RESEARCH

Paper 3 in applied machine learning

Aakash Nepal and Cuong Gia Pham

Full list of author information is available at the end of the article
*Equal contributor

Abstract

Goal of the project: The project aims to develop and evaluate the performance of an SVM model for a somatic variant refinement problem and compare it to the Random Forest model from Ainscough Paper[1] using AUC values and feature importances.

Main results of the project: Machine learning models are able to improve the detection of clinically actionable variants mislabeled by manual refinement strategies. From that, it will have experiments to reduce the time-consuming and resources for validating variants.

Personal key learning :

1 Pham Gia Cuong: Understanding the manual review for variant calling.
2 Aakash Nepal: was introduced to somatic refinement problem, Multilabelbinarizer, learned more about SVM and RF.

Estimation working hours:

1 Pham Gia Cuong: 7 hours per week
2 Aakash Nepal: 7 hours per week.

Project evaluation: 1

Number of words : Approx. 1350(without abstract, captions and appendix)

1 Scientific Background

In cancer genomics and clinical cancer assessments, somatic variant refinement is essential to removing false positive variant calls. In order to assure correct downstream analysis, particularly in therapy-guiding clinical contexts, it incorporates heuristic filtering and manual evaluation. Underreported refining processes, however, result in discrepancies and poor reproducibility. Automated somatic variant refinement techniques using machine learning algorithms have been developed to address this issue, saving time and money while enhancing repeatability in genomics and clinical applications.

2 Goal of the project

The goal of the project is to develop a Machine Learning(ML) model (Support Vector Machine in our case) to automate the somatic variant refinement problem, and also evaluate its performance, and then compare it to the Random Forest model (RF) described in the Ainscough paper[1], both in terms of AUC value and feature importance analysis.

3 Data and Preprocessing

3.1 Data assembly

The data used in this project is based on a training dataset of 41000 variants obtained from 21 research studies, including 440 individual tumor cases taken from nine cancer subtypes. Capture sequencing (14234 variants), exome sequencing (14044 variants), and genome sequencing (12722 variants) were used equally. Besides that, the information from the manual evaluation of each variant has been added to this dataset. From that manual evaluation, 18381 were determined to be somatic, 10643 to be ambiguous, 8854 to be failed, and 3122 to be germline. Both hematopoietic (10583 variants) and solid tumors (30417 variants) are included in the training data, which frequently have unique features during manual variant refinement[1].

3.2 Preprocessing

The training pickle dataset was imported using the read_pickle() function from the pandas library. The data was then sorted and filtered by removing an unwanted case named "AML31". Germline calls were reclassified as failed calls, creating a three-class classification problem. The features and labels were separated into arrays. The dataset was then scaled using StandardScaler() function from the scikit−learn library. The dataset was divided using the train_test_split() function into two parts: training and testing (33%). Labels were one-hot encoded in a single pass for multi-class classification by using MultiLabelBinarizer() from scikit−learn.

4 Methods

For the model in this project, we utilized the random forest and linear support vector machine (SVM-lin) for somatic variant refinement. RF uses multiple unique classifiers called decision trees and averages them to generate a single prediction. According to the given script, the random forest was trained using the different parameters n_estimater from 100 to 10000 and trees max_features = 8, where n_estimators describes the number of trees and max_features indicates the maximum features at each split. For SVM-lin, the model was trained using different regularization parameters from 0.0005 to 0.01. The regularization parameter is a control of fitting parameters. As the magnitudes of the fitting parameters increase, there will be an increasing penalty on the cost function. After that, the parameter corresponding for the highest AUC of each model has been chosen.

After choosing the appropriate parameter, to evaluate the performance of the learning models, the Scikit-learn library was used to construct one-versus-all receiver operator characteristic curves and calculate the area under the curve metrics (AUC) in order to compare model performance. Two-thirds of the data were randomly chosen to serve as a training set, while the final third was used as the hold-out test set. We used tenfold cross-validation for model selection and hyper-parameter adjustment on the training set. A model was trained on the training set and compared to the hold-out test set after the selection of models and hyperparameters to understand model performance.

The linear support Vector Machine(SVM-lin) model used coefficients, whereas the random forest(RF) model used its built-in feature importance function to determine the feature importance.

5 Results and Discussion

5.1 Refinement

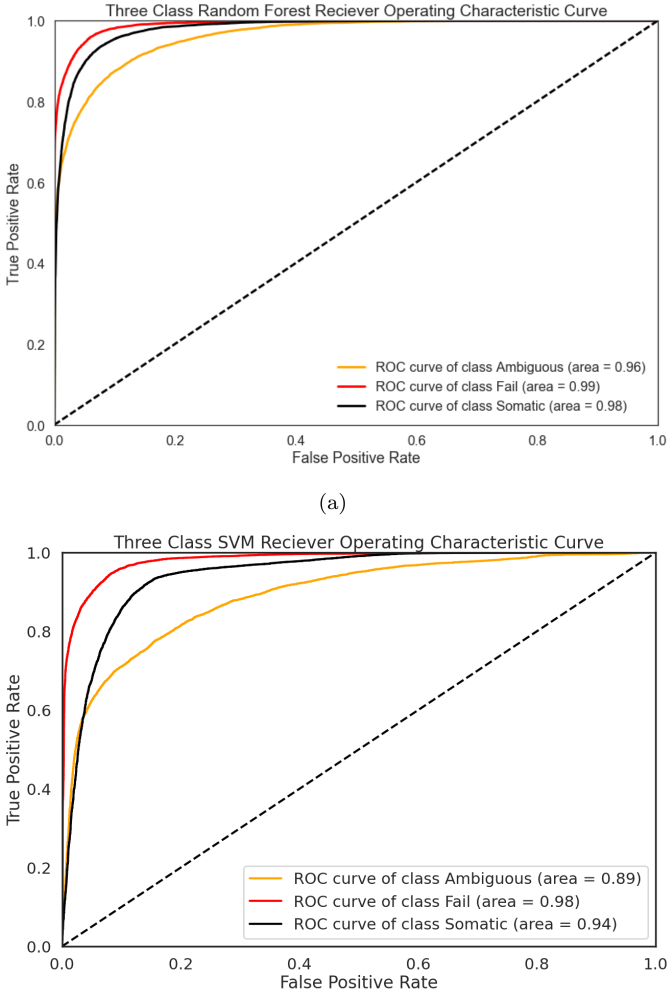
In the context of cancer genomics, somatic variants are those which have a genetic mutation and occur in tumor cells only. This is a reason that further investigation of these variants is potentially useful to discover new cancer therapies and diagnoses. However, the reliable detection of the somatic variants is difficult due to the high amount of noise in the sequencing data. The refinement process is therefore a critical step that helps to increase the accuracy of somatic variant calls by detecting and filtering out the false positives during the somatic variant detection. Furthermore, the refinement step ensures that only true positive somatic variants are used for further downstream analysis.

5.2 Manual refinement and automatic refinement

For the refinement step, first, some variant caller software like MuTect, SAMtools, and VarScan are used, after that by using some simple criteria such as removal of variants with low VAF (for example, $< 5\%$) or low coverage (for example, $<$ 20X), the satisfied above criteria variants founded by automated variant caller were subjected. However, there are still many more false positives variants, which are hard to be detected by simple criteria. Hence manual refinement step came into the process. Using the Integrative Genomics Viewer (IGV) and IGVNav, the manual refinement process includes visually inspecting and determining each variant(Somatic, ambiguous, germ-line, or failed). The primary aim is to thoroughly investigate the variant-supporting reads, estimate their number and quality, identify any artifacts or discrepancies, and make informed decisions regarding the validity of each variant. Furthermore, this procedure involves the evaluation of coverage, variant allele frequencies, mapping quality, base quality, and lack of support in normal samples, and synthesizing available information to make a file with an abundance of information about each variant[2]. Moreover, Machine learning algorithms such as Random Forest and Support Vector machines can improve the automated refinement step by reducing the false positive variants more effectively which is discussed in the next section(5.3).

5.3 Evaluation of machine learning models

We have used the 4100 variant dataset on two models (Random forest and linear support vector machine). To prevent overfitting, one-third of the dataset was selected as a test set, and the models were trained on the remaining two-thirds. Besides that, to have a reliable result, a tenfold cross-validation strategy has been used. After running cross-validation, the random forest model demonstrated a better performance than SVM in all three classes. The linear SVM model's performance indicates the limited ability of the model in the classification of the ambiguous class (AUC=0.89). In contra, its performance in the classification of failed and somatic classes (corresponding to AUC= 0.98 and AUC=0.94) is comparable with the performance of RF (corresponding to AUC=0.99 and AUC=0.98) (Figure 1).



(a)

(b)

Figure 1 (a) ROC curve of Random forest on 3 classes (b) ROC curve of linear support vector machine on 3 classes

5.4 Feature importance

The ranking of the feature importance was normalized and a heatmap as shown in Figure 2 was generated. When compared between the ML models and the manual reviewer, it was seen that although there were some common features for the classification decisions, there were also many differences. For example, the correlation between SVM-lin and the manual reviewer was very weak(Pearson r = 0.05) compared to the moderate correlation(Pearson r = 0.52) between RF and the manual reviewer. Also, the correlation between both models was weak(Pearson r = 0.13). When going more into the heatmap, we can see that, some feature importance like ”tumor variant average mapping quality” and ”tumor variant average base quality” were marked important by both manual reviewers and the ML models. Both ML models highly ranked the features ”tumor variant average sum mismatch qualities” and ”tumor variant average clipped length” as important, but not the reviewers. Some features like ”tumor variant count” and ”tumor VAF” were marked impor-

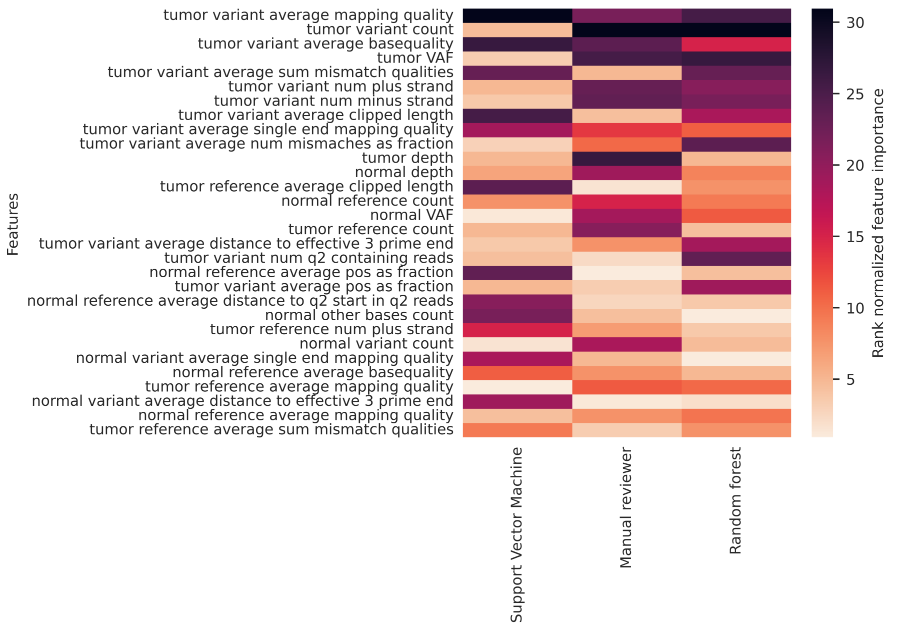


Figure 2 Heatmap depicting feature importance as identified by SVM-lin, RF, and manual reviewer. The darker the color, the more important the normalized feature importance. RF seems to be more comparable to manual reviewer than SVM-lin, which has a poor correlation with manual reviewer.

tant by both manual reviewers and the RF, but not SVM-lin. Although there were a few similarities in feature importance, we may conclude that the above-mentioned features like ”tumor variant average mapping quality” and ”tumor variant average base quality” are important for variant analysis.

6 Contributions

1 Pham Gia Cuong: Overall 50% report and did the work setup the environment, running Random forest.

2 Nepal Aakash: Overall 50% report and ran SVM-lin plus extracted the feature importance.

References

[1] Benjamin J. Ainscough et al. "A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data". In: *Nature Genetics* 50.12 (2018), pp. 1735–1743. DOI: 10.1038/s41588-018-0257-y. URL: https://doi.org/10.1038/s41588-018-0257-y.

[2] Erica K. Barnell et al. "Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples". In: *Genetics in Medicine* 21.4 (2019), pp. 972–981. DOI: 10.1038/s41436-018-0278-z.