Automated deduplication of pedigree datasets: A tree-based machine learning approach

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Dana-Farber Cancer Institute

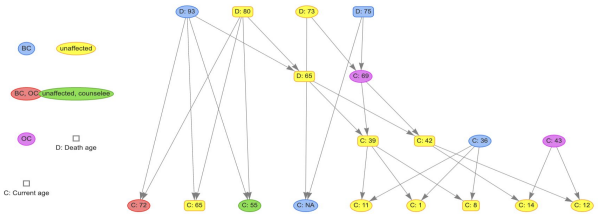
Maria Sol Rosito PhD[1,2], Aleck Cervantes MS[3], Christine Hong MS[3], Joseph Bonner PhD[3], Stephen Gruber MD PhD[3], Danielle Braun PhD[1,2] ; [1]Department of Data Science Dana-Farber Cancer Institute [2]Department of Biostatistics Harvard T.H. CHAN School of Public Health [3]Center for Precision Medicine City of Hope National Medical Center

INTRODUCTION

- Due to the importance of pedigree family data in genetic studies, ensuring quality of datasets is a challenge that needs to be addressed.
- Duplicated records are common due to repeated visits or multiple clinician entries, potentially introducing biases in estimations.
- Variations in record-keeping practices, naming conventions, and inconsistencies in visit records complicate duplicate detection.
- We focus on families carrying TP53 mutations, linked to Li-Fraumeni syndrome, which increases the risk of early-onset cancers.

Goal: Explore tree-based algorithms to identify pedigree duplicates and enhance datasets.

PEDIGREE DATA



- Each pedigree is a family tree in which arrows represent parent-child relationships, and nodes represent individuals.
- Individual information includes sex (depicted by the shape of the symbols), age and deceased status, and cancer diagnoses (e.g. breast and ovarian cancer), among others.
- Duplicated families may vary in the individuals they include and in the data recorded about each individual.
- In our method, pedigrees are converted to direct graphs and the following features are computed:
  - Structural characteristics: number of nodes, number of edges, average degree, graph density, number of connected components, and average shortest path length.
  - Summary of node characteristics: proportion of males, deceased individuals, individuals with any cancer and specific cancer diagnoses, people with positive genetic testing, average age, and set of mutation variants present in the family.

METHODS

Real clinical data of families + Artificially generated duplicates through random individual removal

→ Conversion to pedigree graphs + Computation of features

→ Features for each pair of graphs in each set:
- Absolute difference of each structural characteristic
- Absolute difference of proportions and averages of node characteristics
- Intersection of each set of variants where 0 indicates non-empty and 10 empty

→ Split into training (80%)/test(20%) pedigrees using determinantal point process (DPP) sampling

→ Labels for each pair of graphs in each set: "duplicate" and "non-duplicate"

→ Decision tree/Random Forest + Hyperparameter tuning using cross validation and F1 Score + Analysis of feature importance

RESULTS: Experiment 2

Training and test using families with x.yyyyX>X variant

The dataset includes known duplicate family pairs. One such example is a pair sharing the x.yyyyX>X mutation. One family consist of 56 individuals and the other of 100 individuals. Similar families are generated via data augmentation.

- Decision tree model for all families with x.yyyyX>X
- F1 = 0.7323
- All features related to the number of individuals have importance equal to 0 once the model is trained
- The most important features (61%) are breast cancer fraction and graph density

RESULTS: Experiment 1

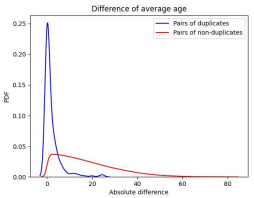
Training and test using the whole dataset

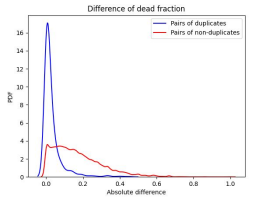
- Random forest, 100 estimators
- F1 in cross-validation: 0.7000 and F1 test: 0.6875
- 84% of pairs identified as duplicates share mutation variant, as expected.
- High precision, low recall. A low precision (many false positives) may lead to information loss due to the removal of false duplicates.

Test confusion matrix:

Prediction / Actual	Negative	Positive
Negative	56890	0
Positive	30	33

Most important features (31%):





TAKEAWAYS

- Tree-based models effectively identify duplicated families with high precision and F1 comparable to state-of-the-art methods (e.g. SNIP algorithm).
- These models are fast and highly interpretable.
- These models can act as feature selectors, removing irrelevant features when trained on reliable data.
- Criteria for selecting the most representative family must be defined after identifying duplicated families.

ACKNOWLEDGEMENTS

This work was supported by the grant NIH/NCI R01CA242218 (Gruber, Amos, Garber)