A Brief Thesis Summary: Large-Scale Classification and Clustering Methods with Applications in Record Linkage

Samuel L. Ventura
Carnegie Mellon University
Department of Statistics

January 17, 2016

Abstract

Record linkage, or the process of linking records corresponding to unique entities within and/or across data sources, is an increasingly important problem. We frame record linkage as a clustering problem, where the objects to be clustered are the records in the data source(s), and the clusters are the unique entities to which the records correspond. We use the following three-step approach to record linkage.

First, records are partitioned into blocks of loosely similar records to reduce the comparison space and ensure computational feasibility, while also avoiding inducing false negative linkage errors. We propose a sequential blocking approach that iterates through a nested set of decreasingly strict blocking criteria to reduce the comparison space more efficiently. This approach, when used in conjunction with hierarchical clustering, offers theoretical guarantees on the probability that records matched at different blocking iterations should be linked, allowing for inference on the linkage results. Second, we adopt a supervised learning approach to estimate the probability that a pair of records matches. When training datasets are prohibitively large, we train ensemble classifiers on subsets of the training data. We propose a new adaptive prediction approach for ensemble classifiers (specifically, random forests) that extracts and incorporates summary statistic information from the distribution of estimated probabilities. Third, we propose a framework for hierarchical clustering in the presence of multiple dissimilarity estimates. We find that we can use properties of the distribution of dissimilarities to model whether or not observations should be linked. In our record linkage context (with distributions of pairwise dissimilarity estimates), we find that we can better estimate the true match probability between pairs of records when using the skew of the distribution and properties of the Beta distribution.

We apply these approaches to three labeled record linkage datasets: a set of labeled inventors from the United States Patent and Trademark Office database, a compilation of lists of death records from the Syrian Civil War contract, and a multi-season roster of National Hockey League players.

1 Introduction to Supervised Record Linkage Approaches

Record linkage, or the process of linking records of unique individuals or entities within and/or across data sources, is an increasingly important problem in a data-rich world. In many real-world applications, records referring to unique entities may exist in multiple data sources (or multiple times within a single data source). These records, however, are all potentially subject to typographical errors; variations in names, addresses, and other identifying information; name changes; repetition of common names; and other features that make it difficult to determine which pairs should be linked, and which should not be linked.

In this thesis, we detail a number of theoretical, methodological, and empirical improvements to algorithms using supervised learning for record linkage. Specifically, we introduce a theoretical framework for a commonly used supervised approach to record linkage in Chapter 1.4. This approach can be simplified into three steps:

1. Partition the data into groups of similar records ("blocking")
2. Within each "block" of records, compare every pair of records, and calculate the pairwise probability of matching using a supervised learning approach
3. Transform these pairwise match probabilities into pairwise dissimilarity estimates and apply (agglomerative) hierarchical linkage clustering to determine which records should be linked

Consequently, the contributions of this thesis align closely with these three steps. These main contributions are outlined in the following three sections.

2 Sequential Blocked Hierarchical Clustering (SBHC)

The idea behind our sequential blocked hierarchical clustering (SBHC) approach is to sequentially partition and cluster the data with a set of increasingly "looser" blocking rules and thresholds, allowing for unique entities that are initially split across blocks to be linked at subsequent iterations. At each iteration of blocking, only "prototype records" from the clusters in the previous iteration are used. This process is repeated until the a priori blocking rules have been exhausted. The algorithm is given in Chapter 6 of the thesis.

Acknowledging that other authors have implicitly used similar multi-stage blocking approaches in some record linkage applications, we introduce an explicit theoretical framework for SBHC in this thesis. While the introduction of a clear, unified notation for this approach is important, the main contributions here are the five theoretical properties we derive when SBHC is used in conjunction with minimax linkage hierarchical clustering Bien and Tibshirani (2012), a particularly useful type of hierarchical clustering for record linkage. These properties are detailed in Chapter 6.3. Briefly, the properties are:

1. Let $x_{v,b,c}$ be a record from cluster C_c in block b at iteration v, and let $p_{v,b^*,c'}$ be the prototype record from cluster c' in block b^* at iteration v. Then, $d(x_{v,b,c}, p_{v,b^*,c'}) \leq \tau_v + \tau_{v+1}$.
2. Let $x_{v,b,c}$ be a record from cluster c in block b at iteration v, and let $x_{v,b^*,c'}$ be a record from cluster c' in block b^* at iteration v. If $p_{v,b,c}$ and $p_{v,b*,c'}$ are linked at iteration $v+1$, then $d(x_{v,b,c}, x_{v,b^*,c'}) \leq 2\tau_v + \tau_{v+1}$.

3. If prototype representation is used in SBHC, inversions are not allowed. (Corollary: If prototypes are not used in SBHC, inversions are allowed.)
4. SBHC with minimax linkage is monotone admissible.
5. SBHC with minimax linkage is point-proportion admissible.

These properties are particularly important in the context of record linkage.

In the context of record linkage, property (1) means we can offer theoretical guarantees about the probability that a record matches the prototype record from another block to which it is subsequently linked. Property (2) means we can offer theoretical guarantees about the probability that a record matches a record from another block to which its prototype record is subsequently linked. Property (3) means that records linked at lower dissimilarity thresholds at earlier stages of SBHC (higher estimated probabilities of matching) would also be linked at higher dissimilarity thresholds at later stages of SBHC (lower estimated probabilities of matching). Property (4) guarantees that the choice of $h(\bullet)$ will not have any effect on the clustering results (as long as the thresholds $\{\tau_v\}_{v=1}^{V}$ are likewise transformed). (In SBHC, pairwise probabilities are transformed via a monotone decreasing function $h(\bullet)$ into pairwise dissimilarities to be used for clustering.) Finally, property (5) is especially useful, since exactly-duplicated records are common. This property guarantees that these duplicated records will not have any effect on the record linkage results. This property also helps us computationally: Since duplicate records will be automatically linked in the record linkage step and will not effect the clustering results, we can link them a priori to reduce the number of pairwise match probabilities that need to be estimated.

3 Prediction with Ensembles using Distribution Summary Statistics (PREDS)

In Chapter 5 of the thesis, methods for taking advantage of additional information from a random forest are introduced, with the goal of making more accurate predictions. Specifically, summary statistics for the distribution of estimated tree probabilities are calculated separately for each observation (here, a pair of records). These summary statistics are used (in conjunction with the known truth/labels from the training dataset) to build a second classifier that models the true class given features of the distribution of estimated tree probabilities. This "stacking"-like approach is flexible with respect to the specific summary statistics used.

The main contributions of this chapter are two-fold:

1. A methodological contribution, introducing a new stacking-like approach for prediction that avoids some potential pitfalls of traditional stacking methods, called "Prediction with Ensembles using Distribution Summary Statistics (PREDS)"
2. Empirical demonstrations of the efficacy of this approach.

Briefly, we demonstrate that our PREDS approach for ensemble prediction outperforms standard prediction approaches (e.g. majority vote, mean predicted probability) and other stacking approaches. Figure 1 shows the misclassification error rates by number of trees for the standard majority vote approach and the PREDS$_{RF}$ approach.

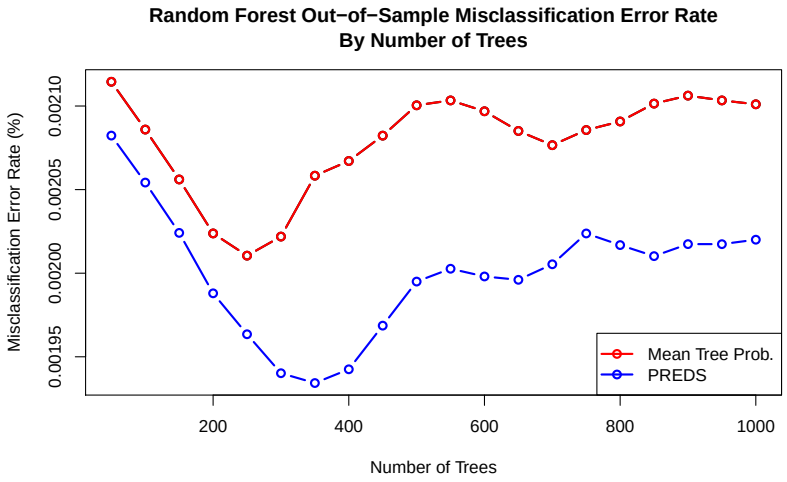

Figure 1: Convergence of the Misclassification Error Rates for Different Random Forest Prediction Methods as the Number of Trees Increases, USPTO Record Linkage Dataset.

The misclassification error rates for the PREDS$_{RF}$ prediction approach are consistently lower than those of the standard majority voting approach and the mean tree probability prediction approach in the USPTO record linkage application. The results shown in this figure reflect the average misclassification error rates across 50 iterations of re-sampling and retraining the models on different subsets of the training/testing data for each value of the number of trees in the random forest. This indicates that the PREDS approach offers consistent improvements to prediction accuracy over standard prediction approaches for tree ensembles.

4 Distribution Linkage Hierarchical Clustering

In Chapter 4 of the thesis, we introduce a method for hierarchical clustering when faced with distributions of distances or distributions of estimated dissimilarities. That is, instead of a single estimate of dissimilarity to be submitted to hierarchical clustering, we instead have a set (or distribution) of dissimilarities. We introduce methodology for summarizing a distribution of dissimilarity estimates for use in hierarchical clustering. This approach was the predecessor of PREDS (discussed above), so we will not detail it here. A more detailed discussion of this approach, along with an associated classification algorithm designed for large training datasets, can be found in Ventura et al (2014).

5 Software

In addition to the contributions outlined above, we plan to release an R package with all code associated with the supervised record linkage approaches from this thesis. This package is in progress, expected to be released in 2016. More details about the software, including an in-depth discussion of the statistical computing techniques used in the software, can be found in the Appendix of the thesis.