probabilistic approaches pose a set of complex issues related to the choice of attributes, the technique to be used for matching decisions, and the accuracy of resulting data sets. The absence of gold standards to validate probabilistic linkage methods leads the user to the usage of manual review over dubious records (those not classified as true positives or true negatives), but this approach is limited by the amount of data to be considered and is subject to human error.

This paper presents an approach based on supervised and unsupervised machine learning techniques to improve the accuracy of probabilistic record linkage. We discussed recent results obtained within the “100 million cohort” project, a joint collaboration Brazil-UK started in 2013 to conduct a set of epidemiological studies concerning the assessment of health outcomes of individuals receiving cash transfer incomes.

We have performed tests using incremental samples from which we got very high accuracy results during the linkage process further validated through manual review. These samples were used as training data sets for our methods aiming to allow us to scale to 100 million records without the need of using manual review. We also plan to use these methods to perform probabilistic linkages within other two ongoing projects comprising a surveillance platform for Zika and microcephaly and some predictive analytical methods and data integration routines applied over Malaria data.