21/09/2020

A machine learning platform to generate interpretable medically approved, data-responsive clinical decision algorithms

Liamarcia Bifano

ABSTRACT

* Difficult to accurately differentiate generic symptoms or predict diagnoses without advanced probabilistic guidance
  + WHO developed clinical guidelines to direct clinicians to the “most” probable diagnoses and use of resources / antimicrobials
    - Not from representative populations, outdated data, often based on consensus rather than data
    - Static, generic
* Need to directly involve clinicians in machine learning based tool so they understand the black box.
  + Justify why a model is giving a certain prediction
  + Be aware of edge cases where the model underperforms and why

BACKGROUND

* Preventable childhood mortality in sub-Saharan Africa is disproportionate
* Integrated Management of Childhood Illnesses (IMCI) is paper-based, known sort of computerized. Use ePOCT, or other eCDAs.
* When uncertain, clinicians usually err on the side of caution: over-prescription and financial burden
* ML algorithms in ePOCT+ will:
  + Predict diagnostics either as well as ePOCT but with less resource consumption
  + Predict diagnostics to a gold standard
  + Allow clinicians to understand
    - Choose features
    - Explore how inputs affect outputs
    - Help interpret the predictive potential
    - Different algorithms should be standardized so that similar conclusions can be drawn from different models

METHODS

* Remove highly correlated features only if they are also redundant.
* Features should not be removed only based on their individual predictive power (the way it impacts other features should also be taken into account)
* Stopping overfitting: regularization, early-stopping, ensembling, etc
* Least Absolute Shrinkage and Selection Operator (LASSO): linear model where the parameter estimation is restricted to a set using the L1 penalty. The solution to this problem does not have a closed form or analytical solution 🡪 first order optimization estimation. Features with coefficient = 0 can be selected out. Ridge regression can only approach zero for coefficients, so is not as good for automatic feature dropping.
* Boruta: based on random forest model. Shuffle every single column in dataset multiple times, then train a random forest model and compute the feature importance, and compare it with the feature importance of the non-shuffled feature. Feature importance is based on information gain (e.g. gini metric). Then run a hypothesis test and conclude.
* Unbalanced data: under-sample majority class (loss of information) or over-sample minority class (risk of overfitting), or add white noise or random noise.
* Missing data: replace missing values with aggregates (mean / median) ONLY if the values are missing completely at random (MCAR)… otherwise guess them by aggregates using kNN, SVD, etc…

INTEPRETABILITY IN MACHINE LEARNING

* Static algorithms: unable to adapt to seasons, epidemics, resources, etc…
* Clinicians need to know the importance of features to the machine learning model because in scarce-resource environments information is expensive.
* Linear models are well-liked because of their interpretability
* Know which features are useful
* Know which features you can’t do without
* Feature importance: how strongly a feature is associated with the general prediction relative to other features
* feature shape: how the predictions behave when a certain feature changes
* feature interaction: how features behave in presence of other features
* instance level explanation: this task comprises identifying similar samples in the train set and exploring the model’s behavior around that neighborhood
* Partial dependency plot: PDP shows the marginal effect of a set of features in any type of model. Calculates average predicted values across range of feature values
* Individual conditional expectation: ICE - same as above, instead of taking mean of curves, it pulls down every instance of tow subsets (?)
* Accumulated local effects: ALE computes the variation in the output over a feature set neighborhood.
* Forward selection: not the same as above. Model with no features, added features are ordered by performance improvement e.g. using the area under curve (AUC).
* Flatness of Partial Dependency Plot: PDPs with a higher level of variability are more likely to be more important to the response than those with lower variability.
* Permutation Feature Importance: Model reliance is the rational of expected loss under noise over expected loss without noise.
* Local interpretable model agnostic explanations: LIME tries to explain the behavior of a single outcome by modelling the behavior of a black box model in a certain neighborhood around a give single observation.
* Shapley values: aims to divide fairly the values that individual features marginally contributes with a given outcome.
* Shapley Additive Explanations: SHAP is a combination of Shapely values and LIME.
* H-statistic: can capture any form of dependency – not just linear

DATASET AND POPULATION

* For now use imputation based on median / mode, but looking to use matrix factorization
* Use stratification of samples
* Hidden feedback loops affecting data 🡪 need a monitoring system associated to each model

RESULTS:

* No hyperparameter tuning (would take too long)

DISCUSSION: