Consumer Data Research Centre

An ESRC Data Investment

How competitive is propensity score matching using 'address embeddings' with supervised classification for record linkage tasks?

Sam Comber[1], Dani Arribas-Bel[1] and Ronald Nyakairu[2]
[1]University of Liverpool, [2]Local Data Company

Project Background

The growing availability of Open Data has opened the opportunity for businesses to enrich their existing insight from data sources. In particular, a competitive advantage exists in the analysis of integrated data compared to analysing data in isolation. Yet, in reality, most real-world databases are noisy, inconsistent or replete with missing values; these issues complicate the integration of data. The process for disambiguating record pairs that represent the same entity into matches and non-matches is known as record linkage.

In this project, an innovative technique is compared with a supervised method for predicting the match status of address pairs.

Data and Methods

Our record linkage methods are tested on commercial address databases maintained by the Local Data Company (LDC) and Valuation Office Agency (VOA). Given the Cartesian product of the 700,077 and 1,978,830 records for the LDC and VOA databases creates a 1,311 comparison space of candidate record pairs, we require the addresses to be 'blocked' due to computational complexity. This means we partition the number of comparisons for linkage to within mutually exclusive blocks by postcode value.

Once partitioned, two methods for linkage are employed. Our innovative technique uses propensity score matching on the vector representations of addresses learnt from a neural-based language model. Here a 100 dimension vector is learnt using the word2vec algorithm to create an 'address embedding' for every address. Propensity score matching on these vectors is then used to facilitate the linkage.

Our second method begins with using the libpostal address parser to segment each address string into feature columns – for example, business name, street number, street name. These features become the basis for generating 'comparison vectors' for each candidate record pair. Here, each element in the vector is a similarity score between the features columns. This score might, for example, indicate how similar two street names are from one another. The comparison vectors are then classified as matched or non-matched using decision tree ensemble learners. These models are trained on address pairs with a known match status obtained from a previous round of matching between LDC and VOA addresses.

Key Findings

Precision, recall and F1 scores in Table 1 are interpreted as percentages. Our ideal outcome is to minimise misclassification error – i.e. the number of false positives and false negatives. For this reason particular attention is paid to the F1 score which balances the trade-off between false negatives and false positives. As shown, the random forest and XGBoost model far outperform propensity score matching on address embeddings.

Classifier	Precision	Recall	F1 Score
Logit	0.993	0.989	0.991
Random forest	0.994	0.996	0.995
XGBoost	0.994	0.996	0.995
PSM	0.000	0.000	0.000
PSM (blocking)	0.144	0.245	0.181

Table 1. Quality metrics for classifications

Value of the Research

The untidiness of Open Data complicates the process of linkage. Innovation in the methods employed for record linkage problems has the potential to improve the richness of knowledge discovery from data sources when this data can be successfully linked.

Our proposed method for learning the vector representations of addresses for linkage via propensity score matching was only able to match 18.1% of address pairs. Nevertheless, our supervised classification workflow was able to match up to 99.5% of addresses correctly. We attribute this to libpostal's segmentation of addresses into accurate feature columns for comparison. In summary, the availability of open source libraries such as libpostal represents an opportunity for companies to achieve high quality match rates for record linkage tasks.