Large scale biomedical data analysis with tree-based automated machine learning

Trang T. Le
University of Pennsylvania
Philadelphia, PA 19104
ttle@pennmedicine.upenn.edu

Weixuan Fu
University of Pennsylvania
Philadelphia, PA 19104
weixuanf@pennmedicine.upenn.edu

Jason H. Moore
University of Pennsylvania
Philadelphia, PA 19104
jhmoore@upenn.edu

ABSTRACT

Tree-based Pipeline Optimization Tool (TPOT) is an automated machine learning (AutoML) system that recommends optimal pipeline for supervised learning problems by scanning data for novel features, selecting appropriate models and optimizing their parameters. However, like other AutoML systems, TPOT may reach computational resource limits when working on big data such as whole-genome expression data. We develop two novel features for TPOT, Feature Set Selector and Template, which leverage domain knowledge, greatly reduce the computational expense and flexibly extend TPOT's application to biomedical big data analysis.

CCS CONCEPTS

• Computing methodologies → Genetic programming;

KEYWORDS

autoML, genetic programming

ACM Reference Format:

Trang T. Le, Weixuan Fu, and Jason H. Moore. 2020. Large scale biomedical data analysis with tree-based automated machine learning. In *Proceedings of the Genetic and Evolutionary Computation Conference 2020 (GECCO '20)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3377929.3397770>

1 INTRODUCTION

For many bioinformatics problems of classifying individuals into clinical categories from high-dimensional biological data, performance of a machine learning (ML) model depends greatly on the problem it is applied to [6, 7]. Automated machine learning (AutoML) systems were developed to automate this challenging and time-consuming process.

Tree-based Pipeline Optimization Tool (TPOT) is a EA-based AutoML system that uses genetic programming (GP) [1] to optimize a series of feature selectors, preprocessors and ML models with the objective of maximizing classification/regression accuracy. While most AutoML systems primarily focus on model selection and hyperparameter optimization, TPOT also pays attention to feature selection and feature engineering by evaluating the complete pipelines based on their cross-validated score. Applying GP with the NSGA-II Pareto optimization [2], TPOT optimizes the accuracy

achieved by the pipeline while accounting for its complexity. Implementation details can be found at TPOT's active Github repository <https://github.com/EpistasisLab/tpot>. Given no *a priori* knowledge about the problem, TPOT has been shown to frequently outperform standard machine learning analyses [5]. Effort has been made to specialize TPOT for human genetics research, but, at the current stage, TPOT still requires great computational expense to analyze large datasets such as in genome-wide association studies (GWAS) or gene expression analyses. Consequently, the application of TPOT on real-world datasets has been limited to small sets of features [3].

In this work, we introduce two new features implemented in TPOT that helps increase the system's scalability. First, the Feature Set Selector (FSS) lets the users pass specific subsets of the features, reducing the computational expense of TPOT at the beginning of each pipeline to only evaluate on a smaller subset of data rather than the entire dataset. Second, Template allows users to specify a desired structure of the resulting machine learning pipeline and thus helps reduce TPOT computation time and potentially provide more interpretable results.

2 METHODS

We introduce a new operator called Feature Set Selector (FSS) that enables biologically guided group-level feature selection. From pre-defined subsets of features, the FSS operator allows TPOT to select the best subset that maximizes average accuracy in k -fold cross validation (5-fold by default). With FSS, users can specify groups of features of interest to reduce the feature space's dimension at pipeline initialization. TPOT learns and selects the most relevant group of features for outcome prediction. Compared to TPOT's existing Selector operators, FSS selects features at the group level instead of individual level.

Parallel with the establishment of the Feature Set Selector operator, we now offer TPOT users the option to define a Template that provides a way to specify a desired structure for the resulting machine learning pipeline, which will reduce TPOT computation time and potentially provide more interpretable results. With a Template defined, each node in the tree pipeline is assigned one of the five major operator types: feature set selector, feature selector, feature transformer, classifier or regressor. Moreover, besides the major operator types, each node can also be assigned more specifically as a method of an operator, such as decision trees as a classifier. An example Template is Feature set selector → Feature transform → Decision trees (Fig. 1).

We apply TPOT with the new FSS operator on both simulated datasets and a real world RNA expression dataset [4] to acquire a comprehensive view of the strengths and limitations of TPOT in the next generation sequencing domain. Detailed simulation

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
GECCO '20, July 8–12, 2020, Cancun, Mexico
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7127-8/20/07.
<https://doi.org/10.1145/3377929.3397770>

Figure 1: Template's general workflow with TPOT-FSS and example pipelines. Final pipelines with optimized parameters are shown for simulated data (top, green) and real-world RNA-Seq data (bottom, mauve).

and analysis code needed to reproduce the results has been made available on the GitHub repository <https://github.com/lelaboratoire/tpot-fss>.

3 RESULTS

For simulated data, across all 100 model fits, the optimal TPOT-FSS pipeline yields an average holdout prediction accuracy of 0.65, while the standard TPOT without FSS and tuned XGBoost models respectively report an average holdout accuracy of 0.48 and 0.49 (Fig. 2, left). This overfitting in the performance of these other two models is likely due to the models' high flexibility such that these models over-learn the training data, especially with the presence of many noisy background features. In 100 replications, TPOT-FSS correctly selects subset with most functional features in 75 resulting pipelines, with the highest average holdout accuracy (0.69 across all 75 pipelines).

For the real-world RNA-Seq data, the optimal TPOT-FSS pipeline yields an average holdout prediction accuracy of 0.68, while the standard TPOT without FSS and tuned XGBoost models produce average holdout accuracies of 0.60 and 0.59 respectively across all 100 model fits (Fig. 2, right). In summary, the optimal models from standard TPOT and XGBoost perform worse than TPOT-FSS. In both cases, separate Welch two-sample one-sided t-tests show TPOT-FSS optimal pipelines significantly outperform those of XGBoost and standard TPOT (all p values $< 10^{-15}$). In 100 replications, TPOT-FSS selects the subset DGM-5 (291 genes) 64 times to be the subset most predictive of the diagnosis status with the highest average holdout accuracy of 0.636 across 64 pipelines.

4 CONCLUSIONS

To our knowledge, TPOT-FSS is the first AutoML tool to offer the option of feature selection at the group level. Previously, it was computationally expensive for any AutoML program to process biomedical big data. TPOT-FSS is able to identify the most meaningful group of features to include in the prediction pipeline. Template allows users to pre-specify a particular pipeline structure based on their knowledge, which speeds up AutoML process and provides more interpretable results. Template was utilized in this study to specify the FSS as the first step of the pipeline, which enables the

Figure 2: In both simulated and real-world expression datasets, TPOT-FSS optimal pipelines significantly outperform those of XGBoost and standard TPOT.

comparison between the two TPOT implementations, with and without FSS.

A limitation of the FSS analysis is the required pre-definition of subsets prior to executing TPOT-FSS. While this characteristic of an intelligent system is desirable when *a priori* knowledge on the biomedical data is available, it might pose as a challenge when this knowledge is inadequate, such as when analyzing data of a brand-new disease. Nevertheless, one can perform a clustering method such as k -means to group features prior to performing TPOT-FSS on the data.

Future extension and integration of these two operators have the potential to enrich the application of AutoML on different real world biomedical problems.

ACKNOWLEDGEMENTS

This research was supported by National Institutes of Health Grant Nos. LM010098, LM012601, A1116794.

REFERENCES

- [1] Wolfgang Banzhaf, Frank D. Francone, Robert E. Keller, and Peter Nordin. 1998. *Genetic Programming: An Introduction: on the Automatic Evolution of Computer Programs and Its Applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [2] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (April 2002), 182–197. <https://doi.org/10.1109/4235.996017>
- [3] Trang T Le, Nigel O Blackwood, Jaclyn N Taroni, Weixuan Fu, and Matthew K Breitenstein. 2018. Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients. *AMIA ... Annual Symposium proceedings. AMIA Symposium 2018* (Dec. 2018), 1358–1367. <https://www.ncbi.nlm.nih.gov/pubmed/30815180>
- [4] Trang T. Le, Jonathan Savitz, Hideo Suzuki, Masaya Misaki, T. Kent Teague, Bill C. White, Julie H. Marino, Graham Wiley, Patrick M. Gaffney, Wayne C. Drevets, Brett A. McKinney, and Jerzy Bodurka. 2018. Identification and replication of RNA-Seq gene network modules associated with depression severity. *Translational Psychiatry* 8, 1 (Sept. 2018), 180. <https://doi.org/10.1038/s41398-018-0234-3>
- [5] Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. 2016. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. *ACM Press*, 485–492. <https://doi.org/10.1145/2908812.2908918>
- [6] Randal S Olson, William La Cava, Zairah Mustahsan, Akshay Varik, and Jason H Moore. 2018. Data-driven advice for applying machine learning to bioinformatics problems. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 23* (2018), 192–203. <https://www.ncbi.nlm.nih.gov/pubmed/29218881>
- [7] Randal S. Olson, William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, and Jason H. Moore. 2017. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining* 10, 1 (Dec. 2017). <https://doi.org/10.1186/s13040-017-0154-4>