Embedding covariate adjustments in tree-based automated machine learning for biomedical big data analyses

Elisabetta Manduchi[1,2], Weixuan Fu[2], Joseph D. Romano[1], Stefano Ruberto[1], Jason H. Moore[1,2,*]

[1]Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

[2]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

[*]Corresponding author: jhmoore@upenn.edu

Abstract

Background

A typical task in bioinformatics consists of identifying which features are associated with a target outcome of interest and building a predictive model. Automated machine learning (AutoML) systems such as the Tree-based Pipeline Optimization Tool (TPOT) constitute an appealing approach to this end. However, in biomedical data, there are often baseline characteristics of the subjects in a study or batch effects that need to be adjusted for in order to better isolate the effects of the features of interest on the target. Thus, the ability to perform covariate adjustments becomes particularly important for applications of AutoML to biomedical big data analysis.

Results

We present an approach to adjust for covariates affecting features and/or target in TPOT. Our approach is based on regressing out the covariates in a manner that avoids 'leakage' during the cross-validation training procedure. We then describe applications of this approach to toxicogenomics and schizophrenia gene expression data sets. The TPOT extensions discussed in this work are available at https://github.com/EpistasisLab/tpot/tree/v0.11.1-resAdj.

Conclusions

In this work, we address an important need in the context of AutoML, which is particularly crucial for applications to bioinformatics and medical informatics, namely covariate adjustments. To this end we present a substantial extension of TPOT, a genetic programming based AutoML approach. We show the utility of this extension by applications to large toxicogenomics and differential gene expression data. The method is generally applicable in many other scenarios from the biomedical field.

2

Keywords

AutoML, covariate adjustment, genetic programming, pathways, feature importance.

Background

Automated machine learning (AutoML) refers to methods which assist (potentially non-expert) users in the optimization of model construction steps such as data preprocessing, feature selection, feature transformations, model selection, and hyperparameter tuning.

The Tree-based Pipeline Optimization Tool (TPOT) [1, 2] is a genetic programming (GP) based AutoML which has been successfully used in biomedical applications including genetics [1], metabolomics [3, 4], and transcriptomics [5]. TPOT explores learning pipelines consisting of arbitrary combinations of selectors, transformers, and estimators (classifiers or regressors). In order to extend its scalability and provide more interpretable results, it is possible to specify a Template for the searched pipelines and to incorporate a Feature Set Selector (FSS) at the beginning of each pipeline to slice the input data set into smaller sets of features allowing the GP to select the best subset in the final pipeline [5].

A desirable feature of TPOT is the ability to adjust for relevant covariates, as this is particularly important in the biomedical context where there are often either baseline characteristics of the subjects or batch effects whose influence on the target or the features needs to be removed so to isolate the actual effects of the features on the target. It is important to note that, while common in biostatistics and epidemiology, covariate adjustment is uncommon and understudied in machine learning.

3

A typical covariate adjustment approach consists in 'regressing the covariates out' of the relevant variables in a data set, which could be the target, or some or all features or both (in case of confounding covariates). Essentially, to regress a set of covariates out of a variable, a predictor model for this variable based on such covariates is first built (e.g. using linear regression if the variable is continuous or logistic regression if it is binary or multiclass) and then the variable is replaced by the residuals from this model [6]. A simplified example of this can be found in [3]. However, the GP optimization in TPOT is based on cross-validation (CV). Consequently, simply regressing out the covariates from the relevant variables *before* feeding the data to TPOT suffers from 'leakage' because, for each CV split, a model built on the training part gets access to information it otherwise would not know, as it uses residuals calculated from a regression which employed the entire data set, i.e. both the training and the testing parts of the CV split. Such leakage can result in overfitting thus reducing the generalizability of the model.

In this work, we present an extension of the TPOT framework, referred to as 'resAdj TPOT' in what follows, which allows covariate adjustment without leakage. To illustrate the usefulness of this, we analyzed a toxicogenomic data set extracted from TG-GATEs [7] representing gene expression array data on kidney tissue from rats exposed to individual drugs known to cause kidney injury. Combining resAdj TPOT with the FSS and Template features, we identified pathways and genes whose expression is associated with creatinine levels in rat kidney, after removing confounding effects such as study batch, compound treatment, dose, and sacrifice time. Creatinine levels provide a broad image of overall kidney health and our findings are very consistent with known kidney biology. Moreover, if we apply TPOT with no covariate

4

adjustments to this data set, the results are different and only tangentially associated with kidney disease. We also show an application to a gene expression dataset from PsychENCODE [8], where the selected covariates are not expected to bear much relevance. In this case we obtain similar results with resAdj and classic (i.e. no adjustments) TPOT, as desirable. The results of our comparisons between resAdj and classic TPOT confirm that our adjustment strategy has been useful in the first case and correctly neutral in the second. Interestingly, we also noted that in the second data set TPOT identified known pathways associated with differential expression between schizophrenic and controls which were not detected by the more typical Gene Set Enrichment Analysis (GSEA) [9].

Methods

TPOT leakage-free covariate adjustment

TPOT leverages the scikit-learn framework [10] and uses GP to evolve machine learning pipelines that consist of selectors, transformers and estimators [1, 2]. The GP seeks to optimize machine learning pipelines with respect to a specified score (e.g. 'balanced accuracy', 'r-squared', etc.) using CV to avoid overfitting on the provided data. Thus, in the pipeline optimization phase, the training set consists of a subset of the input samples (we used 75% at each CV split). At the end of the pipeline optimization procedure, the best pipeline is then trained on the entire set of input samples.

Suppose that, for each of m subjects, we have values for a target variable (binary, multiclass, or continuous) and a collection of n features (binary, multiclass, or continuous). We represent the target values by an m-dimensional vector y and the feature values by an $m \times n$ matrix X. Suppose

5

also that we have values for a collection of covariates that need to be adjusted for. Given a covariate, depending on the context of the data set, it may make sense to adjust only the target y by it, or only a subset of the features (columns in X), or both the target and a subset of the features (the latter is appropriate when the covariate is a confounder). We adjust the values of a variable v (which may be either the target or a feature) by a collection of covariates by 'regressing the covariates out', with the typical approach of fitting an estimator to v on the covariates (e.g. linear regression if v is continuous, or logistic regression otherwise) then replacing the values of v by the corresponding residuals. The latter are obtained by subtracting from the values of v the values predicted by the estimator (if v is continuous) or the expected values based on the estimator predicted class probabilities (if v is binary or multiclass). However, for each CV train-test split, to avoid leakage, *the estimator must be fitted only using the training data for that split* (see Additional file 1 for details). This can be easily achieved for feature covariate adjustments within the current TPOT/scikit-learn framework, but for target covariate (y) adjustment it is necessary to substantially extend the framework.

For feature covariate adjustments, we have added a transformer (*resAdjTransformer*) to TPOT. This needs to be either the first step of any pipeline or the second step after an FSS (TPOT Template can be used to specify these). The initial input to TPOT adds the covariate columns to X. One hyperparameter of this transformer is a file specifying which columns of X should be adjusted by which covariate columns. The transformer applies the no leakage residual adjustments to these columns and removes the covariate columns before passing its output on to the other steps. If no covariate adjustment on the target is needed, classic TPOT can then be run as usual (Figure 1 path A).

For target covariate adjustment, we have added a pre-processor, which creates predefined CV splits and, for each split, adds 2 columns to the dataset; an indicator column to denote the training-testing rows and a column with precomputed no-leakage residuals for *y* using that split. In addition, *y* is fully replaced by the residuals (this is only used for the final Testing Score reported by TPOT, after the optimal pipeline has been determined via CV). The output of this pre-processor (whose structure is outlined in Additional file 2) can then be passed on to TPOT, which must be run with the same CV splits. For each relevant scorer, selector, transformer, and estimator in the classic TPOT, we have added a corresponding scorer, selector, transformer, and estimator which, by using the indicator and corresponding residual columns, enables the pipeline to flow utilizing the appropriately adjusted target at each CV split. The pre-processor can optionally also prepare a hold-out testing set to assess the optimized pipelines output by resAdj TPOT runs.

Note that, when the only needed adjustments are for features, classic TPOT can be run as long as the *resAdjTransformer* is incorporated within every pipeline (using Template, see Figure 1 path A). If the only needed adjustment is for the target, then the pre-processor and the new scorer, selectors, transformers, and estimators must be used with predefined CV splits (Figure 1 path B). If both features and target need adjustment, then all the above can be used together (Figure 1 path C).
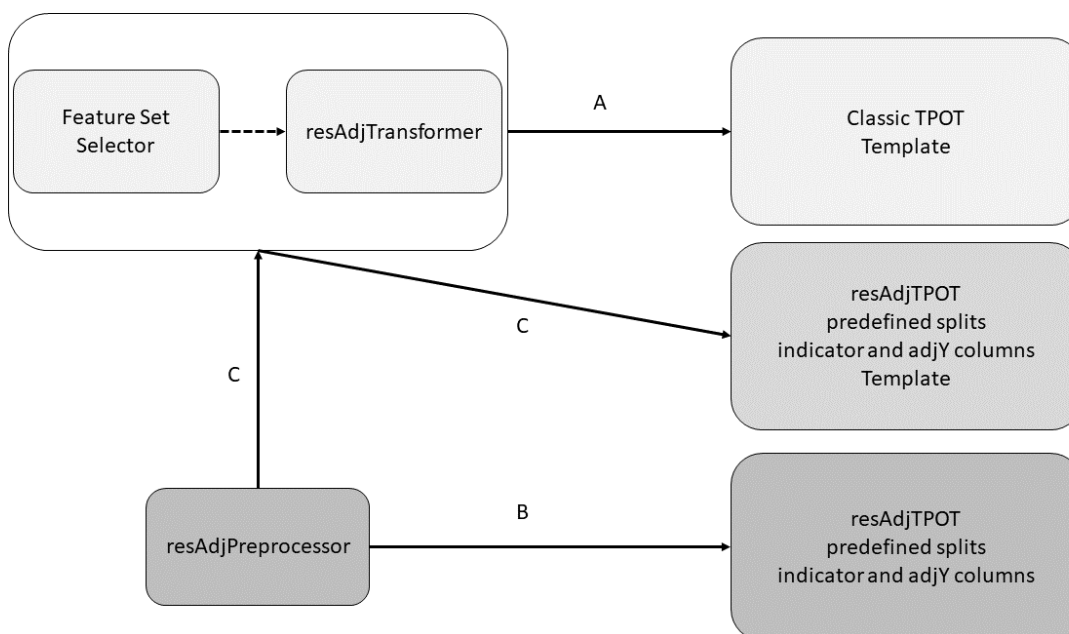
7

Figure 1. The 3 possible workflows for resAdj TPOT. (A) Workflow when the only needed adjustments are for features. The Feature Set Selector (FSS) step is optional. (B) Workflow when the only needed adjustments are for the target. adjY denotes the no-leakage adjustment of the target for each predefined CV split. (C) Workflow when adjustments are needed for both the target and features.

Data sets

TG-GATEs

We selected from TG-GATEs [7] the 933 microarray gene expression assays on kidney tissue from rats in the interquartile range of kidney weight, where each rat was treated with one of 41 compounds in a single dose. The dose was one of 'control' (untreated animal), 'low', 'medium', or 'high' (concentrations varied based on the compound) and the rat was sacrificed after a duration of 3, 6, 9 or 24 hours. We chose creatinine level measured at sacrifice time as the target of interest, because this provides a broad image of overall kidney health. After obtaining the CEL files from accession E-MTAB-799 in Array Express [11], we summarized and normalized (with RMA) the expression values using the Transcriptome Analysis Console (TAC; Affymetrix)

8

software. We encoded dose and sacrifice time by ordinal variables with 4 levels and creatinine measurements by an ordinal variable with 7 levels. We encoded the 41 compounds with 6 binary variables using the binary encoder from the Category-Encoders (http://contrib.scikit-learn.org/category_encoders/). We defined 154 Feature Sets corresponding to the pathways obtained by mapping to rat the MSigDB [12] canonical pathways from KEGG and the pathways annotated in the Rat 230 2.0 Affymetrix Array. The features in our analyses were the 1632 probesets mapping to these Feature Sets.

PsychENCODE

We downloaded the prefrontal cortex normalized RNAseq gene expression data from http://resource.psychencode.org/Datasets/Derived/DER-01_PEC_Gene_expression_matrix_normalized.txt. We extracted 1072 control and schizophrenic individuals with either Caucasian or African American ethnicity, from 3 studies (LIBD_szControl, CMC, and BrainGVEX). Disease status was our target. We encoded each of disease status, sex and ethnicity by binary variables and study by 2 binary variables (using one-hot encoding). We defined 186 Feature Sets corresponding to the MSigDB canonical pathways from KEGG. The features in our analyses were the 4952 genes mapping to these Feature Sets.

Results

TG-GATEs

To fully exploit this large expression data set to identify pathways and genes directly associated to creatinine levels, we needed to factor out the confounding effect of compound treatment. To explore the latter, we clustered the expression data from the 933 assays using *k*-means [13, 14].

We leveraged NbClust [15] to identify the value of k (between 30 and 200) maximizing the Dunn index [16], a measure of cluster quality defined as the ratio between minimal intercluster distance to maximal intracluster distance. The best Dunn index was 0.52 for k=171. We then measured the Biological Homogeneity Index (BHI) of the resulting clustering with respect to various annotations, where the BHI (defined in https://cran.r-project.org/web/packages/clValid/vignettes/clValid.pdf) is a value in the range [0, 1], reflecting the average proportion of pairs in the same cluster with identical annotation. Thus, larger BHI values correspond to more homogeneous clusters in terms of the annotation attribute. We calculated the BHI for each of compound, dose, and sacrifice time. The values were 0.77, 0.31, and 0.66, which indicates that these are important covariates to adjust for, as they are associated to how the data are clustering. We also note that, since TG-GATEs combines data collected from different studies through the years, adjusting all features and target by compound has the additional benefit of removing study batch effects, as typically studies revolved around specific compounds.

To assess the robustness of TPOT with respect to both GP stochasticity and selection of training and testing portions, it is crucial to run the program repeatedly with different random train/test splits and verify the consistency of the results across such runs, in particular in terms of selected pathways. We generated 100 random splits of the data into training (75%) and testing (25%) parts. For each split, we ran resAdj TPOT, adjusting target and all features by the encoded compounds, doses, and sacrifice times (following path C in Figure 1), using 500 generations and a population of 500 in the GP. For each run, the training data set underwent 5 CV splits and the adjustments during training were leakage-free thanks to the approach we described in Methods.

10

Since the target is replaced by residuals, it represents a continuous outcome. We used the Template 'FeatureSetSelector-resAdjTransformer-Transformer-Regressor' and the coefficient of determination (R^2) to score the models. For each run we noted the optimal pipeline, including the selected pathway (i.e. Feature Set) and its score on the held-out testing set. Figure 2 summarizes the results. The 'G-protein signaling' and 'Integrin-mediated cell adhesion' pathways were consistently selected in most of the runs (54% and 28%, respectively) with a median R^2 of ~0.57. Both pathways play known roles in kidney function: G-protein signaling is responsible for mediating response to various types of physical damage to the cell in a broad range of renal diseases [17, 18]. Integrin-mediated cell adhesion plays a crucial structural role in the portions of the kidneys responsible for collecting waste from the circulatory system [19]. Therefore, considering that creatinine is an indicator of kidney health and is not restricted to a specific mechanism of kidney injury, when adjusting for relevant covariates, TPOT preferentially selects biologically plausible pathways.

We also assessed features and covariates by calculating their permutation importance using eli5 v0.10.1 (https://github.com/TeamHG-Memex/eli5) across the 100 runs and then computing the weighted (by testing score) average of the mean score decrease as a percentage of the score . As illustrated in Figure 3, the adjustments by compound and sacrifice time are highly relevant. Moreover, the top 20 features (ranked by importance scores) include genes that are biologically significant in terms of modulating kidney function. For example, the gene with the highest importance score, *Akap9*, codes A-Kinase Anchoring Protein 9, which is highly expressed in kidney tissue, is implicated in retention of T-lymphocytes in kidney tissue under states of inflammation [20], and is one of the most commonly mutated genes in metastatic renal

carcinoma [21]. The gene *Gnb1*, which codes for G Protein Subunit β 1, comprises both the second and third most important features (at two genetic loci), and is highly expressed in kidney glomeruli and tubules [22]. Like *Akap9*, *Gnb1* is also implicated in kidney disease, where downregulation of the gene is associated with worsened prognosis of clear-cell renal cell carcinoma [23], as well as in resistance to tyrosine kinase inhibitor drugs in human models of kidney cancer [24]. Similar patterns can be seen in other important features, although we omit them for brevity.
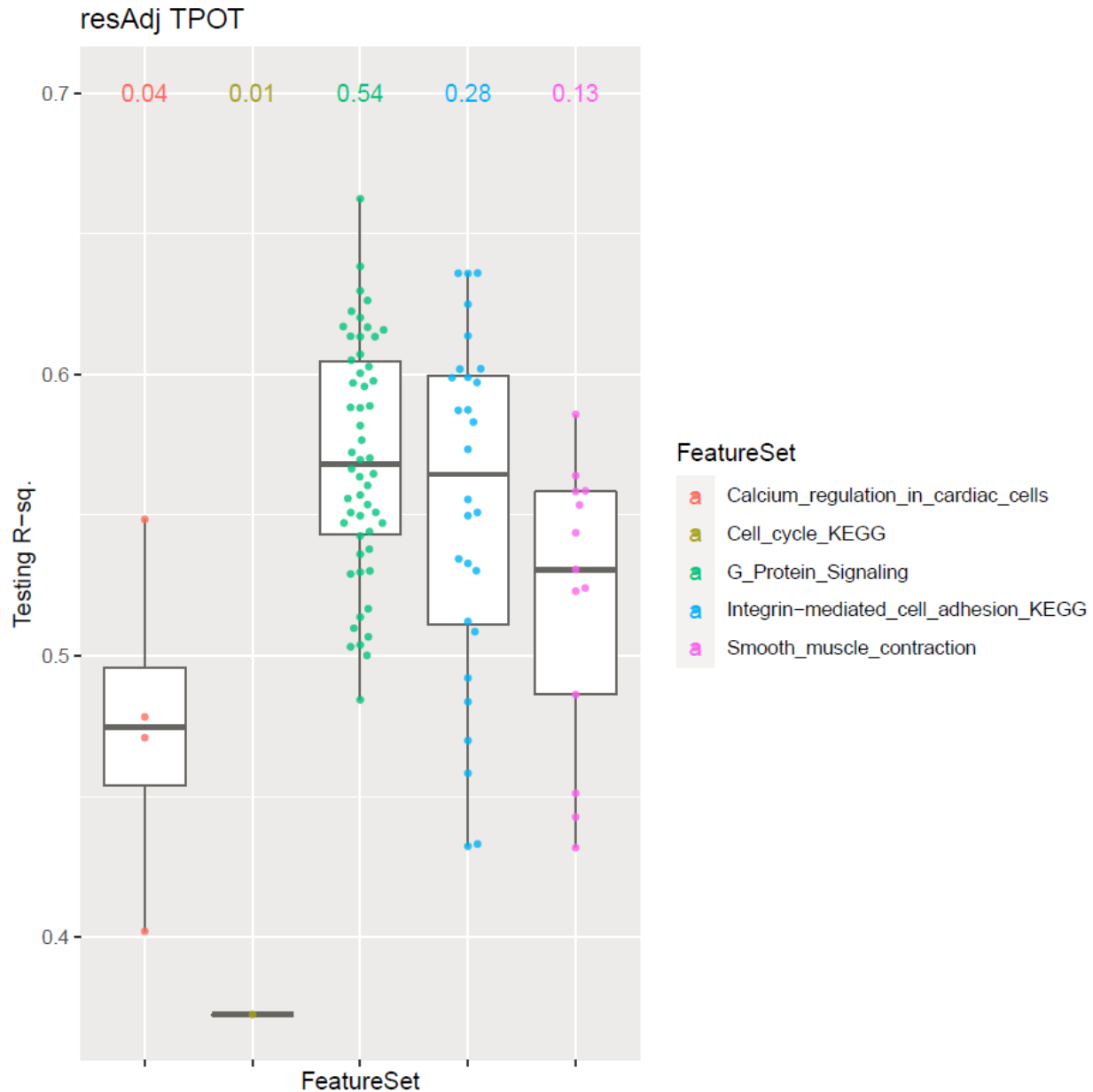
Figure 2. Boxplots for the results of 100 runs of resAdj TPOT on the TG-GATEs data set. Each point corresponds to one run, where the *x*-coordinate indicates the pathway (Feature Set) selected in the optimal pipeline for that run and the *y*-coordinate indicates the R^2 on the held-out Testing dataset.

The clustering results underscore how adjusting for covariates in this data set is crucial to

properly examine the net effect of genes on creatinine levels. Indeed, running classic TPOT (i.e.

no covariate adjustments) on the expression features, yields quite different results. We used the

13

same 100 random splits into training and testing parts and, for each split, we ran TPOT with the template 'FeatureSetSelector-Transformer-Classifier' and balanced accuracy as a scorer since, without adjustments, the target is the original multiclass variable (7 levels of creatinine). Again, we used 500 generations and a population of 500 in the GP. As illustrated in Additional file 3, the most frequently selected pathway was 'Smooth-muscle contraction' (45% of the runs) followed by 'mRNA-processing' (14%). Parenthetically, scores cannot be directly compared with the runs in resAdj TPOT since the scorers are different (balanced accuracy versus R^2). Even though both pathways that were identified as important in the resAdj runs ('G-protein signaling' and 'Integrin-mediated cell adhesion') also show up in some of the classic runs, they are partially obscured by a number of other pathways that are largely uninformative or irrelevant in terms of kidney disease, primarily pathways that mediate the functionality of the heart and blood vessels. Although these are peripherally involved in kidney function due to the kidneys' role in the circulatory system, there are no obvious connections between these pathways and kidney injury in particular.

In terms of feature importance, the difference between classic and resAdj TPOT is striking. Each of the top two most important genes in the covariate adjusted runs (*Akap9* and *Gnb1*, but not the second locus of *Gnb1*) are notably absent from the list of the top 20 features from the unadjusted runs (see Additional file 4), implying that the covariate adjustment filters out spurious correlations that would otherwise obscure mechanistically significant features, at least when applied to data sets with a focus on toxicogenomic response. Rather, the top gene from the classic runs is *Sf3b1*, which is only tangentially associated with kidney disease via copy number variation, through no known mechanism [25].
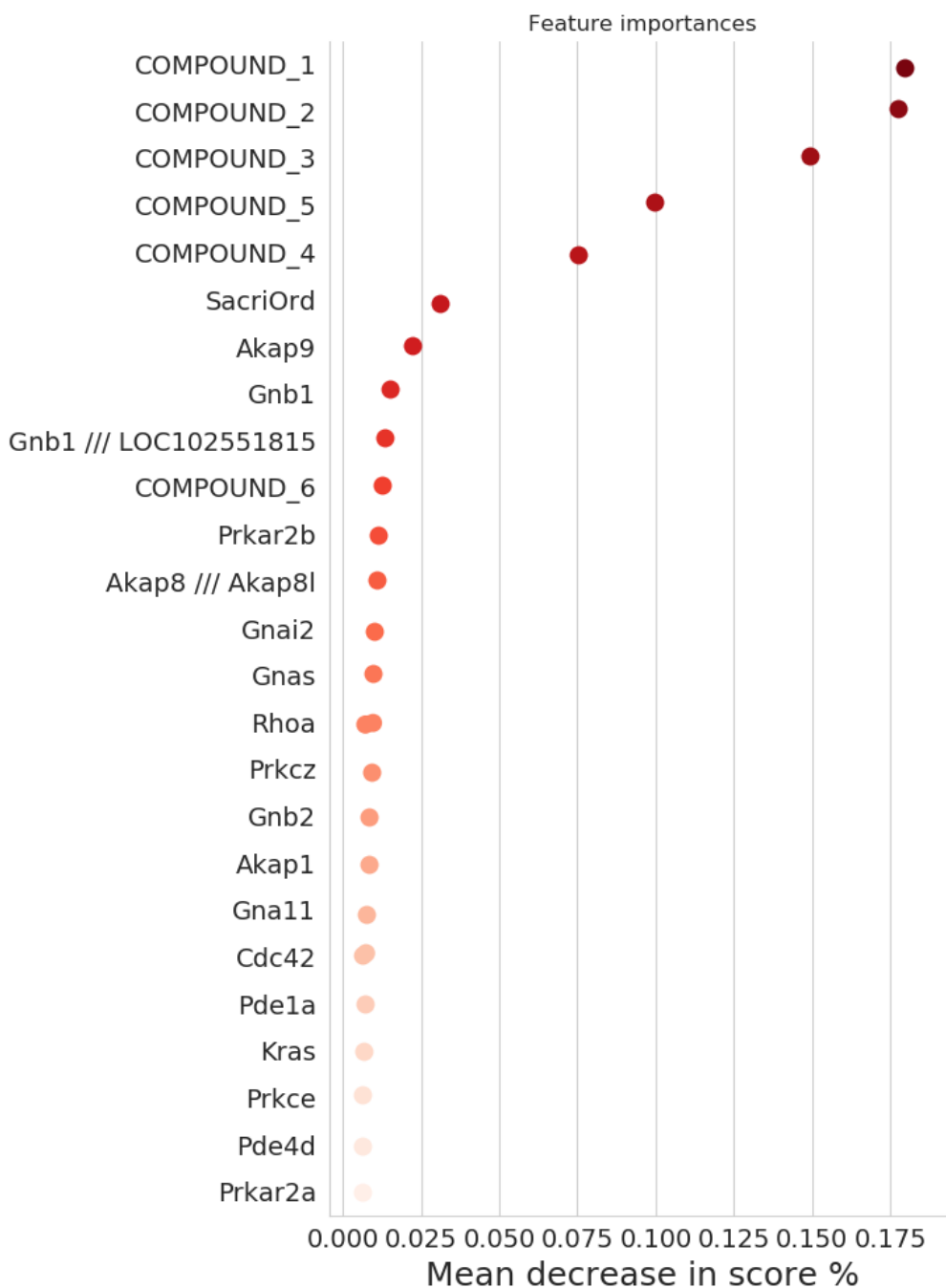
14

Figure 3. Permutation importance from 100 runs of resAdj TPOT on the TG-GATEs data set. The top 20 features and 7 covariates are shown. The gene names are displayed on the *y*-axis and the weighted (by testing score) averages of the mean score decrease as a percentage of the score are displayed on the *x*-axis. The analyses were done at the probeset level, and for *Rhoa* there were two probesets among the top 20 features.

15

3.2. PsychENCODE

In this data set, we considered sex and ethnicity as potential covariates affecting the target (disease status), and ethnicity and study as potential covariates affecting the features (measured gene expression). We therefore ran similar clustering analyses and BHI calculations to those done on TG-GATEs. Varying k from 10 to 240 for clustering of the 1072 expression assays, the best Dunn index was 0.47 corresponding to k=201. The BHI values of the corresponding clustering for the potential covariates affecting gene expression were 0.04 for both ethnicity and study, close to 0 and much smaller than the covariates we analyzed in TG-GATEs, indicating that these covariates do not have a strong effect on the expression. Thus, on these data, it is desirable to get similar results when running resAdj and classic TPOT. Indeed, this was the case. We proceeded similarly to TG-GATEs to set up 100 runs for each of resAdj and classic TPOT, noting the pathways identified in the optimal pipelines and computing permutation importance scores for the genes (see Additional files 5-8 for corresponding plots). In either classic or resAdj modality, for about 80% of the runs the pathway from the optimal pipeline was 'Calcium Signaling'. The latter is an important pathway for the pathophysiology of schizophrenia, as implicated by several previous studies [26–31]. This pathway contained the top strongest genes based on permutation importance in both TPOT modalities, namely *PTAFR*, *GNAL*, *CAMK2G*, and *PRKCG.* The remaining runs, in both modalities, identified three other pathways whose association with schizophrenia is supported in the literature: 'neuroactive ligand-receptor interaction' [27, 30, 32], 'MAPK signaling' [33–35], and 'pathways in cancer' [35]. In addition to these, one of the classic runs identified 'Long-Term potentiation' (albeit with a lower score), which too has been reported as associated with schizophrenia [36]. We note that none of these 5 pathways were identified when we ran the typical GSEA [9] analyses on this data set. GSEA did

not identify any of the MSigDB canonical pathways as significant at FDR<25%. Additional file 9 lists the pathways with GSEA unadjusted p-values < 0.05.

Discussion

An important need for applications of AutoML methods, and machine learning in general, to biomedical data analyses is the ability to incorporate covariate adjustments. In this work, we have presented an approach to adjust the target and/or any subset of the features by a collection of relevant covariates in the context of TPOT, a GP-based AutoML approach. Our method enables adjustments while properly avoiding leakage during pipeline training. We have illustrated its usefulness by an application to a toxicogenomic dataset where adjustments were necessary in order to identify pathways and genes associated to creatinine levels in the presence of confounders. In this application we obtained results which were robust (across 100 runs) and very consistent with known kidney biology. We then applied resAdj TPOT to a different gene expression data set (schizophrenia), obtaining results which were extremely stable and supported by biological literature. In the latter data set we had an indication that confounding was less of an issue, and indeed we observed that resAdj led to the same results as classic TPOT, as desired. In general, though, selecting which covariates are important for adjusting which of the features or target is a delicate step, as different choices could significantly impact the results. Thus, developing approaches and tools aimed at assisting users in this endeavor is as an important area of research that needs to move in parallel with the refinement of approaches like ours.

Another area that needs to be further explored is how to best encode nominal variables (i.e. categorical variables with no natural order relationship) in the context of our method. When the

number of values N of a nominal variable is small, one-hot encoding can be a simple solution. In this case, the variable is replaced by N-1 new binary 'dummy variables'. However, for large values of N, this encoding may not work well. In fact, when we tried it for the compound variable in our TG-GATEs data set, we obtained models with very low R^2 (data not shown). In our case, switching to a binary encoding with six variables led to good results with high testing scores. But a systematic study of the effects of different encoding choices on our method is an important aspect for future research.

Our no-leakage adjustments leverage the typical approach of replacing a variable by its residuals obtained by regressing out the covariates via linear or logistic regression, depending on the type of variable. However, in principle, other estimators could be used, i.e. other classifiers for binary or multiclass variables and other regressors for continuous variables. Possibly, a first pass assessment of the best estimators to use for the adjustments could be done using TPOT itself. Refining and extending resAdj TPOT to this end is another interesting path for future developments.

Conclusions

Our resAdj TPOT approach represents a first step towards addressing a relevant need for AutoML applicability to biomedical big data analyses, where covariate adjustments are often necessary. The applications we presented in this work leveraged toxicogenomics and differential gene expression data. But there are many other scenarios where resAdj TPOT could provide a useful analysis option. For example, it could aid in epistasis analyses of genotype data, where covariate adjustments (e.g. genetic principal components, age, sex, etc.) is typically crucial.

18

More generally, the increased availability of resources such as the UK Biobank [37] providing a rich and large basket of phenotypes and genotypes, enables a plethora of interesting data explorations for which methods such as resAdj TPOT represent a very useful tool. We anticipate this work will have broader impact on machine learning.

List of abbreviations

AutoML = Automated Machine Learning

CV = Cross Validation

FSS = Feature Set Selector

GP = Genetic Programming

GSEA = Gene Set Enrichment Analysis

resAdj = residual adjustment

TPOT = Tree-based Pipeline Optimization Tool

Declarations

Availability of data and materials

We originally obtained the TG-GATEs CEL files used in this study from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-799. These files are currently available from ftp://ftp.biosciencedbc.jp/archive/open-tggates/LATEST/Rat/in_vivo/Kidney/Single/. The complete sample annotation can be obtained from ftp://ftp.biosciencedbc.jp/archive/open-tggates/LATEST/Open-tggates_AllAttribute.zip.

The list of 933 assays we used for our analyses is in Additional file 10.

The PsychENCODE data can be obtained from http://resource.psychencode.org/.

The TPOT extensions discussed in this work are available at

https://github.com/EpistasisLab/tpot/tree/v0.11.1-resAdj.

Funding

Authors' Contributions

EM conceived the TPOT extension, provided code, searched for data sets, analyzed the data, and drafted the manuscript. WF worked on the code and github repository and contributed to method discussions. JDR suggested the TG-GATEs data set, summarized the CEL files, and provided biological interpretations of the TG-GATEs results. SR contributed to several method discussions. JHM supervised the project and contributed to method discussions. All authors reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (USC), Mark Gerstein (Yale), Daniel Geschwind (UCLA), Thomas M. Hyde (LIBD), Andrew Jaffe (LIBD), James A. Knowles (USC), Chunyu Liu (UIC), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Nenad Sestan (Yale), Pamela Sklar (Icahn School of Medicine at Mount Sinai), Matthew State (UCSF), Patrick Sullivan (UNC), Flora Vaccarino (Yale), Sherman Weissman (Yale), Kevin White (UChicago) and Peter Zandi (JHU).

References

1. Olson RS, Urbanowicz RJ, Andrews PC, Lavender NA, Kidd LC, Moore JH. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. In: Squillero G, Burelli P, editors. Applications of Evolutionary Computation. Cham: Springer International Publishing; 2016. p. 123–37.

2. Olson RS, Moore JH. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In: Hutter F, Kotthoff L, Vanschoren J, editors. Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer International Publishing; 2019. p. 151–60. doi:10.1007/978-3-030-05318-5_8.

3. Orlenko A, Moore JH, Orzechowski P, Olson RS, Cairns J, Caraballo PJ, et al. Considerations for automated machine learning in clinical metabolic profiling: Altered homocysteine plasma concentration associated with metformin exposure. In: Biocomputing 2018. WORLD SCIENTIFIC; 2017. p. 460–71. doi:10.1142/9789813235533_0042.

4. Orlenko A, Kofink D, Lyytikäinen L-P, Nikus K, Mishra P, Kuukasjärvi P, et al. Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning. Bioinformatics. 2020;36:1772–8.

5. Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics. 2020;36:250–6.

6. Harrell F. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer-Verlag; 2001. doi:10.1007/978-1-4757-3462-1.

7. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res. 2015;43 Database issue:D921–7.

8. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362. doi:10.1126/science.aat8464.

9. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34:267–73.

10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

11. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update – from bulk to single-cell expression data. Nucleic Acids Res. 2019;47:D711–5.

12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

13. MacQueen J. Some methods for classification and analysis of multivariate observations. The Regents of the University of California; 1967. https://projecteuclid.org/euclid.bsmsp/1200512992. Accessed 29 May 2020.

14. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Appl Stat. 1979;28:100.

15. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. J Stat Softw. 2014;61:1–36.

16. Dunn† JC. Well-Separated Clusters and Optimal Fuzzy Partitions. J Cybern. 1974;4:95–104.

17. Weiss RH. G protein-coupled receptor signalling in the kidney. Cell Signal. 1998;10:313–20.

18. Park F. Accessory proteins for heterotrimeric G-proteins in the kidney. Front Physiol. 2015;6. doi:10.3389/fphys.2015.00219.

19. Viquez OM, Yazlovitskaya EM, Tu T, Mernaugh G, Secades P, McKee KK, et al. Integrin alpha6 maintains the structural integrity of the kidney collecting system. Matrix Biol J Int Soc Matrix Biol. 2017;57–58:244–57.

20. Herter JM, Grabie N, Cullere X, Azcutia V, Rosetti F, Bennett P, et al. AKAP9 regulates activation-induced retention of T lymphocytes at sites of inflammation. Nat Commun. 2015;6:10182.

21. Kim SH, Park WS, Chung J. Tumour heterogeneity in triplet-paired metastatic tumour tissues in metastatic renal cell carcinoma: concordance analysis of target gene sequencing data. J Clin Pathol. 2019;72:152–6.

22. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357.

23. Chen C, Chi H, Min L, Junhua Z. Downregulation of guanine nucleotide-binding protein beta 1 (GNB1) is associated with worsened prognosis of clearcell renal cell carcinoma and is related to VEGF signaling pathway. J BUON Off J Balk Union Oncol. 2017;22:1441–6.

24. Zimmermannova O, Doktorova E, Stuchly J, Kanderova V, Kuzilkova D, Strnad H, et al. An activating mutation of GNB1 is associated with resistance to tyrosine kinase inhibitors in ETV6-ABL1 -positive leukemia. Oncogene. 2017;36:5985–94.

25. Ohashi R, Schraml P, Batavia A, Angori S, Simmler P, Rupp N, et al. Allele Loss and Reduced Expression of CYCLOPS Genes is a Characteristic Feature of Chromophobe Renal Cell Carcinoma. Transl Oncol. 2019;12:1131–7.

26. Lidow MS. Calcium signaling dysfunction in schizophrenia: a unifying approach. Brain Res Brain Res Rev. 2003;43:70–84.

27. Liu Y, Li Z, Zhang M, Deng Y, Yi Z, Shi T. Exploring the pathogenetic association between schizophrenia and type 2 diabetes mellitus diseases based on pathway analysis. BMC Med Genomics. 2013;6:S17.

28. Berridge MJ. Calcium signalling and psychiatric disease: bipolar disorder and schizophrenia. Cell Tissue Res. 2014;357:477–92.

29. Mizoguchi Y, Kato TA, Horikawa H, Monji A. Microglial intracellular Ca2+ signaling as a target of antipsychotic actions for the treatment of schizophrenia. Front Cell Neurosci. 2014;8. doi:10.3389/fncel.2014.00370.

30. Hu Y, Fang Z, Yang Y, Rohlsen-Neal D, Cheng F, Wang J. Analyzing the genes related to nicotine addiction or schizophrenia via a pathway and network based approach. Sci Rep. 2018;8:2894.

31. Nanou E, Catterall WA. Calcium Channels, Synaptic Plasticity, and Neuropsychiatric Disease. Neuron. 2018;98:466–81.

32. Adkins DE, Khachane AN, McClay JL, Åberg K, Bukszár J, Sullivan PF, et al. SNP-based analysis of neuroactive ligand-receptor interaction pathways implicates PGE2 as a novel mediator of antipsychotic treatment response: Data from the CATIE Study. Schizophr Res. 2012;135:200–1.

33. Kyosseva SV, Elbein AD, Griffin WS, Mrak RE, Lyon M, Karson CN. Mitogen-activated protein kinases in schizophrenia. Biol Psychiatry. 1999;46:689–96.

34. Funk AJ, McCullumsmith RE, Haroutunian V, Meador-Woodruff JH. Abnormal activity of the MAPK- and cAMP-associated signaling pathways in frontal cortical areas in postmortem brain in schizophrenia. Neuropsychopharmacol Off Publ Am Coll Neuropsychopharmacol. 2012;37:896–905.

35. Maschietto M, Tahira AC, Puga R, Lima L, Mariani D, da Silveira Paulsen B, et al. Co-expression network of neural-differentiation genes shows specific pattern in schizophrenia. BMC Med Genomics. 2015;8:23.

36. Frantseva MV, Fitzgerald PB, Chen R, Möller B, Daigle M, Daskalakis ZJ. Evidence for impaired long-term potentiation in schizophrenia and its relationship to motor skill learning. Cereb Cortex N Y N 1991. 2008;18:990–6.

37. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.

Additional files

Additional file 1. *Additional methods*. Details for the leakage-free covariate adjustment. (AdditionalFile1.pdf).

Additional file 2. *resAdj TPOT pre-processor output.* Structure and column description of the output from this pre-processor (AdditionalFile2.pptx).

Additional file 3. *Boxplots for the results of 100 runs of classic TPOT on the TG-GATEs data set.* Each point corresponds to one run, where the *x*-coordinate indicates the pathway (Feature Set) selected in the optimal pipeline for that run and the *y*-coordinate indicates the balanced accuracy on the held-out Testing dataset. (AdditionalFile3.pdf).

Additional file 4. *Permutation importance from 100 runs of classic TPOT on the TG-GATEs data set.* The top 20 features are shown. The gene names are displayed on the *y*-axis and the weighted (by testing score) averages of the mean score decrease as a percentage of the score are displayed on the *x*-axis. The analyses were done at the probeset level, and for some of the genes there were two probesets among the top 20 features. (AdditionalFile4.png).

Additional file 5. *Boxplots for the results of 100 runs of resAdj TPOT on the PsychENCODE data set.* Each point corresponds to one run, where the *x*-coordinate indicates the pathway (Feature Set) selected in the optimal pipeline for that run and the *y*-coordinate indicates the R^2 on the held-out Testing dataset. (AdditionalFile5.pdf).

Additional file 6. *Permutation importance from 100 runs of resAdj TPOT on the PscychENCODE data set.* The top 20 genes (features) and 2 covariates are shown. The gene names are displayed on the *y*-axis and the weighted (by testing score) averages of the mean score decrease as a percentage of the score are displayed on the *x*-axis. (AdditionalFile6.png).

Additional file 7. *Boxplots for the results of 100 runs of classic TPOT on the PsychENCODE data set.* Each point corresponds to one run, where the *x*-coordinate indicates the pathway (Feature Set) selected in the optimal pipeline for that run and the *y*-coordinate indicates the balanced accuracy on the held-out Testing dataset. (AdditionalFile7.pdf).

Additional file 8. *Permutation importance from 100 runs of classic TPOT on the PscychENCODE data set.* The top 20 genes (features) are shown. The gene names are displayed on the *y*-axis and the weighted (by testing score) averages of the mean score decrease as a percentage of the score are displayed on the *x*-axis. (AdditionalFile8.png).

Additional file 9. *GSEA results on the PsychENCODE data set.* Pathways with unadjusted p-values<0.05. (AdditionalFile9.xlsx).

Additional file 10. *TG-GATEs assays used.* The list of 933 assays used in this work (AdditionalFile10.txt).