Analysis of the Study on Automated Refinement of Somatic Variant Calling using ML algorithm

Shubhangi Kaushik

Abstract

Goal:- Exploration and evaluation of the results of various Machine Learning algorithms, on somatic variant refinement of sequencing data obtained from studies conducted at the McDonnell Genome Institute (MGI).

Results:- Comparison of the important features obtained from Gradient boasting, against the features obtained from the Random forest approach used in the previous study.

The project gave a glimpse on utilizing files like pickle file for data analysis. Identification and resolution of class imbalance was studied, along with the encoding methods like one hot encoding. Exposure to the implementation of gradient boasting method was obtained. Moreover, the interpretation of the graphs like Reliability graphs was studied.

Goal

Exploration of Random forest method, along with, the implementation of Gradient boosting in automated refinement of somatic variant calling was conducted. The subset of data obtained from studies conducted at the McDonnell Genome Institute (MGI) was used. The features obtained from both methods were compared, followed by, the interpretation and comparison of ROC curves and reliability diagrams. The importance of manual review was checked in the data.

Data and Pre-processing

Data

The analysis of the study on automated refinement of somatic variant calling is based on the paper **A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data**[1]. The data used was a subset comprised of 41,000 called and manually reviewed variants obtained from the studies conducted at the McDonnell Genome Institute (MGI).

The data was composed of 440 individual tumors representing nine different cancer types, along with information on genetic variants that were manually reviewed and classified as somatic, germline, ambiguous, or failed. These labels were the composition of the class column in the data.

The rows in the data set represent individual genetic variants, and columns were 71 features such as class, cancer type, sample type, tumor read depth and normal read depth.

Pre-processing

During pre-processing, the data related to acute myeloid leukemia case (AML31) was separated. While conducting pre-analysis of the data, an imbalance was found in the class labels, hence the label ’germline’ was merged into ’fail’ to resolve the issue.
As the next step, the class column was separated from the rest of the data and was formed into a data set ’Y’, where as, the remaining data formed a data set ’X’. Following this, one hot encoding binarized the categorical labels in the ’Y’ data set and stored it in a data set ’Y one hot’. Splitting the ’X’ and ’Y’ data-sets into train and test sets concluded the pre-processing steps.

Methods

The analysis started with finding the optimized hyper-parameters for the chosen machine learning algorithms, Random Forest and Gradient Boosting Classifier. The model was created for each whose performances were later evaluated. Finally, after the evaluation, the most significant features were found for each model.

0.1 Hyper-parameter Optimisation

The hyper-parameters of the models were fine-tuned using 5-fold cross-validation on pre-processed data. In n-fold cross-validation, multiple rounds of training and testing are conducted on a range of user-defined parameters. The parameters giving the best AUC scores[4] were chosen for the final model.

0.2 Machine Learning Models

The optimized parameter values were utilized to train the models using the following algorithms:

0.2.1 Random Forest

Random Forest[2] is a bagging method that uses Decision Trees. In addition to the random subsets of observations, random subsets of features are also used to create these Trees. The final prediction is made from the majority vote after each tree casts a vote.

0.2.2 Gradient Boosting

Gradient Boosting[3] trains weaker algorithms like Decision Trees sequentially, such that where every later model corrects the mistakes made by the former model. A weight based on the performance of the model is assigned and the final prediction is made based on these weights.

0.3 Evaluation

The machine learning models were evaluated using the prediction probabilities calculated using K-Fold cross-validation technique. The number of folds was set to 10, which means the data set splits into 10 equal-sized folds and each fold act as a test set while the rest acts as the training set. The data was also shuffled randomly before the split to avoid any biases. The probabilities for each label for every row were used to calculate various scores of the evaluation matrix which included accuracy, precision, recall and f1-score.

ROC (Receiver Operating Characteristic) Curve[4] for each class was also generated using the probabilities. It plots the true positive rate against the false positive rate at different classification thresholds.

Another graph called the ”Reliability Diagram” was also plotted through which the trend of correctly and incorrectly predicted classifiers could be visualized.

Results and Discussion

The performance of two algorithms was assessed using Evaluation Scores obtained from the Confusion Matrix. Based on the data presented in Table 1 and Table 2, both algorithms performed similarly overall. However, Random Forest demonstrated slightly better scores compared to Gradient Boosting.

Model	Accuracy	
	Random Forest	Gradient Boost
	0.8927193301783765	0.8817983254459411

Table 1: Accuracy of the Random Forest and Gradient Boosting Algorithms after 10-fold cross-validation

Model	Random Forest				Gradient Boost			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Ambiguous	0.85	0.78	0.82	7189	0.83	0.77	0.80	7189
Fail	0.91	0.92	0.92	8015	0.91	0.91	0.91	8015
Somatic	0.90	0.94	0.92	12266	0.89	0.93	0.91	12266
micro avg	0.89	0.89	0.89	27470	0.88	0.88	0.88	27470
macro avg	0.89	0.88	0.88	27470	0.88	0.87	0.87	27470
weighted avg	0.89	0.89	0.89	27470	0.88	0.88	0.88	27470
samples avg	0.89	0.89	0.89	27470	0.88	0.88	0.88	27470

Table 2: Classification Report of the Random Forest and Gradient Boosting Algorithms after 10-fold cross-validation. It includes the scores of Precision, Recall, F1 Score and Support for each label and also for overall averages.

ROC curves were generated for both models, and the AUC was calculated for all three classes(Figure 1). Both the models seem to be working extremely ell. In terms of the ”Ambiguous” class, the Random Forest model covered 98% of the AUC while Gradient Boosting covered 97%. For the ”Somatic” class, both models exhibited even greater coverage, with Random Forest covering 1% more area than Gradient Boosting. In the ”Fail” class, both models achieved a 99% AUC. Random Forest showed slightly better performance compared to Gradient Boosting, although the difference was minimal.

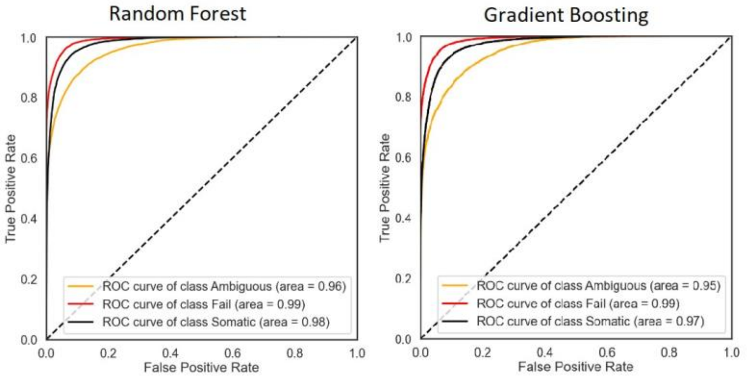

Figure 1: ROC curves for each Label for Random Forest and Gradient Boosting Model. Performance was checked for the probabilities obtained during cross-validation for the three classes, ambiguous, fail, and somatic.

Reliability Diagrams were also generated for the two models (Figure 2). Gradient Boosting seems to be working better based on r value. Gradient Boosting showed better results here.

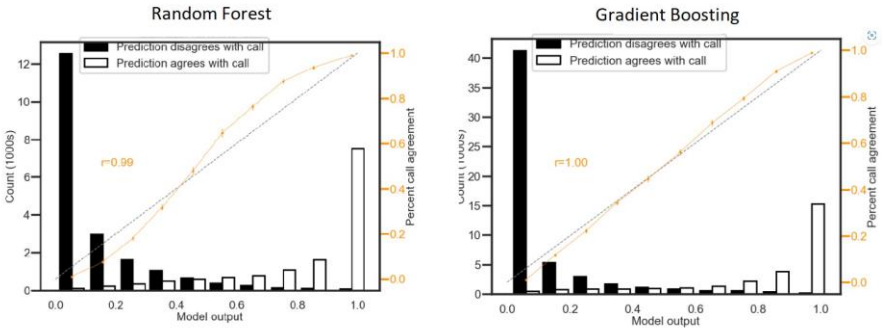

Figure 2: Reliability Diagram for Random Forest and Gradient Boosting models. Reliability diagram illustrates the alignment between model outputs and probabilities (ranging from 0 to 1) using cross-validation data. It consists of 2 main parts, (a)the bar plot and (b)the diagonal line. (a)The bar graph consists of 10 equidistant bins of model's output. The black bars indicate the model's calls disagree with the manual review calls while the white bar indicates they agree. (b)The black diagonal line represents a perfect prediction of probabilities. The coloured line represents the model. The closer the coloured line is to the black line, the better the prediction. The output is based on Pearson's correlation. In this analysis, it could be noticed that for both of the ML methods the prediction disagreement was high and agreement probability was very low for 0.0 model output. The disagreement probability was observed to decrease drastically, along with a gradual increase in agreement probability along the x axis. Eventually for 1.0 model output the agreement probability was the highest and disagreement probability the lowest.

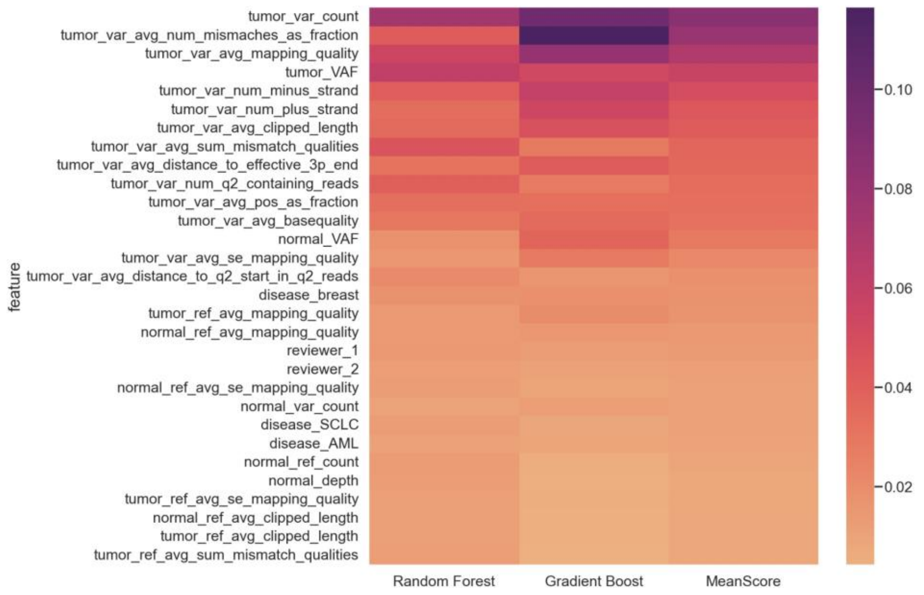

Figure 3: Important features predicted by the random forest and gradient boost are represented here using Heatmap. The mean of both models for each feature is calculated and then the data is sorted on it. ’tumor var count’ is the most important feature overall. ’tumor var avg num mismatches as fraction’ was the most important feature for Gradient Boost, but for Random Forest it gave lower score comparatively. Either similar or vice-versa patterns could be observed for other features as well.

The most important features were identified for both models. It was observed that most of the important features were the same for both models. Hence, the importance score was averaged and Figure 3 was obtained. Two of the reviewers can be observed in the list.

The importance of somatic variant refinement is highlighted in the above analysis, where both algorithms placed reviewers’ decisions high on the list of significant features. This suggests that manual review plays a critical role in improving the accuracy of final variant calls. By carefully filtering and reviewing the list of identified variants, researchers can gain a more accurate understanding of the genetic basis of a patient’s cancer and develop personalized treatment plans that target specific targets in the tumor cells while minimizing side effects for the patient.

In the steps involving manual refinement of somatic variant calling[1] , the entire data was manually reviewed by an experienced reviewer. Based on the review, they labeled the classes for the respective rows as
‘somatic’—a variant that has sufficient sequence read data support in the tumor. ‘ambiguous’ —a variant with insufficient sequence read data support; ‘germline’—a variant that has sufficient support in the normal sample beyond what might be considered attributable to tumor contamination of the normal; ‘fail’—a variant with low variant sequence read data support and/or reads.

The main idea of doing this manual review was to obtain a list of features deemed important by the domain experts, which could be compared against the important feature prediction conducted by the ML methods; In order to check the accuracy of the automated refinement and conclude whether it could be used to extract important features when provided with the data.

During the automated refinement step using ML method, the main idea is to build a automated way to predicted the important features, which are as accurate or closely accurate to the features obtained from the manual reviews of the experts in the domain. The process, in the attempt to automate, involved applying ML methods like Random forest and Gradient boosting, on the manually reviewed data set. After splitting the data set in test and train, the ML algorithms are trained on the train set along with conducting hyper parameter optimisation. The evaluation of the model's working is done using the test set. Ultimately, the important features are extracted, followed by the validation using the feature obtained from the manual review. This enabled determining whether the somatic variant calling could be automated with good accuracy.

Based on the important feature obtained from both of the methods. According to figure 3, Features like ”tumor var count” and ”tumor VAF” have been given somewhat similar importance by both of the methods. Nonetheless Gradient boosting gave higher score to ”tumor var avg num mismatches as fraction” unlike random forest.All over the results were similar to the feature importance obtained through manual review in the previous study.

Appendix

Author details

References

1. Peter Ronning Katie M. Campbell Alex H. Wagner Todd A. Fehniger Gavin P. Dunn Ravindra Uppaluri Ramaswamy Govindan Thomas E. Rohan Malachi Griffith Elaine R. Mardis S. Joshua Swamidass Benjamin J. Ainscough, Erica K. Barnell and Obi L. Griffith. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. page 14.
2. Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
3. Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.
4. James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.