A Benchmarking System to Evaluate the Effectiveness and Efficiency of Machine Learning Algorithms for Record Linkage

Gurudev. Ilangovan, MS\textsuperscript{1,2}, Mahin. Ramezani, MS\textsuperscript{1,2}, Hye-Chung. Kum, PhD\textsuperscript{1,2}

\textsuperscript{1}Computer Science & Engineering, Texas A&M University, College Station, TX, USA; \textsuperscript{2}Population Informatics Lab, Texas A&M University, College Station, TX, USA

Introduction

Record linkage, often called entity resolution or patient matching, refers to identifying the same person across several databases. In most cases, there is no unique identifiers common to all the databases; so, the available fields common to the databases are compared and a decision is made on whether the two records refer to the same person or not. Most often hybrid systems are used where automated methods are applied in the first pass, followed by manual resolution of the more complex records that the algorithms are uncertain on. In this research we developed a benchmarking system to systematically evaluate the performance of different automated methods under different conditions on different outcomes and to use the system to compare three popular machine learning algorithms for record linkage.

Methods

We developed open source code to infuse different levels of data heterogeneity most often found in record linkage projects (duplicates, twins, suffixes, day-month swaps, first-last name swaps, nick names, last name change due to marriages, typos on names and dates) into any given data (\texttt{github.com/pinformatics/rlErrorGenerator}). The system allows the user to control the overall rate of heterogeneity in the data making it easy to run systematic controlled experiments. We used the system to compare the impact of (1) the data heterogeneity rate and (2) the amount of training data on two performance measures, F1 score and the percentage of the database that require manual review, in three popular machine learning algorithms for record linkage- Random Forests, SVM (RBF kernel) and Dense Neural Networks (DNN with 2 hidden layers). Since the predictions of algorithms are given in terms of probabilities, selecting the uncertain pairs can be achieved by selecting two thresholds T1 and T2, defined based on two measures: Positive Predictive Value (PPV) and Negative Predictive Value (NPV), between which predictions fall into the uncertain category. The two thresholds T1 and T2 are chosen such that the predictions with predicted probabilities greater than T1 or lesser than T2 have to meet the pre-specified performance requirement. All models used 10-fold cross validation for tuning model parameters, and we repeated each experiment 10 times.

Results and Discussion

We found that the record linkage benchmarking system worked well to systematically compare different ML algorithms. There were several interesting findings in our comparison. First, it is very clear that in a hybrid system the price for perfect algorithmic performance in the first pass has to be paid by a lot of manual review in the second pass. We found that when the PPV and NPV requirements are dropped by as little as 1%, the manual review percentage dropped by as much as 50\%. Thus, negligible errors should be allowed during the tuning phase whenever possible to ensure more stable and generalizable models that are not over fit to the training data. Second, acquiring quality labeled data can be expensive and is often resource-intensive. Doing a simple benchmarking study with our proposed open source system during the tuning phase can provide guidance on the optimal training set size. Finally, we found that Random forest outperforms the other models in terms of both F1-score and the heterogeneity in the data as well as requiring the least amount of training data.

Conclusion

Proper use of computing resources typically requires understanding the trade-offs between different algorithms, and selecting the most appropriate algorithm for your data and context. We demonstrate that the benchmarking system is critical to understanding these trade-offs in record linkage algorithms and supports optimal choices. The benchmarking system and machine learning algorithms released on github supports the ecosystem needed for good patient matching.