Tyler Folkman - GenCo – Machine Learning Entity Resolution

Entity resolution is the problem of identifying and linking different manifestations of the same real-world object. For Ancestry, we have many manifestations of the same person across our users' genealogical trees. Being able to resolve when people are in fact the same allows us to make connections across users' trees and generate what we refer to as a Big Tree. This Big Tree is an attempt to connect all of our users' trees and records in a de-duplicated manner - where all of the same people form a single entity cluster. We have developed a robust, scalable machine learning method to perform entity resolution for a person in a genealogical tree – called GenCo. We find substantial improvements over the previous rule-based system and demonstrate the flexibility of the model to produce more or less precise clusters depending on the needs.

Entity resolution is a long-studied problem that still poses some interesting challenges. This problem is defined as identifying and linking different manifestations of the same real-world object. Many organizations collect data on entities with potential duplicate data. For Ancestry, each member has a personal genealogical tree. These trees often have overlap with other members' trees which creates duplicate entities in our system. One of the steps in the process of generating a Big Tree - a tree consisting of all of our users' tree and record data - is to resolve multiple entities into one cluster.

There are many criteria to consider when establishing what constitutes a match. Previously studied methodologies include techniques such as exact match, distance match, and TF/IDF matching for text data. In this paper, we will propose a novel approach that attempts to replicate the matching criteria of a genealogist and uses a machine learned algorithm to combine all these criteria into a prediction score. To learn such a model we train our algorithm on thousands of pairwise entity comparisons to determine whether two persons are the same or not. We find this approach to provide significant improvements over previous rule-based approaches used for entity recognition.

The machine learning model we use is an open-source version of gradient boosting decision trees called XGBoost. This implementation is highly scalable and provides libraries in multiple languages.

Our algorithm implements a two-step learning process. First, we train a model using only features from the person of interest - excluding family relations. Once trained, this model allows us to evaluate the strength of comparisons between family members. For example, given a comparison between two persons, we can use this model to determine how well their mothers match. This model is learned on only the same and different classes and thus returns a probability that two entities are the same.

Once this model has extracted the most similar relations for mothers, fathers, spouses, 3 children, we can then extract the same features from all these relations. We then train another

XGBoost model using this extended feature set. We refer to this model as the family model because it leverages information from the family relations of the entities.

XGBoost has many hyper-parameters that are important to tune in order to achieve the best results possible. To efficiently tune these parameters we make use of the Python library Hyperopt [9]. Lastly, we use thresholding to convert the probability distribution across classes to actual class labels.

This hierarchical model of boosted trees obtains F1 scores over 25% better than previous models and improves both precision and recall across all classes.