CBM 2020 Project 2 Report

Piersilvio De Bartolomeis, Anej Svete, Mirko De Vita

December 2020

1 Introduction

This report summarizes our work on the second project at the course of Computational Biomedicine. We opted for a machine learning based approach due to our computer science background.

To implement it, we closely read through several publications dealing with similar goals and took inspiration from those. In the end, we came up with two models to complete our task due to the multitude of SNP types present in the files. Namely, we separated the variants into 3 main categories: nonsynonymous SNP's, synonymous SNP's, and "other", where "other" includes for example intronic and intergenic SNP's. Since the first two categories include the variants that are by far the most important for disease prediction and are best covered in the literature, we focused on predicting the effect of those alone and left the last category as uncertain.

In the following, the model and the results for the first 2 categories are presented. Some additional figures can be found in the "figures" directory of the repository (in the P2 folder).

2 Model Selection

In both cases, we used a boosting ensemble model from the well-known XGBoost library. Due to limited amount of training data and noise, we avoided too much model tuning to prevent over-fitting and left most of the parameters as default. We modeled the task as a binary prediction problem of pathogenic or benign in the end. Although we had additional labels at hand (e.g. possibly pathogenic/benign), we decided against using them, again because of relatively small training data set. We did, however, experiment with multi-class classification in the case of nsSNP because the quality of the data was highest in that case and we report on some of the results below.

3 Results

3.1 Nonsynonymous SNP effect prediction

We had the biggest room for creativity here, but modelling was still difficult because of the diverse data sources and inconsistent annotations. In the end, we opted combine various features from Polyphen2 [Adzhubei, Jordan, and Sunyaev 2013], Gerp [Cooper et al. 2005], Rhapsody [Ponzoni et al. 2020], and Cadd [Rentzsch et al. 2018]. Most of them were obtained from MyVariant [Xin et al. 2016], while the Polyphen2 were downloaded directly from the service and Rhapsody ones via its Python API.

As mentioned, we experimented with both multiclass and binary classification and in the end opted to only keep the binary version. The multi-class one suffered some performance drop and did not handle the more similar classes well, as can be seen in figures

The results of our final classifier are presented in figure 1. All of them correspond to a 5-fold CV average. We think the model performed reasonably well and the performance was stable across the folds with the standard deviation of the presented scores in the order of $\mathcal{O}(10^{-3})$. Reading the features importance's given by XGBoost, we see that, by far, the most important features turn out to be those corresponding to (in order of importance) the PHAST and PhyloP vertebrate conservation scores, PolyPhen2 score 1, the PolyPhen2 score difference, and the BLOSSUM score, in both binary and multi-class models. The multi-class model did not turn out to offer much more information and had trouble distinguishing between the similar classes.

3.2 Synonymous SNP effect prediction

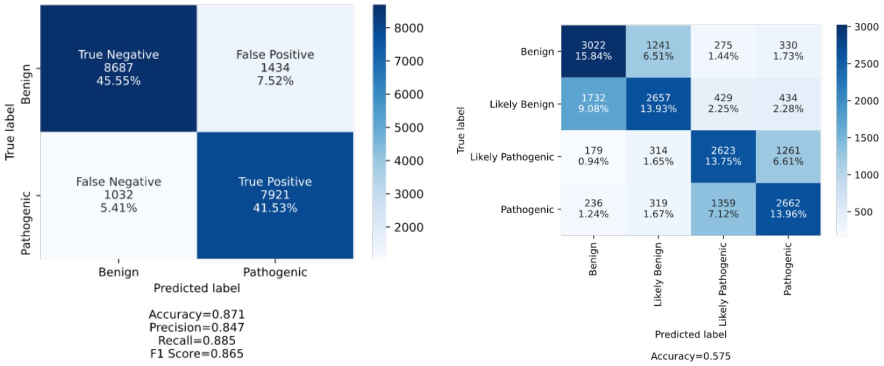
In order to obtain the annotations for sSNPs we relied on two different sources, namely: MyVariant.info [Xin et al. 2016] and SeattleSeq Annotation 138.
In the end, we combined features from Gerp [Cooper et al. 2005], Cadd [Rentzsch et al. 2018] and general splicing annotations. Of note, the feature's importance computed by XGBoost identified the three most important features:

1. PhyloP mammalian conservation score

2. Distance to the splice site

3. Gerp S score

The challenge with sSNP was the highly imbalanced dataset. We performed random undersampling (with ratio $65 : 100$) and validated our classifier with 5-fold CV (undersampling was not applied to the validation folds). The results of our final classifier are presented in Figure 3. We argue that the classic metrics used in the literature for validation do not take into account the high class imbalance. Hence, we propose the Balanced Accuracy [Kay H. Brodersen and Buhmann 2010] as validation metric and we plot the probability density function using the inverse cdf method in Figure 4. Overall, we think that there is not



(a) Binary

(b) Multi-class

Figure 1: Validation confusion matrices of the binary and the multi-class classifiers for nsSNP classification.

enough data available for the model to perform quite well in the context of sSNP interpretation. Nevertheless, it is well suited for the purpose of identifying an interesting subset of sSNP and might be used by biologist to pinpoint significant variants that require further research.

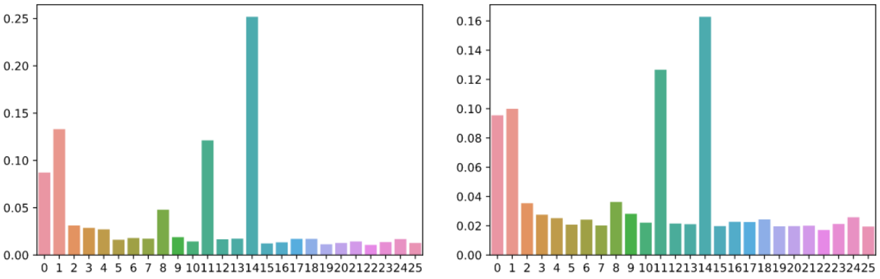
4 Conclusion

Variations in the DNA sequence can affect how humans develop diseases and respond to drugs, pathogens, vaccines and other agents.
This makes SNPs valuable for biomedical research and for developing pharmaceutical products or medical diagnostics.
We elaborated on the current state of the art machine learning methods applied to the study of SNPs to develop our predictor.
The main challenge was collecting all the useful features from different sources in order to improve the accuracy of our predictions.
We hope that the availability of data will increase in the years to come, so that it would be much easier to work on this exciting challenge!

References

[AJS13] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev. “Predicting functional effect of human missense mutations using PolyPhen-2”. In: *Curr Protoc Hum Genet* Chapter 7 (Jan. 2013), Unit7.20.

[Coo+05] G. M. Cooper et al. “Distribution and intensity of constraint in mammalian genomic sequence”. In: *Genome Res* 15.7 (July 2005), pp. 901–913.



(a) Binary

(b) Multi-class

Figure 2: Feature importances of the binary and the multi-class classifiers for nsSNP classification.

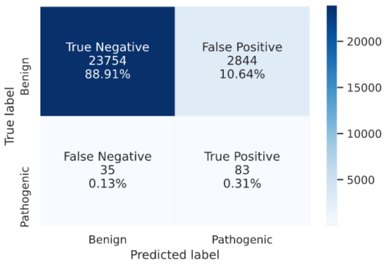


Figure 3: Validation confusion matrix of the binary classifier for sSNP classification.

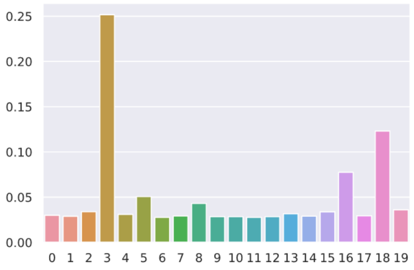


Figure 4: Feature importances of the binary classifier for sSNP classification.

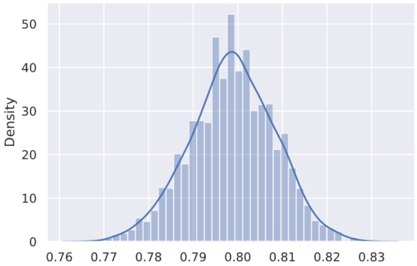


Figure 5: Balanced Accuracy Posterior Distribution.

[KB10] Klaas E. Stephan Kay H. Brodersen Cheng Soon Ong and Joachim M. Buhmann. “The balanced accuracy and its posterior distribution”. In: *International Conference on Pattern Recognition* (2010).

[Pon+20] Luca Ponzoni et al. “Rhapsody: predicting the pathogenicity of human missense variants”. In: *Bioinformatics (Oxford, England)* 36.10 (May 2020), pp. 3084–3092. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btaa127`. URL: `https://europepmc.org/articles/PMC7214033`.

[Ren+18] Philipp Rentzsch et al. “CADD: predicting the deleteriousness of variants throughout the human genome”. In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D886–D894. ISSN: 0305-1048. DOI: `10.1093/nar/gky1016`. eprint: `https://academic.oup.com/nar/article-pdf/47/D1/D886/27436395/gky1016.pdf`. URL: `https://doi.org/10.1093/nar/gky1016`.

[Xin+16] Jiwen Xin et al. “High-performance web services for querying gene and variant annotation”. In: *Genome Biology* 17.1 (May 2016), p. 91. ISSN: 1474-760X. DOI: `10.1186/s13059-016-0953-9`. URL: `https://doi.org/10.1186/s13059-016-0953-9`.