"Predicting a Biological Response of Molecules from Their Chemical Properties Using Diverse and Optimized Ensembles of Stochastic Gradient Boosting Machine"

By
Tarek Abdunabi and Otman Basir
Electrical and Computer Engineering Department
University of Waterloo
Canada

Outline

- Introduction.
- Stochastic gradient boosting.
- Data and computational resources.
- Building and tuning the ensembles.
- Performance evaluation.
- Fusing ensembles' decisions.
- Conclusion and future work.
- Supplementary material (R code).

Introduction

- The development of a new drug largely depends on trial and error.
- It typically involves synthesizing thousands of compounds that finally becomes a drug.
- As a result, this process is extremely expensive and slow.
- Therefore, the ability to accurately predict the biological activity of molecules, and understand the rationale behind those predictions are of great value.

Stochastic gradient boosting

- Gradient Boosting Machines (GBMs) are powerful ensemble learning techniques.
- In boosting methods, new models are added sequentially to the ensemble.
- At each iteration, a new base-leaner model is trained with respect to the error of the ensemble learned in the previous iterations.
- Despite their high accuracy, GBMs suffer from major drawbacks such as high memory consumption.

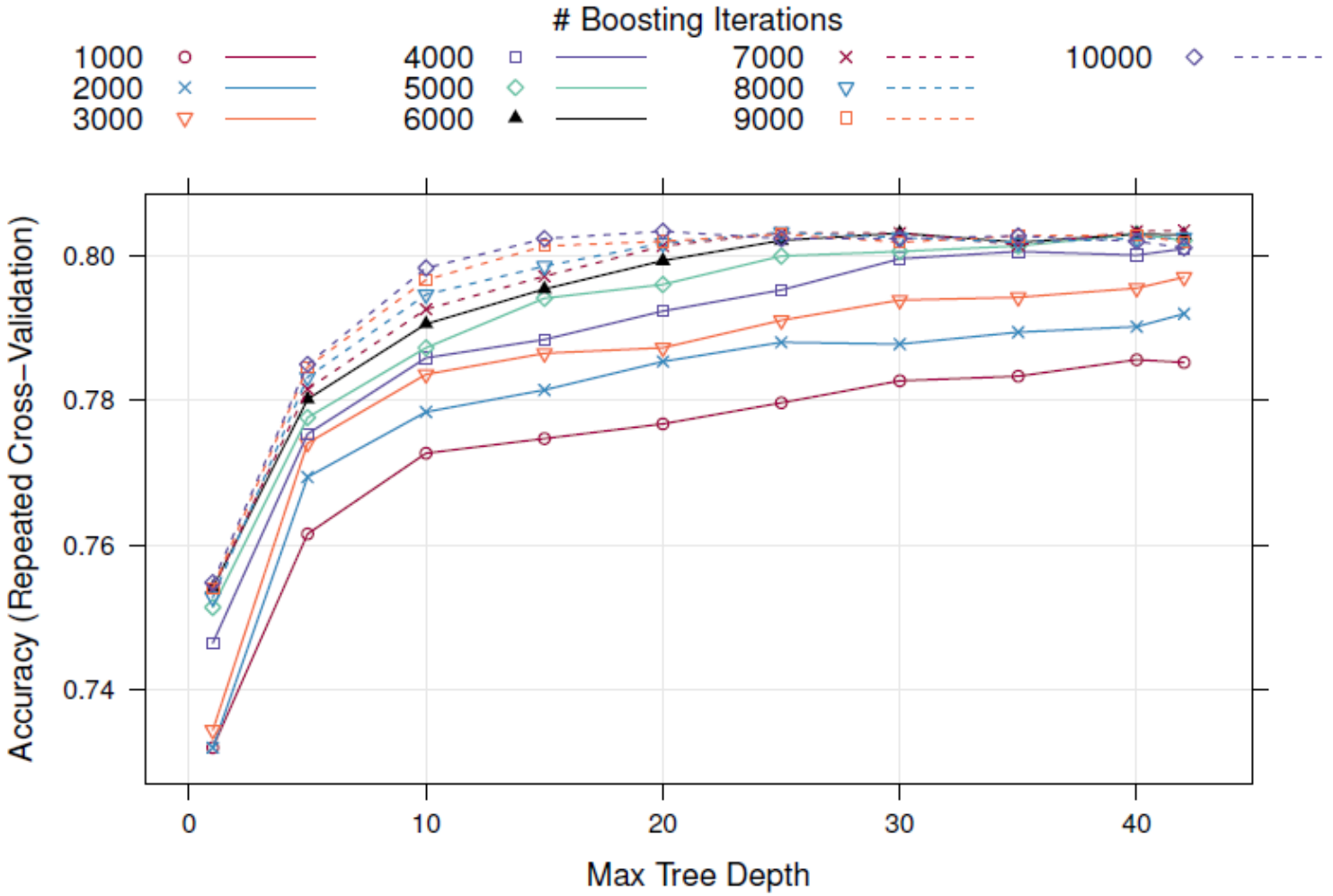
Data and computational resources

- The data was obtained from the Kaggle.com's competition: "Predicting a Biological Response" held between March16, 2012 and June 15, 2012.
- The objective of the competition was to build a predictive model to optimally relate molecular information to an actual biological response.
- The first column contains experimental data describing an actual biological response (Active/Inactive).
- The remaining columns represent molecular descriptors (D1 through D1776) e. g. Size, shape, etc.

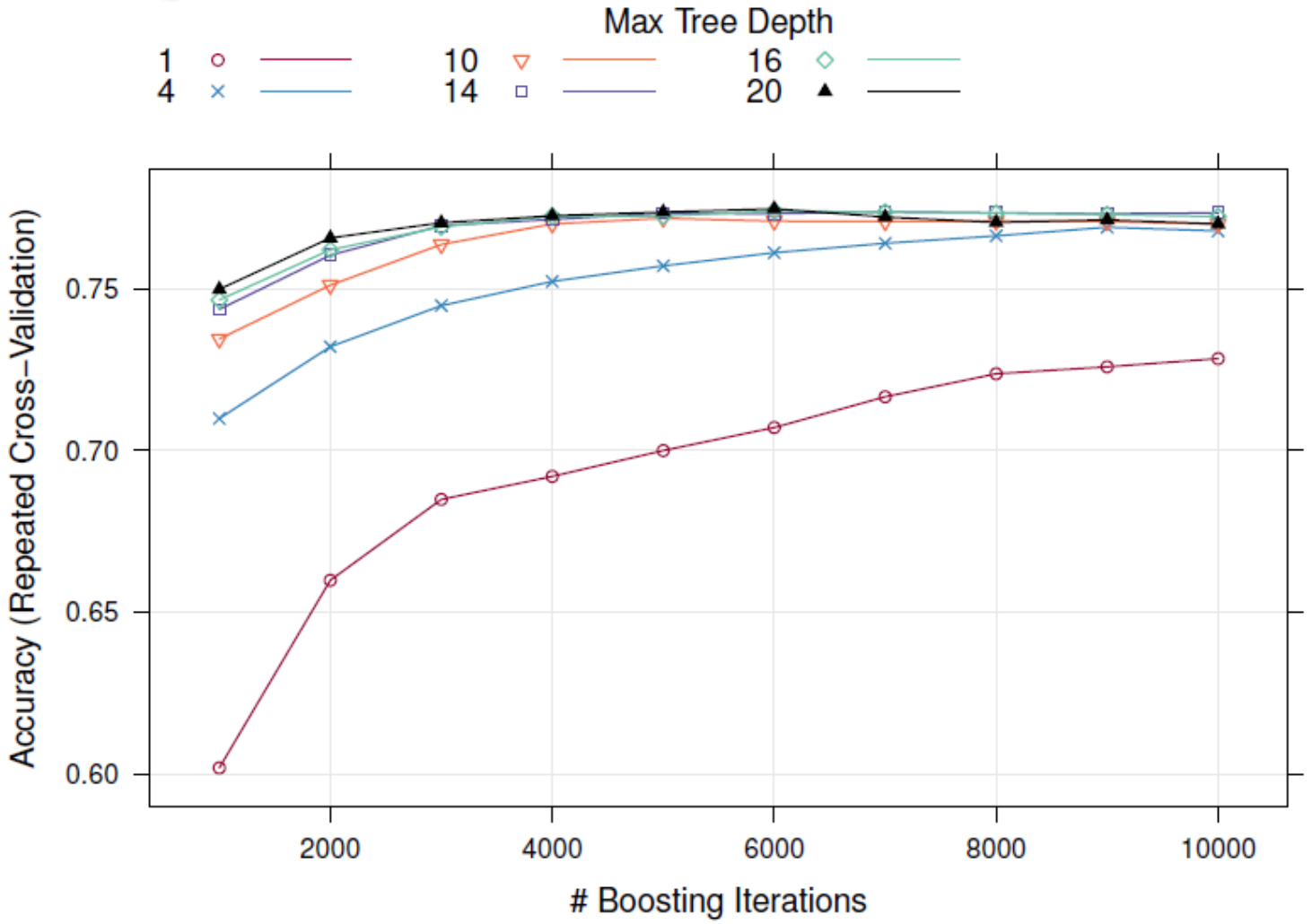
Building and tuning the ensembles

- We build optimized GBMs ensembles using several feature selection/reduction techniques.
- The performance of each ensemble is compared to the performance of the optimized ensemble built using all predictors (1776).
- Then, we use two fusion techniques to obtain better prediction accuracy.

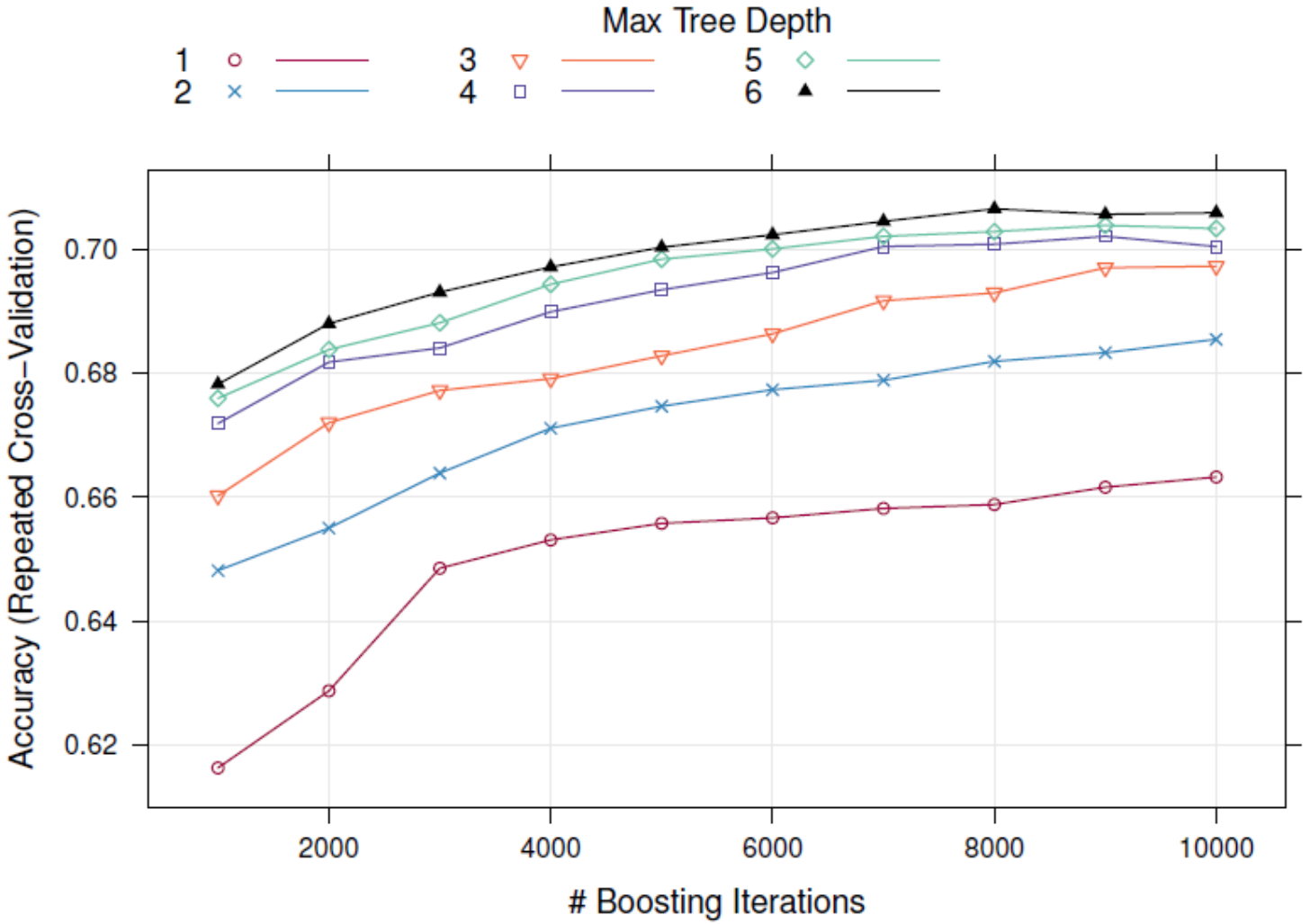
Building and tuning the ensemble using all predictors



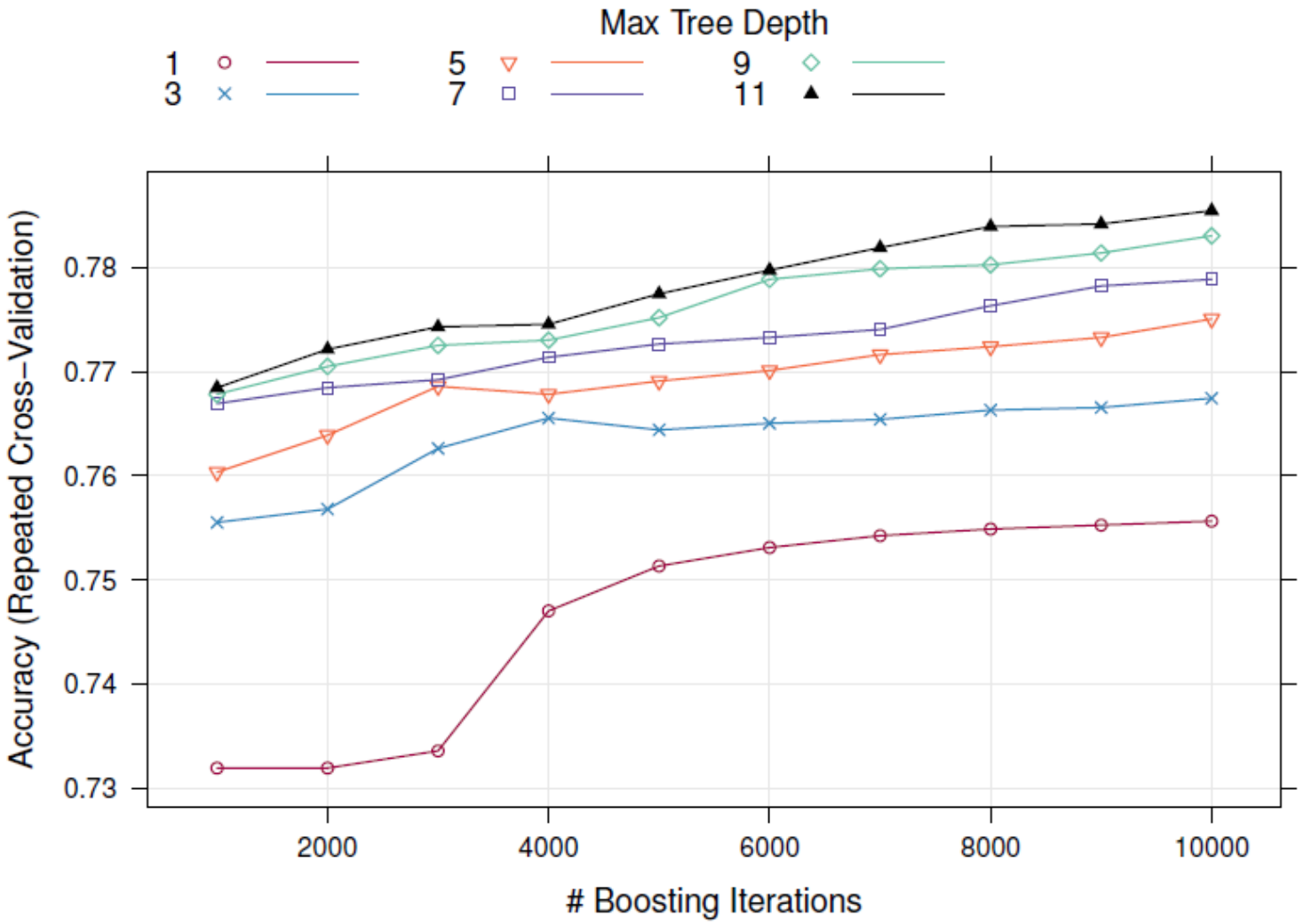
Building and tuning the ensemble using PCA (first 255 with 95% var.)



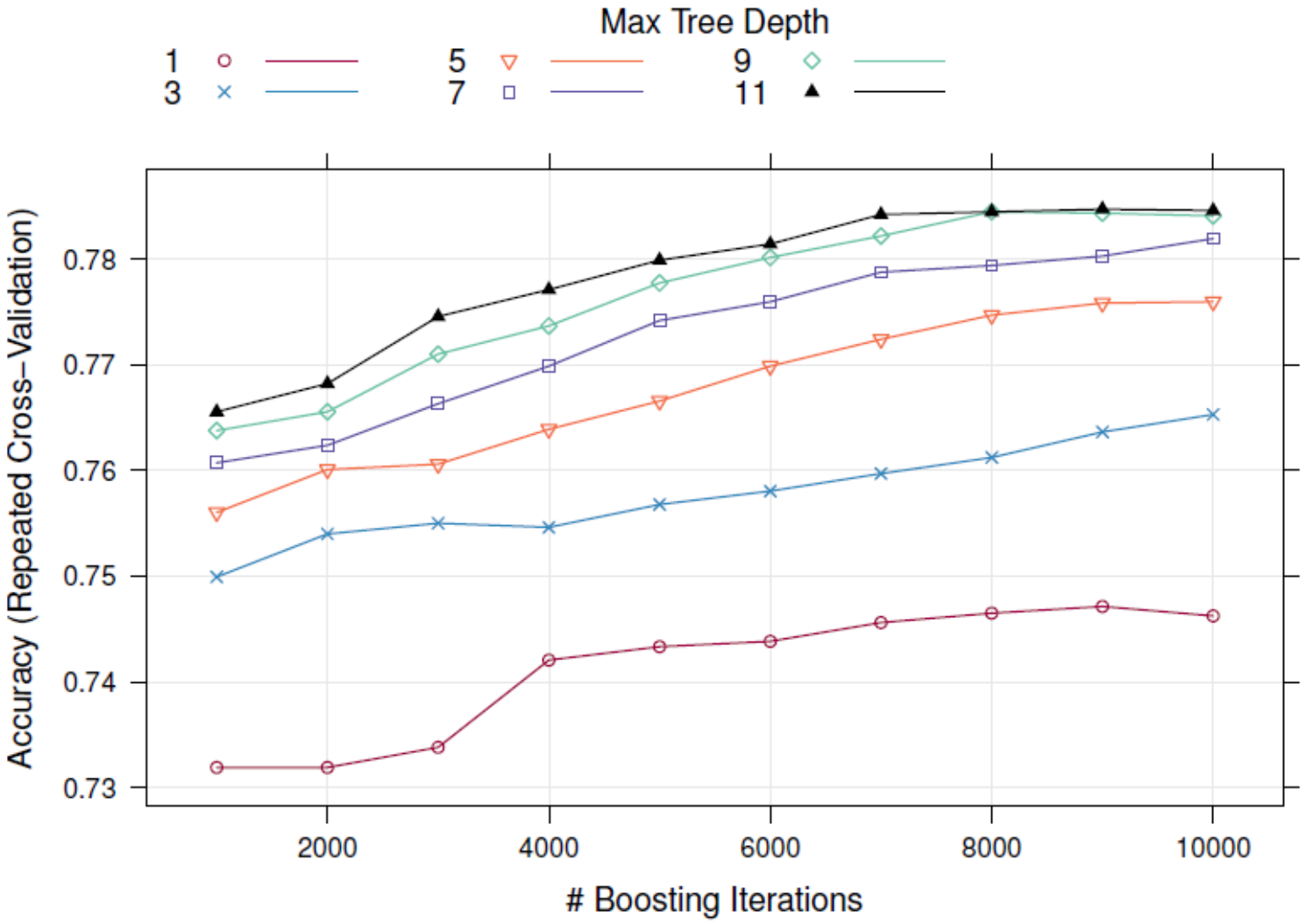
Building and tuning the ensemble using PCA (Kaiser-Guttman rule)



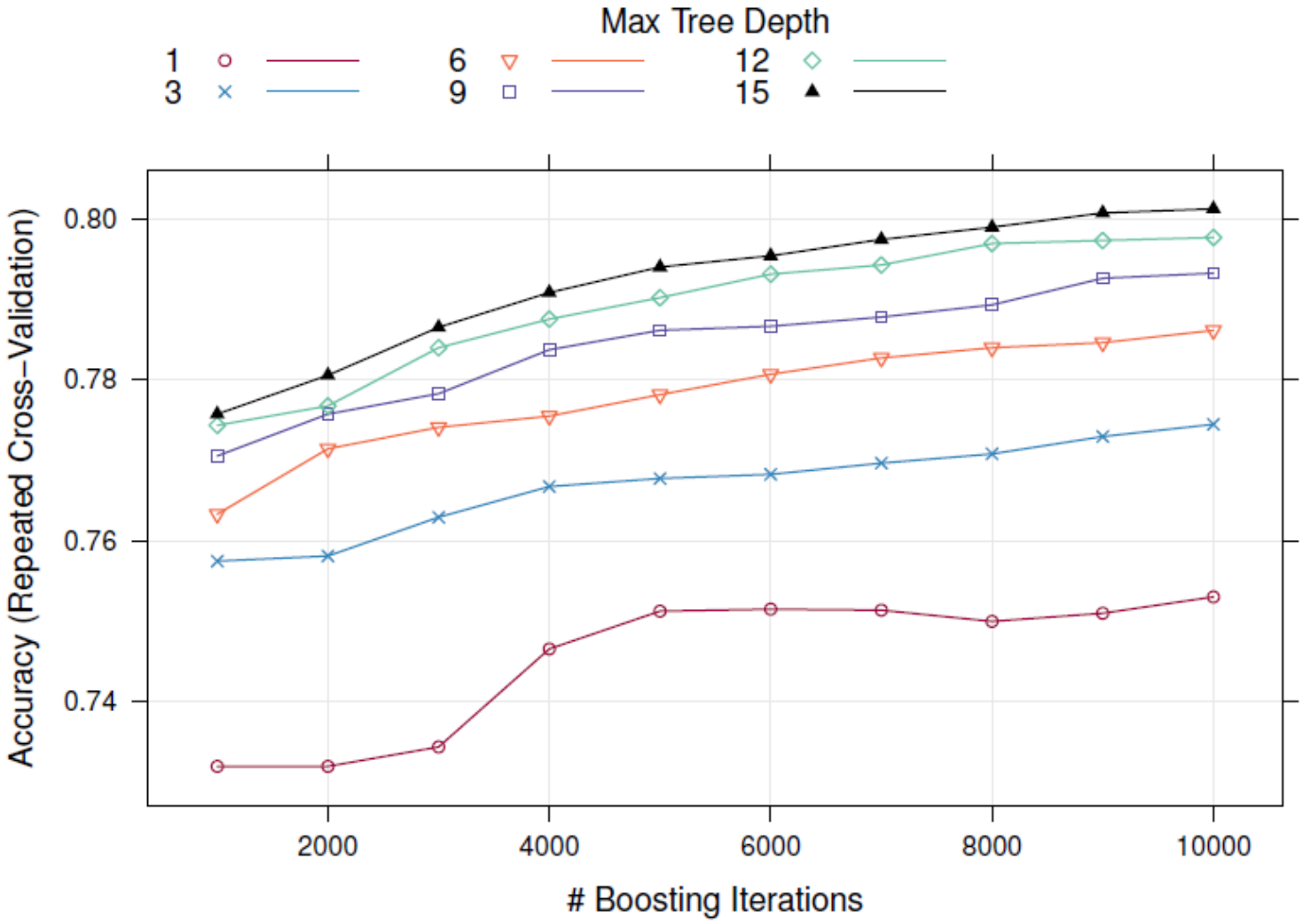
Building and tuning the ensemble using predictors' area under ROC



Building and tuning the ensemble using the Relief algorithm



Building and tuning the ensemble using GBM's predictor importance



Performance evaluation

In-sample performance

Metric	Ensemble 1 (all predictors)	Ensemble 2 (PCA 95%)	Ensemble 3 (PCA Kaiser)	Ensemble 4 (ROC area)	Ensemble 5 (Relief score)	Ensemble 6 (GBM importance)
Accuracy	99.62%	98.13%	79.93%	91.32%	90.40%	96.53%
95% Confidence interval	(99.30, 99.82)%	(97.54, 98.62)%	(78.35, 81.45)%	(90.17, 92.37)%	(89.21, 91.50)%	(95.76, 97.20)%
Sensitivity	99.72%	98.81%	82.94%	93.19%	92.49%	97.33%
Specificity	99.50%	97.34%	76.37%	89.10%	87.94%	95.59%
Kappa	99.23%	96.24%	59.47%	82.47%	80.62%	93.01%
No Information Rate (NIR)	54.23%	54.23%	54.23%	54.23%	54.23%	54.23%
P-Value [Accuracy > NIR]	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$
Positive predictive value	99.58%	97.78%	80.61%	91.02%	90.08%	96.32%
Negative predictive value	99.67%	98.57%	79.07%	91.70%	90.81%	96.80%
Prevalence	54.23%	54.23%	54.23%	54.23%	54.23%	54.23%
Detection rate	54.07%	53.58%	44.97%	50.53%	50.15%	52.78%
Detection prevalence	54.30%	54.80%	55.79%	55.52%	55.67%	54.80%
Balanced accuracy	99.61%	98.07%	79.65%	91.14%	90.21%	96.46%

Performance evaluation

Out-of-sample performance

Metric	Ensemble 1 (all predictors)	Ensemble 2 (PCA 95%)	Ensemble 3 (PCA Kaiser)	Ensemble 4 (ROC area)	Ensemble 5 (Relief score)	Ensemble 6 (GBM importance)
Accuracy	79.82%	78.40%	71.56%	79.11%	78.84%	78.67%
95% Confidence interval	(77.36, 82.23)%	(75.88, 80.77)%	(68.82, 74.18)%	(76.62, 81.45)%	(76.34, 81.20)%	(76.15, 81.03)%
Sensitivity	82.62%	82.46%	76.56%	81.80%	82.13%	81.64%
Specificity	76.50%	73.59%	65.63%	75.92%	74.95%	75.15%
Kappa	59.26%	56.30%	42.41%	57.84%	57.26%	56.92%
No Information Rate (NIR)	54.22%	54.22%	54.22%	54.22%	54.22%	54.22%
P-Value [Accuracy > NIR]	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$
Positive predicted value	80.64%	78.72%	72.52%	80.10%	79.52%	79.55%
Negative predicted value	78.80%	77.98%	70.27%	77.89%	74.95	77.56%
Prevalence	54.22%	54.22%	54.22%	54.22%	54.22%	54.22%
Detection rate	44.80%	44.71%	41.51%	44.36%	44.53%	44.27%
Detection prevalence	55.56%	56.80%	57.24%	55.38%	56.00%	55.64%
Balanced accuracy	79.56%	78.03%	71.09%	78.86%	78.54%	78.39%

Performance evaluation

Computations time

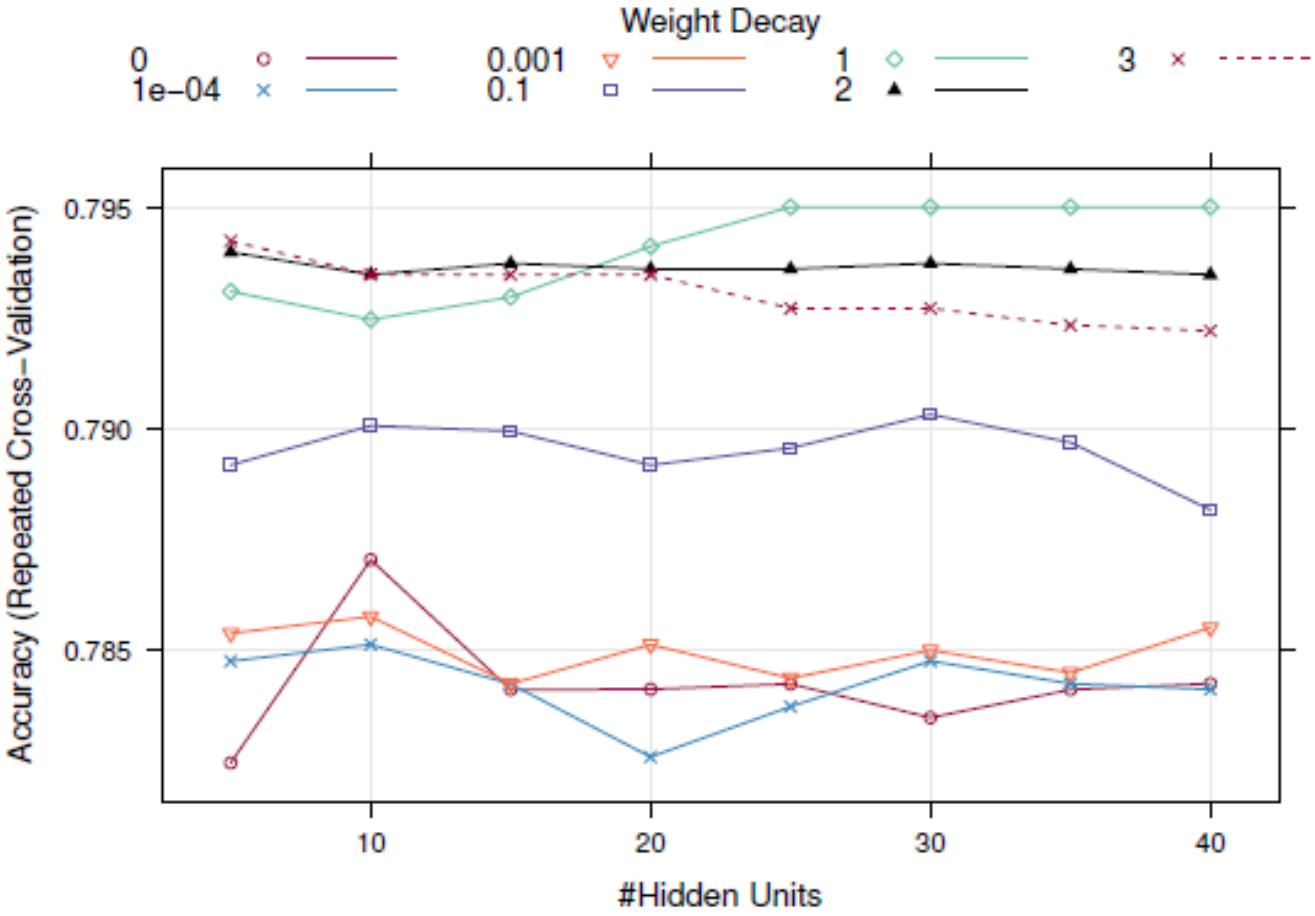
- The time required to build and tune our ensembles using Amazon cloud computing (32 cores) was:
- Ensemble 1 (all predictors) ~ 43.2 hours
- Ensemble 2 (PCA 255 comp.) ~ 2.9 hours
- Ensemble 3 (PCA Kaiser) ~ 3.58 minutes
- Ensemble 4 (ROC area) ~ 31.32 minutes
- Ensemble 5 (Relief score) ~ 32.06 minutes
- Ensemble 6 (GBM var. Impor.) ~ 1.1 hours

Fusing the ensembles' decisions

- Two techniques were used, Majority Vote and an optimized feedforward neural network.
- For unbiased results, we used stratified random sampling to split the test dataset to training and testing datasets.
- The prediction accuracy of ensembles 1 to 6 are 79.53%, 78.04%, 72.7%, 79.28%, 79.23%, and 79.53% respectively.
- The prediction accuracy using the majority vote is 79.53% (better than ensembles 2 to 5, and equals to ensembles 1 and 6)

Fusing the ensembles' decisions

- Predictions of ensembles 2-6 on the training dataset are used to build and tune a feedforward neural network.
- Prediction accuracy on the test data is 80.12%



Conclusion and future work

- The ability to accurately predict the biological activity of molecules, and understand the rationale behind those predictions are of great value in drug discovery.

- Using different feature selection/reduction techniques, diverse and optimized Tree-based Gradient Boosting Machines were built for a real, high-dimensional dataset.

Conclusion and future work

- By using these techniques, the computations time for building an optimized GBM using all predictors is significantly reduced at a slight drop in prediction accuracy.

- A better prediction accuracy is obtained by fusing ensembles decisions using a majority vote and an optimized feedforward neural network.

- For future work, more feature selection, models, and fusion techniques will be investigated.

Supplementary material (R code)

- Complete R code and data analysis steps

https://www.box.com/s/yjmaiizovjxufdiq7zxl

- Or

- Email Tarek Abdunabi at:

  tabdunab@uwaterloo.ca