Variant relevance prediction in extremely imbalanced training sets

Max Schubach[1,2*], Matteo Re[3], Peter N Robinson[4], and Giorgio Valentini[3]
[1]*Berlin Institute of Health, Berlin, Germany*
[2]*Institute for Medical and Human Genetics, Charite – Universitatsmedizin Berlin, Berlin, Germany*
[3]*Department of Computer Science, University of Milan, Milan, Italy*
[4]*The Jackson Laboratory for Genomic Medicine, Farmington CT, USA*
*Correspondence: max.schubach@charite.de

Abstract The interpretation of non-coding variants still constitutes a major challenge in the application of whole-genome sequencing. For example for Mendelian diseases only several hundreds of pathogenic regulatory mutations are known but millions of possible neutral sites can be derived. In this context, machine learning (ML) methods for predicting disease-associated non-coding variants are faced with a chicken and egg problem - such variants cannot be easily found without ML, but ML cannot begin to perform effective until a sufficient number of instances have been found. Recent ML-based methods for variant prediction do not adopt specific imbalance-aware learning techniques to deal with imbalanced data that naturally arise in several genome-wide variant scoring problems, thus resulting in relatively poor performance with reduced sensitivity and precision. Here we present a ML algorithm based on resampling techniques and a hyper-ensemble approach, called hyper SMOTE Undersampling with Random Forests (hyperSMURF), which is able to deal with extremely imbalanced datasets. HyperSMURF outperforms previous methods on two different published imbalanced variant datasets: regulatory Mendelian mutations and classification of microRNA/SNP pairs into eQTLs or non-eQTLs. We show that imbalance-aware ML is a key issue for the design of robust and accurate prediction algorithms and the provided method hyperSMURF can be effectively applied to discover disease-associated variants out of millions of neutral sites from whole genome sequencing.

Proposal Computational algorithms for the analysis of non-coding deleterious variants are faced with special challenges owing to the rarity of confirmed pathogenic mutations. In this setting, classical learning algorithms, such as support vector machines (SVM) [1] or artificial neural networks [2] tend to generalize poorly, because they usually predict the minority class with very low sensitivity and precision [3].

HyperSMURF is a method specifically conceived to deal with extreme imbalanced data. To achieve high coverage of the available input data as well as high accuracy of the predictions, hyperSMURF simultaneous oversamples the minority class and undersamples the majority class to generate balanced training sets, each one used to train a different random forest (RF) ensemble. The predictions of the trained models are finally combined through an hyper-ensemble approach (ensemble of ensembles) to obtain an

overall consensus prediction. The burden of the computation can be reduced by using parallel computation techniques, since the learning processes of the RF ensembles are largely independent and easily parallelizable. Moreover the diversity between the learners and the balancing between positive and negative examples introduced by the over and undersampling techniques avoid the bias toward the majority class and promote good generalization behavior, while the hyper-ensemble approach provides more accurate base learners and robust predictions.

For performance validation we retrained two published datasets. The first dataset is highly imbalanced and consists of 406 regulatory Mendelian mutations compared to 14,755,199 non-deleterious sites [4]. The second dataset contains published cis-microRNA-eQTL data for the human GM12878 cell line together with promoter predictions, and other functional annotations from Budach et al. [5] to determine the relationship between functional elements and microRNA regulation (4785 SNPs associated with microRNA–eQTL and 1,997,341 SNPs associated with non-microRNA–eQTL observations).

On the eQTL data hyperSMURF was compared to the original trained logistic regression model defined by Budach et al. [5]. For regulatory Mendelian mutations the underlying machine learning methods of different well known non-coding scores were retrained on the complete imbalanced Mendelian dataset. We used the SVM underlying the CADD score [6], the modified version of the RF algorithm of GWAVA [7], and the unsupervised approach of the Eigen and EigenPC score [8]. Figure 1 shows that in terms of the precision/recall curve hyperSMURF achieves significantly better results than the other retrained methods with both the Mendelian and eQTL data.

We think that this method could be helpful to discover trait and disease-associated variants in contexts characterized by imbalanced genomic data. Moreover our work shows that imbalance-aware machine learning is a central issue of method development for the prediction of deleterious variants in non-coding regions.

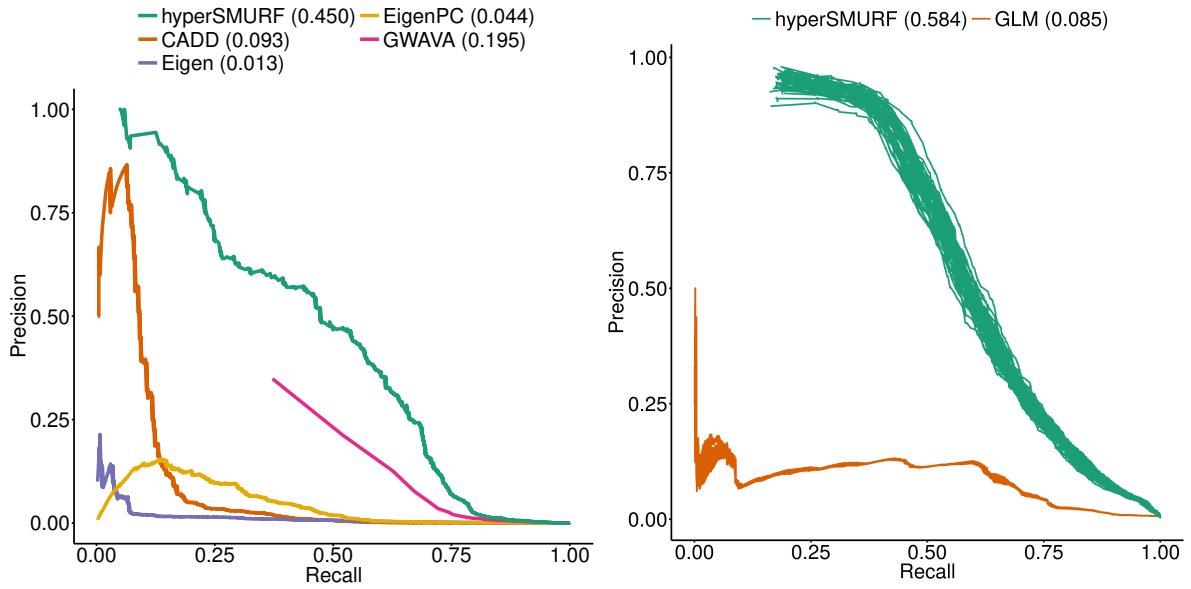
References

[1] Cortes, C. & Vapnik, V. Support vector networks. *Machine Learning* **20**, 273–297 (1995).

[2] Bishop, C. M. *Neural Networks for Pattern Recognition* (Clarendon Press, Oxford, 1995).

[3] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. & Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **42**, 463–484 (2012).

[4] Smedley, D. *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am. J. Hum. Genet.* **99**,



(a) Mendelian data **(b)** eQTL data

Figure 1: **(a)** Precision/recall curves of hyperSMURF compared to the retrained methods of CADD, GWAVA, Eigen and EigenPC on the Mendelian data. Performance was measured using a cytogenetic band-aware 10-fold cross-validation introduced by Smedley et al. [4]. Values in brackets show the area under the precision/recall curve. **(b)** Precision/recall curves of hyperSMURF compared to the logistic regression model (GLM) used in Budach et al. [5] on the eQTL data. Performance was measured using a 50-times repeated experiment with a random data split of 75 % for training and 25 % for testing. Values in brackets show the average area under the precision/recall curve over all 50 curves.

595–606 (2016).

[5] Budach, S., Heinig, M. & Marsico, A. Principles of microRNA Regulation Revealed Through Modeling microRNA Expression Quantitative Trait Loci. *Genetics* **203**, 1629–40 (2016).

[6] Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).

[7] Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294–296 (2014).

[8] Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**, 214–20 (2016).