How does TPOT perform on specific biomedical data set problems and how can it be improved?

Technische Universiteit Eindhoven University of Technology

TPOT's performance for Biomedical Data

Tim Beishuizen
Supervisor: Joaquin Vanschoren

Biomedical Data

Challenges
Several biomedical data challenges are known:
- **Volume:** A large or small data set
- **Dimensionality:** A large number of features
- **Complexity:** The data is stored in a complex way
- **Heterogeneity:** The data has differences to different origins
- **Quality:** The quality of the values is low

Skin disease data set

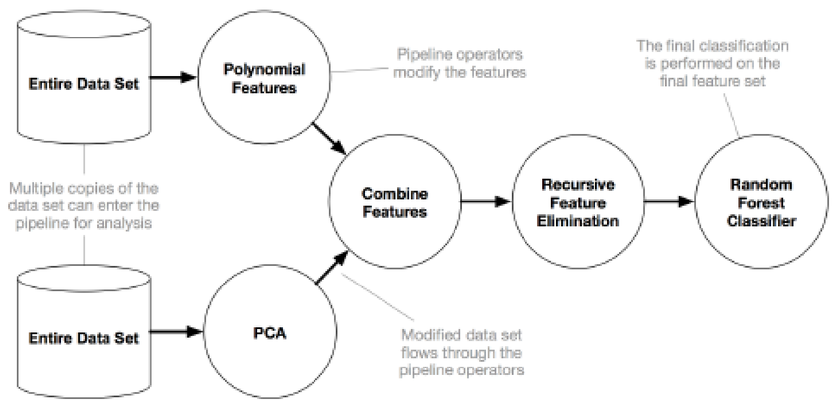
Details:
- Originated from **NCBI**
- Psoriasis and Atopic Dermatitis (Eczema)
- Nine data sets
- The feature set: 54676 genes
- Sample size: between 28 and 180
  - Normal skin
  - Non-lesional skin (not affected by disease)
  - Lesional skin (affected by disease)

The challenges that are available in this data set:
- **Heterogeneity:** Nine data sets needed to be combined
- **Dimensionality:** The 54676 genes
- **Volume:** A very low number of samples

TPOT

Tree-based Pipeline Optimization tool
TPOT is an automated machine learning tool that tries to find the best machine learning pipeline to correclty explain the behaviour of the data.

TPOT Outline



TPOT Algorithms

Algorithm type	Specification	Algorithms
Classifier	Naïve Bayes	GaussianNB, BernoulliNB, MultinomialNB
	Decision Tree	DecisionTree, ExtraTrees, RandomForest, GradientBoosting
	Nearest Neighbor	KNeighbors
	Support Vector Machines	LinearSVC
	Logistic Regression	Logistic Regression
Preprocessors	Scaler	Binarizer, MaxAbsScaler, MinMaxScaler, Normalizer, RobustScaler, StandardScaler
	Feature reduction	PCA, FastICA, RBFSampler, Nystroem, FeatureAgglomeration
	Feature Modifier	Polynomial, OneHotEncoder,ZeroCount
	Feature Selectors	SelectFwe, SelectPercentile, VarianceThreshold, RFE, SelectFromModel

Results

Initialization problems
- **Memory issues**: The high dimensionality slowed down the process significantly
- The feature agglomeration algorithm had to be removed, due to freezing the cpu.
- One data set had an error and needed to be removed.

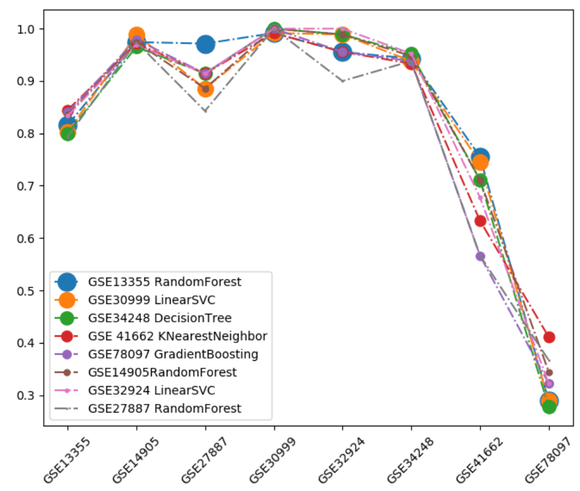
Combined data set pipeline

LogisticRegression(C=0.0001, dual=True, penalty='l2'),
GradientBoostingClassifier(learning rate=0.1, max depth=5,
max features=0.05, min samples leaf=12,
min samples split=19, n estimators=100,
subsample=0.5)

Pipeline Results

Data set	Algorithms	Own score	Combined Data set score
GSE13355	Binarizer RandomForest	0.8333793341383474	0.674509803922
GSE30999	KNearestNeighbor	0.9804301075268818	0.860784313725
GSE34248	RandomForest	0.9666666666666668	0.911764705882
GSE41662	LinearSVC	1.0	0.901960784314
GSE78097	RandomForest	1.0	0.858823529412
GSE14905	LinearSVC	0.9733333333333334	0.905882352941
GSE32924	GradientBoosting	0.8283333333333334	0.835294117647
GSE27887	DecisionTree	0.5142857142857142	0.882352941176

Pipeline scores for each data set



Final score
The final score of the combined data set pipeline was 0.913725490196. It is a reasonable score, considering the values should differ between the different data sets. Comparing this score with the scores of the combined data set with the other data set, this was the best scoring pipeline.

Conclusion

TPOT can tackle the biomedical data challenges reasonably. There are problems regarding memory issues because of the high dimensionality and scoring issues with the low number of samples. However the final result is good.

Since not all biomedical data challenges were present in the data set, future research could be done on those. Also extensions could be made to TPOT to address the found issues to it.