Bioinformatics, 33(9), 2017, 1407–1410
doi: 10.1093/bioinformatics/btw765
Advance Access Publication Date: 6 February 2017
Applications Note

Gene expression

IntegratedMRF: random forest-based framework for integrating prediction from different data types

Raziur Rahman[1], John Otridge[2] and Ranadip Pal[1,*]

[1]Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX 79409, USA and [2]Leidos Biomedical Research Inc, Frederick, MD 21701, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on April 27, 2016; revised on November 18, 2016; editorial decision on November 26, 2016; accepted on December 1, 2016

Abstract

Summary: IntegratedMRF is an open-source R implementation for integrating drug response predictions from various genomic characterizations using univariate or multivariate random forests that includes various options for error estimation techniques. The integrated framework was developed following superior performance of random forest based methods in NCI-DREAM drug sensitivity prediction challenge. The computational framework can be applied to estimate mean and confidence interval of drug response prediction errors based on ensemble approaches with various combinations of genetic and epigenetic characterizations as inputs. The multivariate random forest implementation included in the package incorporates the correlations between output responses in the modeling and has been shown to perform better than existing approaches when the drug responses are correlated. Detailed analysis of the provided features is included in the Supplementary Material.

Availability and Implementation: The framework has been implemented as a **R** package *IntegratedMRF*, which can be downloaded from https://cran.r-project.org/web/packages/IntegratedMRF/index.html, where further explanation of the package is available.

Contact: ranadip.pal@ttu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A primary objective of precision medicine for cancer is the selection of an anti-cancer drug or drug combination that is most effective for the individual patient (Garnett and *et al*., 2012). A diverse set of regression models such as linear regression with regularization, nonlinear regression, kernel based techniques and ensemble based approaches have been applied to the problem of drug response prediction from genetic characterizations for enabling personalized cancer therapy. The NCI-DREAM drug sensitivity prediction challenge was a crowd sourced initiative to apply a diverse set of prediction models on the same set of heterogeneous genetic characterization training data to evaluate model performance on holdout datasets (Costello and *et al*., 2014). The current computational framework development was motivated by the second best performance of random forest based approach in the NCI-DREAM challenge among more than 40 different submissions (Wan and Pal, 2014). Individual predictions from each genetic or epigenetic characterization dataset such as RNAseq, Protein expression or methylation based on Random Forests were combined using a linear regression model to arrive at combined predictions. We have considered further enhancements to the framework by (i) considering Multivariate Random Forests (Segal and Xiao, 2011) that improves prediction accuracy by incorporating the correlation between output responses, (ii) analyzing different error estimation techniques as the estimated error can be significantly different from the true or validation error for small sample scenarios. This application note describes the

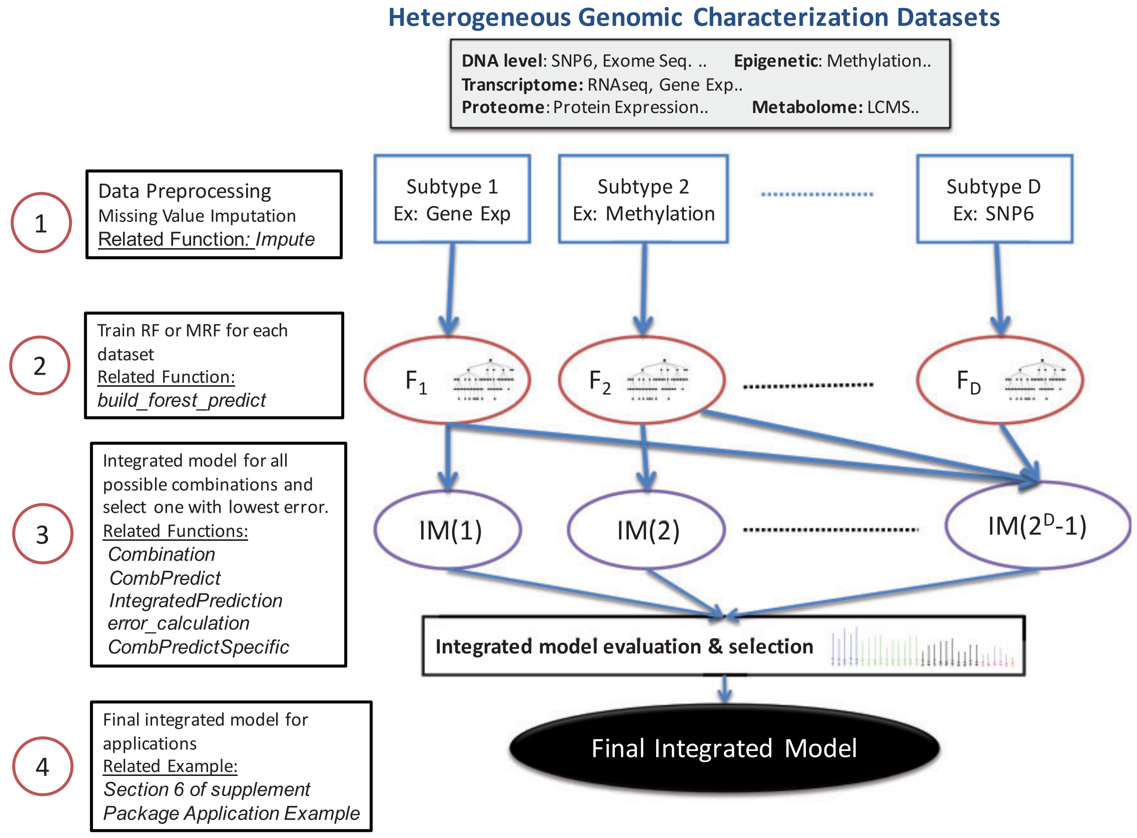
Heterogeneous Genomic Characterization Datasets
DNA level: SNP6, Exome Seq. .. Epigenetic: Methylation..
Transcriptome: RNAseq, Gene Exp..
Proteome: Protein Expression.. Metabolome: LCMS..
1
Data Preprocessing
Missing Value Imputation
Related Function: Impute
2
Train RF or MRF for each dataset
Related Function:
build_forest_predict
3
Integrated model for all possible combinations and select one with lowest error.
Related Functions:
Combination
CombPredict
IntegratedPrediction
error_calculation
CombPredictSpecific
4
Final integrated model for applications
Related Example:
Section 6 of supplement
Package Application Example
Subtype 1
Ex: Gene Exp
Subtype 2
Ex: Methylation
Subtype D
Ex: SNP6
F1
F2
FD
IM(1)
IM(2)
IM(2D-1)
Integrated model evaluation & selection
Final Integrated Model

Fig. 1. Overview of Integrated Prediction Methodology

IntegratedMRF package that has been implemented in *R* and C++. The package provides various functionalities to implement an integrated prediction from different data subtypes using either univariate or multivariate Random Forests along with options for different error estimation techniques to select the weights for generating the integrated model from individual predictions.

2 Methods and results

A conceptual overview of the computational framework is illustrated in Figure 1. For each individual genomic characterization (such as Gene Expression, Methylation, Exome Sequence, RPPA or SNP6 (CNV) denoted by Subtypes 1 to D in Fig. 1), a Random Forest (RF) or Multivariate Random Forest (MRF) model is estimated using the training samples. For generating integrated model predictions for all possible data subtype combinations, a specific error estimation method (such as Leave-one-out, N-fold Cross Validation, Bootstrap or 0.632+ Bootstrap) is applied to generate weights for individual predictions using least square regression. The integrated model that provides the minimum mean absolute prediction error is selected for predicting the testing samples.

Algorithms: The *IntegratedMRF* package provides various options for design of integrated ensemble predictive models, each with the following significant features:

Model Inference: The package provides options for estimation of regular RF or MRF as the predictive modeling approach. If the number of output responses is greater than 1, the default selection is MRF whereas for a single output response, a RF model is generated. The suitability of MRF for prediction as compared to RF is discussed in next section. *We have observed that MRF provides higher accuracy than RF when the output responses are correlated as observed through common targets in the drugs.*

Error Estimation techniques: We provide options for Leave-one-out, n-fold cross validation, Bootstrap and 0.632+ Bootstrap (Bradley Efron, 1997) error estimation approaches to estimate the true error of a designed model. Bootstrap error estimates can sometimes be upward biased (estimate higher than true error) and 0.632+ Bootstrap attempts to correct that bias by linearly combining the high biased Bootstrap estimate with low biased training error estimate. The detailed descriptions of the error estimation techniques are included in the Supplementary Section 2. *Among the considered error estimation approaches, leave-one-out appears closer to validation error for smaller samples based on DREAM challenge results described in next section. For larger sample sizes, it is preferable to use n-fold CV or Bootstrap (details included in* the Supplementary Section 5.3). We recommend the use of Leave-one-out error estimation for small sample scenarios (such as DREAM challenge dataset with 35 samples; Costello and *et al.*, 2014) and the application of n-fold CV or Bootstrap for larger datasets (such as 400 sample dataset of CCLE; Barretina and *et al.*, 2012).

Integrated Model: *IntegratedMRF* provides an option for applying linear least square regression on predictions from models trained on different data types to estimate the weights for combining the model predictions. Based on the selected error estimation approach, the training and testing datasets are created. The training samples are used to estimate the individual model parameters along with the integrated model regression coefficients. The testing samples are used to evaluate the performance of each integrated model in terms of mean absolute prediction error. Note that for *D* different genetic

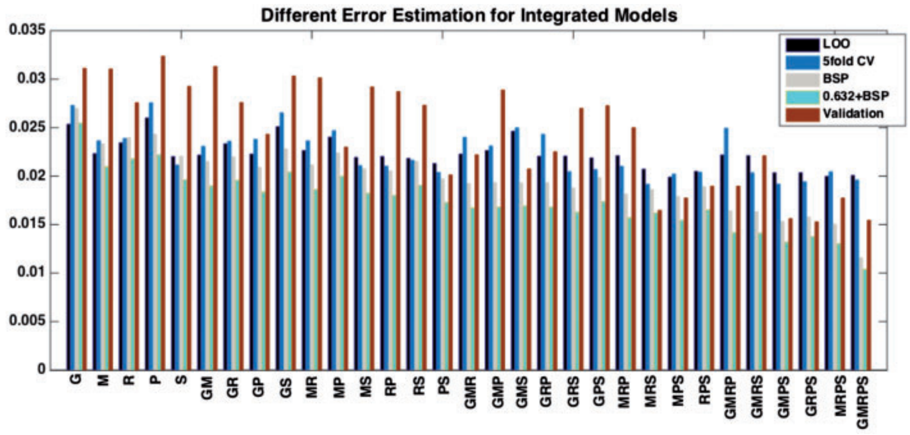

Fig. 2. Multiple Mean Absolute Error estimates (LOO: Leave one out, 5foldCV: 5 fold cross validation, BSP: Bootstrap, 0.632 + BSP: 0.632 + Bootstrap) and mean absolute validation error for drug 3 from NCI Dream Challenge Dataset for 31 ($2^5 - 1$) different data subtype Integrated Models. The datasets are denoted by: G: Gene Expression, M: Methylation, R: RNASeq, P: RPPA and S: SNP6

Table 1. Fivefold Cross validation results for GDSC dataset response prediction for four drug sets (S_1, S_2 and S_3 sets consists of highly correlated responses whereas the set of S_4 has low correlated responses)

			Correlation coefficients			
Drug set	Correlation among responses	Drug name	*EN*	*KBMTL*	*RF*	*MRF*
S_1	0.8439	RDEA119	0.62	0.57	0.63	0.66
		PD-0325901	0.48	0.47	0.61	0.63
S_2	0.8410	BI-2536	0.23	0.23	0.26	0.28
		GW843682X	0.30	0.28	0.31	0.33
S_3	0.8366	CI-1040	0.46	0.51	0.59	0.60
		PD-0325901	0.50	0.52	0.62	0.65
S_4	6.59e-7	Axitinib	0.31	0.33	0.36	0.32
		Mitomycin C	0.28	0.25	0.37	0.38

Correlation coefficients between actual and predicted drug responses using Elastic Net (*EN*), Kernelized Bayesian multitask learning (*KBMTL*), Random Forest (*RF*) and Multivariate Random Forest (*MRF*) are reported. The parameters for EN and KBMTL were same as the parameters used by earlier drug sensitivity prediction studies of Barretina and *et al.* (2012) and Gonen and Margolin (2014) respectively.

or epigenetic characterizations, there will be $2^D - 1$ integrated model combinations. The alternative possibilities of concatenating all features before generating a single RF or MRF model or applying feature selection before model generation is compared to our current approach in the Supplementary Section 5.2. Results indicate that *it is better to generate individual predictive models for each dataset and then combine them to form an integrated model rather than designing a single model from a concatenated dataset*.

2.1 Performance

Error estimation and integrated model performance: The computational framework presented here has been utilized to analyze drug response prediction from NCI-DREAM drug sensitivity prediction challenge (Costello and *et al.*, 2014) and Cancer Cell Line Encyclopedia (CCLE) (Barretina and *et al.*, 2012) datasets. The model training and error estimates are based on 35 samples of the DREAM challenge dataset whereas the validation or true error of an inferred model is estimated based on a separate set of 18 hold-out samples. Figure 2 displays the different error estimates (Leave One Out, 5-fold CV, Bootstrap, 0.632 + Bootstrap) and validation error in terms of Mean Absolute Error for drug 3 from NCI Dream Challenge Dataset for 31 ($2^5 - 1$) different Integrated Models. Figure 2 shows that different error estimates can have variations but the overall trend of the error estimates along with validation error shows a reduction in prediction error with addition of more datasets. We also observe that LOO followed by 5fold CV and BSP provides the closest estimate of the true error. Similar behavior is observed for another scenario (drug 1) as reported in the Supplementary Section 5.1.

We have also compared the various error estimation techniques to validation error for CCLE dataset where 100 random samples out of around 500 samples were used for model generation and error estimations. The remaining hold-out set of around 400 samples were used to estimate the true error or validation error. The results reported in the Supplementary Section 5.3 support the earlier hypothesis of LOO followed by 5-fold CV and BSP provides the closest estimate of the true error. Comparing Figure 2 and Supplementary Figure S6, we observe that the difference between validation error and the error estimates reduces when the initial set of training samples is increased from 35 to 100.

Comparison of MRF with other approaches: In this section, we compare the prediction performance of Multivariate Random Forest (MRF) to Random Forest (RF) (Wan and Pal, 2014), Elastic Net (EN) (Barretina and *et al.*, 2012) and Kernelized Bayesian Multi Task Learning (KBMTL) (Gonen and Margolin, 2014) approaches using *Genomics of Drug Sensitivity in Cancer* (GDSC) gene expression dataset (Yang, 2013). Table 1 shows the 5-fold cross validation performance in terms of correlation coefficient between predicted and actual responses on GDSC dataset for MRF, RF, EN and KBMTL approaches for four sets of drugs (S_1, S_2 and S_3 are the pairs with the highest correlations among training responses while S_4 is the pair with the lowest correlation). Table 1 shows that MRF outperforms RF, EN and KBMTL for S1, S2 & S3 drug sets whereas RF outperforms MRF for S4 drug set.

We have also compared MRF to RF approach using other drug pairs of GDSC, DREAM challenge and CCLE datasets that are reported in the Supplementary Tables S2 and S4–S7. Similar to GDSC scenario, we observed MRF outperforms RF when the drug responses are correlated (Supplementary Tables S2, S5 and S6) and RF outperforms MRF when drug responses are not correlated (Supplementary Tables S4 and S7). Note that, we expect that drug responses will be correlated when they share common primary targets and the use of MRF will likely be advantageous in such scenarios.

3 Conclusions

The presented computational framework provides the following enhanced features of: (i) generation of multivariate random forests that incorporates dependencies in output responses which has been shown to outperform existing approaches for correlated drug responses, (ii) application of multiple error estimation approaches and (iii) integration of predictions from different genetic characterizations that results in a decrease in average error estimates and validation error with additional datasets.

Acknowledgements

We would like to thank Gustavo Stolovitzky, Daniel Gallahan and Joe Gray for organizing the NCI-DREAM drug sensitivity challenge that motivated the generation of the presented computational framework.

Funding

This work has been supported by a contract from National Cancer Institute/ Leidos Biomedical Research Inc 15X073.

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Bradley Efron,R.T. (1997) Improvements on cross-validation: the.632+ bootstrap method. *J. Am. Stat. Assoc.*, **92**, 548–560.

Costello,J.C. *et al.* (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, doi:10.1038/nbt.2877.

Garnett,M. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Gonen,M. and Margolin,A.A. (2014) Drug susceptibility prediction against a panel of drugs using kernelized bayesian multitask learning. *Bioinformatics*, **30**, i556– i563.

Segal,M. and Xiao,Y. (2011) Multivariate random forests. *Wiley Interdisc. Rev. Data Min. Knowl. Discov.*, **1**, 80–87.

Wan,Q. and Pal,R. (2014) An ensemble based top performing approach for NCI-dream drug sensitivity prediction challenge. *PLOS One*, **9**, e101183.

Yang,W.E. (2013) Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.