ECB 2021

sciforum

Proceedings

Machine learning for gene expression-based prediction of individual drug response for cancer patients †

Nicolas Borisov [1,*], Victor Tkachev [2,3], Maxim Sorokin [2,3], and Anton Buzdin [2,3,4]

[1] Moscow Institute of Physics and Technology, 141701 Moscow Oblast, Russia
[2] OmicsWayCorp, 91788 Walnut, CA, USA
[3] I.M. Sechenov First Moscow State Medical University, 119991 Moscow, Russia
[4] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, 117997 Moscow, Russia
* Correspondence: borisov@oncobox.com; Tel.: +7-903-218-7261
† Presented at the 1st International Electronic Conference on Biomedicine, 01–26 June 2021; Available online: https://ecb2021.sciforum.net/.

Published: 31 May 2021

Abstract: (1) Background: Various machine learning (ML) methods are applied for prediction of individual clinical efficiency of cancer drugs and therapeutic regimens. As features for ML, different multi-omics data may be used, such as genomic, transcriptomic, proteomic, and interactomic (activation levels of intracellular molecular pathways) profiles. (2) Methods: We proposed a next-generation ML approach termed FloWPS (FLOating-Window Projective Separator) that uses pre-processing/trimming/filtration of multi-omics features when building the ML models, in order to preclude extrapolation in the feature space. Additionally, FloWPS allows to neglect the influence of preceding cases from the training dataset, which are too distant in the feature space from the case that must be classified. Such extrapolation, as well as too distant instances, can cause model overtraining and results in decreased ML accuracy. (3) Results: Using Gene Expression Omnibus (GEO), The Cancer Genome Archive (TCGA), Tumor Alterations Relevant for GEnomics-driven Therapy (TARGET) project databases, as well as our own data, we selected 27 gene expression datasets for cancer patients, annotated with clinical response status. Each dataset had the same cancer type and treatment regimen. The biggest dataset included 235, and the smallest - only 41 patient cases. To form the robust set of marker features (gene expression levels), we applied the leave-one-out (LOO) cross-validation test that selected genes with the highest AUC values for good-vs-poor responder discrimination. Using the blind/agnostic LOO approach for data trimming, we demonstrated essential improvement of ML quality metrics (AUC, sensitivity and specificity) for FloWPS-based clinical response classifiers for all global ML methods applied, such as support vector machines (SVM), random forest (RF), binomial naïve Bayes (BNB), adaptive boosting (ADA), as well as multi-level perceptron (MLP). Namely, the AUC for the treatment response classifiers increased from 0.61–0.88 range to 0.70–0.97. (4) Conclusion: Considering our ML trial with 27 clinically annotated cancer gene expression datasets, the BNB method showed best performance for data trimming and was the most effective for classifying the clinical response using multi-omics features, with minimal, median and maximal AUC values equal to 0.77, 0.86 and 0.97, respectively.

Keywords: bioinformatics; personalized medicine; oncology; chemotherapy; machine learning; omics profiling.

1. Introduction

Machine learning (ML) methods can offer even a wide spectrum of opportunities by non-hypothesis-driven direct linkage of specific molecular features with clinical outcomes, such as responsiveness on certain types of treatment [1,2].

The high throughput transcriptomic data, including microarray- and next-generation sequencing gene expression profiles can be utilized for building such classifiers/predictors of clinical response to a certain type of treatment. However, the direct use of ML to personalize prediction of clinical outcomes is problematic, due to the lack of sufficient amounts of preceding clinically annotated cases supplemented with the high-throughput molecular data (~thousands or tens thousands of cases per treatment scheme) [3]. As a result, classical ML methods are often not successful in predicting clinical outcomes for several model datasets [4–8].

To improve the performance of ML in biomedicine, we recently developed an approach called flexible data trimming (FDT), which removes or excludes extreme values, or outliers, from a dataset [2,9–11]. Excluding non-informative features helps ML classifiers to avoid extrapolation, which is a well-known problem of ML [12–15]. Thus, for every point of a *validation* dataset, the *training* dataset is adjusted to form a floating window. We, therefore, called the respective ML approach, floating window projective separator (FloWPS) [2].

We investigated FloWPS performance for seven popular ML methods, including linear SVM, *k* nearest neighbors (kNN), random forest (RF), Tikhonov (ridge) regression (RR), binomial naïve Bayes (BNB), adaptive boosting (ADA) and multi-layer perceptron (MLP). We performed computational experiments for 27 high throughput gene expression datasets (41–235 samples per dataset) corresponding to 2192 cancer patients with known responses on chemotherapy treatments. We showed that FloWPS essentially improved the classifier quality for all *global* ML methods (SVM, RF, BNB, ADA, MLP), where the AUC for the treatment response classifiers increased from 0.65–0.85 range to 0.78–0.96. For all the datasets tested, the best performance of FloWPS data trimming was observed for the BNB method, which can be valuable for further building of ML classifiers in personalized oncology.

Additionally, to test the robustness of FloWPS-empowered ML methods against overtraining, we interrogated agreement/consensus features between the different ML methods tested, which were used for building mathematical models for the classifiers. The lack of such agreement/consensus could indicate overtraining of the ML classifiers built, suggesting random noise instead of extracting significant features distinguishing between the treatment responders and non-responders. If ML methods indeed tend to amplify random noise during overtraining, then one could expect a lack of correlation between the features for geometrically different ML models. However, we found here that (i) there were statistically significant positive correlations between different ML methods in terms of relative feature importance, and (ii) that this correlation was enhanced for the ML methods with FloWPS. We, therefore, conclude that the beneficial role of FloWPS is not due to overtraining.

2. Methods

2.1. Clinically Annotated Molecular Datasets

We used 27 publicly available datasets, including high throughput gene expression profiles associated with clinical outcomes of the respective patients [2,11,16]. The biosamples were obtained from tumor biopsies before chemotherapy treatments. The outcomes were response or lack of response on the therapy used, as defined in the original reports.

The datasets preparation for the analysis included the following steps [2,17]:

- Labelling each patient as either *responder* or *non-responder* on the therapy used;
- For each dataset, finding top marker genes having the highest AUC values for distinguishing responder and non-responder classes;

- Performing the leave-one-out (LOO) cross-validation procedure to complete the robust core marker gene set used for building the ML model.

4.2. Principles of Flexible Data Trimming

We used several non-deep ML methods implemented in the Python *sklearn* library. For each ML method we used a data trimming/preprocessing step using FloWPS method (R package flowpspkg.tar.gz) to increase robustness and efficiency due to individual sample specific selection of training dataset [2,10]. Among the ML methods, we applied linear support vector machines (SVM), the *k* nearest neighbors (kNN), random forest (RF), ridge regression (RR), binomial naïve Bayes (BNB), adaptive boosting (APA) and multi-layer perceptron (MLP). To improve performance of ML, we used a recent data preprocessing/trimming technique termed floating-window projective separator (FloWPS). This method increases AUC for most of ML methods in most of the clinically annotated gene expression datasets investigated [2,10,11]. FloWPS improves the classifier robustness by performing dynamic data trimming and selecting sample-specific sets of relevant genes to prevent extrapolation in the feature space (described in detail in [2,10]). It prevents extrapolation in the feature space by excluding the features that cause such extrapolation. Second, it selects only *k* nearest neighbors for the training dataset to build a ML model similarly to the kNN method to avoid confusing interference from too distant points from the training dataset in the feature space.

3. Results

3.1. Performance of FloWPS for Equalized Datasets Using All ML Methods with Default Settings

We used FloWPS in combination with seven ML methods, namely, linear support vector machines (SVM), *k* nearest neighbors (kNN), random forest (RF), ridge regression (RR), binomial naïve Bayes (BNB), adaptive boosting (ADA) and multi-layer perceptron (MLP).

The basic quality characteristics, namely ROC AUC, sensitivity (SN) and specificity (SP), of seven above ML methods for discrimination between responders and non-responders in our 27 cancer datasets are shown in Table 1. Although different values of false positive vs. false negative importance balance factor *B* did not affect the ROC AUC characteristics, they were crucial for sensitivity and specificity.

We found that the use of FloWPS has considerably improved the AUC metric for all global ML methods investigated (SVM, RF, BNB, ADA and MLP), but had no effect on the performance of local methods kNN and RR. For the global ML methods, FloWPS improved the classifier quality and increased AUC from 0.65–0.85 range to 0.78–0.96, and AUC median values—from 0.70–0.77 range to 0.76–0.82 (Table 1). In addition, kNN and RR also showed poor SN and SP for $B > 1$ and $B < 1$, respectively.

These findings are summarized in Table 1. Considering quality criterion of combining the highest AUC, the highest SN at $B = 4$ and the highest SP at $B = 0.25$, the top three methods identified for the default settings were BNB, MLP and RF.

3.2. Correlation Study Between Different ML Methods at the Level of Feature Importance

We showed positive pairwise correlations between the different ML methods at the level of relative importance (I_f, see [36]) of different features tested (Table 2). Greater similarities between I_f marks in the different ML methods reflect more robust applications of the ML. Importantly, the correlations for the ML methods with FloWPS were always higher than for the methods without FloWPS. This clearly suggests the beneficial role of FloWPS for extracting informative features from the noisy data. In this model, the biggest similarity was observed for the pair of RR and BNB methods.

4. Conclusion

Many ML methods which were designed for global separation of different classes of points in the feature space are prone to overtraining when the number of preceding cases is low. Global ML methods may also fail if there is only local rather than global order in the placement of different classes in the feature space [8,36].

Table 1. Performance metrics for seven ML methods with default settings for datasets with equal numbers of responders and non-responders.

ML Method	Method Type	Median AUC without FloWPS	Median AUC with FloWPS	Paired *t*-Test *p*-Value for AUC with-vs.-w/o FloWPS	Advantage of FloWPS	Median SN at $B = 4$	Median SP at $B = 0.25$
SVM	Global	0.74	0.80	1.3×10^{-5}	Yes	0.45	0.42
kNN	Local	0.76	0.75	0.53	No	0.25	0.34
RF	Global	0.74	0.82	1.3×10^{-5}	Yes	0.45	0.42
RR	Local	0.80	0.79	0.16	No	0.36	0.41
BNB	Global	0.77	0.82	2.7×10^{-4}	Yes	0.51	0.58
ADA	Global	0.70	0.76	2.4×10^{-4}	Yes	0.32	0.41
MLP	Global	0.73	0.82	6.4×10^{-5}	Yes	0.53	0.53

Yes–FloWPS is beneficial for ML quality, No–FloWPS is not beneficial for ML quality.

Table 2. Median pairwise Pearson/Spearman correlation at feature (gene expression) importance (I_f) level. Figures above main diagonal: With FloWPS; figures below: Without FloWPS.

	SVM	RF	RR	BNB	MLP
SVM	1	0.53/0.55	0.40/0.39	0.37/0.34	0.46/0.46
RF	0.34/0.40	1	0.51/0.32	0.48/0.31	0.59/0.38
RR	0.19/0.14	0.35/0.04	1	0.93/0.79	0.89/0.52
BNB	0.24/0.14	0.33/0.09	0.88/0.64	1	0.81/0.46
MLP	0.33/0.30	0.40/0.17	0.76/0.06	0.61/0.12	1

To improve performance of ML, FloWPS approach includes some elements of the local methods, e.g., using the flexible data trimming that avoids extrapolation in the feature space for each validation point and by selecting only several nearest neighbors from the training dataset. In such a way, the whole ML classifier becomes hybrid, both global and local [8,36].

In this hybrid approach, for each validation point training of ML models is performed in the individually tailored feature space. Every validation point is surrounded by a floating window from the points of the training dataset, and the irrelevant features are avoided using the rectangular projections in the feature space.

Overtraining, together with extrapolation, is very frequently considered also an Achilles heel of ML. We, therefore, tested if FloWPS helps to extract truly significant features or if it simply adapts to random noise, thus, causing overfitting. We compared four global ML methods (SVM, RF, BNB and MLP) and one local ML method (RR) to check similarities between them in terms of relative importance of distinct individual features. We confirmed that all these five ML methods were positively correlated at the level of feature importance (Table 2). Moreover, using FloWPS significantly enhanced such

correlations in all the cases examined (Table 2). These results clearly suggest that FloWPS is helpful for extracting relevant information rather than merely follows the random noise and overfits the ML model.

Overall, we propose that using correlations between different ML methods at the level of relative importance of distinct features may be used as an evaluation metric of ML suitability for building classifiers utilizing omics data. In this case, the higher is the correlation, the bigger should be the probability that the separation of responders from non-responders is robust and non-overtrained.

Author Contributions: Testing and debugging the computational code, V.T.; identification of relevant gene expression datasets, M.S.; design of the research and preparation of the paper, A.B. and N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Russian Scientific Foundation Grant 21-74-20066.

Acknowledgments: The cloud-based computational facilities were sponsored by Amazon and Microsoft Azure grants.

Conflicts of Interest: Authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Buzdin, A.; Sorokin, M.; Poddubskaya, E.; Borisov, N. High-Throughput Mutation Data Now Complement Transcriptomic Profiling: Advances in Molecular Pathway Activation Analysis Approach in Cancer Biology. *Cancer Inform* **2019**, *18*, 117693511983884, doi:10.1177/1176935119838844.
2. Tkachev, V.; Sorokin, M.; Mescheryakov, A.; Simonov, A.; Garazha, A.; Buzdin, A.; Muchnik, I.; Borisov, N. FLOating-Window Projective Separator (FloWPS): A Data Trimming Tool for Support Vector Machines (SVM) to Improve Robustness of the Classifier. *Frontiers in Genetics* **2019**, *9*, 717, doi:10.3389/fgene.2018.00717.
3. Azarkhalili, B.; Saberi, A.; Chitsaz, H.; Sharifi-Zarchi, A. DeePathology: Deep Multi-Task Learning for Inferring Molecular Pathology from Cancer Transcriptome. *Sci Rep* **2019**, *9*, 16526, doi:10.1038/s41598-019-52937-5.
4. Turki, T.; Wang, J.T.L. Clinical Intelligence: New Machine Learning Techniques for Predicting Clinical Drug Response. *Computers in Biology and Medicine* **2019**, *107*, 302–322, doi:10.1016/j.compbiomed.2018.12.017.
5. Turki, T.; Wei, Z. A Link Prediction Approach to Cancer Drug Sensitivity Prediction. *BMC Systems Biology* **2017**, *11*, doi:10.1186/s12918-017-0463-8.
6. Turki, T.; Wei, Z. Learning Approaches to Improve Prediction of Drug Sensitivity in Breast Cancer Patients. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); IEEE, August 2016; pp. 3314–3320.
7. Turki, T.; Wei, Z.; Wang, J.T.L. A Transfer Learning Approach via Procrustes Analysis and Mean Shift for Cancer Drug Sensitivity Prediction. *J. Bioinform. Comput. Biol.* **2018**, *16*, 1840014, doi:10.1142/S0219720018400140.
8. Turki, T.; Wei, Z.; Wang, J.T.L. Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients. *IEEE Access* **2017**, *5*, 7381–7393, doi:10.1109/ACCESS.2017.2696523.
9. Borisov, N.; Tkachev, V.; Suntsova, M.; Kovalchuk, O.; Zhavoronkov, A.; Muchnik, I.; Buzdin, A. A Method of Gene Expression Data Transfer from Cell Lines to Cancer Patients for Machine-Learning Prediction of Drug Efficiency. *Cell Cycle* **2018**, *17*, 486–491, doi:10.1080/15384101.2017.1417706.
10. Tkachev, V.; Sorokin, M.; Borisov, C.; Garazha, A.; Buzdin, A.; Borisov, N. Flexible Data Trimming Improves Performance of Global Machine Learning Methods in Omics-Based

Personalized Oncology. *International Journal of Molecular Sciences* **2020**, *21*, 713, doi:10.3390/ijms21030713.

11. Borisov, N.; Sergeeva, A.; Suntsova, M.; Raevskiy, M.; Gaifullin, N.; Mendeleeva, L.; Gudkov, A.; Nareiko, M.; Garazha, A.; Tkachev, V.; et al. Machine Learning Applicability for Classification of PAD/VCD Chemotherapy Response Using 53 Multiple Myeloma RNA Sequencing Profiles. *Front. Oncol.* **2021**, *11*, 652063, doi:10.3389/fonc.2021.652063.
12. Arimoto, R.; Prasad, M.-A.; Gifford, E.M. Development of CYP3A4 Inhibition Models: Comparisons of Machine-Learning Techniques and Molecular Descriptors. *Journal of biomolecular screening* **2005**, *10*, 197–205.
13. Balabin, R.M.; Lomakina, E.I. Support Vector Machine Regression (LS-SVM): An Alternative to Artificial Neural Networks (ANNs) for the Analysis of Quantum Chemistry Data? *Physical Chemistry Chemical Physics* **2011**, *13*, 11710–11718.
14. Balabin, R.M.; Smirnov, S.V. Interpolation and Extrapolation Problems of Multivariate Regression in Analytical Chemistry: Benchmarking the Robustness on near-Infrared (NIR) Spectroscopy Data. *Analyst* **2012**, *137*, 1604–1610.
15. Betrie, G.D.; Tesfamariam, S.; Morin, K.A.; Sadiq, R. Predicting Copper Concentrations in Acid Mine Drainage: A Comparative Analysis of Five Machine Learning Techniques. *Environmental monitoring and assessment* **2013**, *185*, 4171–4182.
16. Borisov, N.; Sorokin, M.; Tkachev, V.; Garazha, A.; Buzdin, A. Cancer Gene Expression Profiles Associated with Clinical Outcomes to Chemotherapy Treatments. *BMC medical genomics* **2020**, *13*, 111, doi:10.1186/s12920-020-00759-0.
17. Borisov, N.; Buzdin, A. New Paradigm of Machine Learning (ML) in Personalized Oncology: Data Trimming for Squeezing More Biomarkers From Clinical Datasets. *Frontiers in Oncology* **2019**, *9*, 658, doi:10.3389/fonc.2019.00658.