OpenFE: Automated Feature Generation with Expert-Level Performance

Problem they are trying to solve / Purpose of method

The method tries to improve upon feature generation for ML. Usually feature generation is done manually, which is often very time consuming.

The solution is to automate feature generation. The usual framework is called expand-and-reduce, which creates new features from the base features, then removes ineffective new features. The caveat to this approach is that its hard to evaluate new features, and there is a high amount of candidate features. This is what OpenFE tries to solve.

How does it differ from other methods?

OpenFE does incremental features selection instead of retraining the model on all new features.

How the method works

OpenFE consists of two steps, the Expansion step and the Reduction step.

In the expansion step, features are sorted into numerical and categorical features. Then operators are randomly chosen and applied to the features creating a candidate feature set.

In the reduction step, a model is trained on the original dataset. Then FeatureBoost is used on the predictions together with the candidate feature set. This allows us to evaluate incremental performance of new features instead of a full model retraining. Then a Two-Stage pruning first removes bad features early, based upon the results in FeatureBoost. After that the remaining features are tested together and the best features are selected. Those features are then added to the original dataset.

This process is run until we have all the features we want.