VARIANT RELEVANCE PREDICTION
IN EXTREMELY IMBALANCED TRAINING SETS

Max Schubach[1,2], Matteo Re[3], Peter N Robinson[4], Giorgio Valentini[3]

[1] Berlin institute of Health (BIH), Berlin, Germany
[2] Charité – Universitätsmedizin Berlin, Institute of Medical and Human Genetics, Berlin, Germany
[3] Department of Computer Science, University of Milan, Milan, Italy
[4] The Jackson Laboratory for Genomic Medicine, Farmington CT, USA

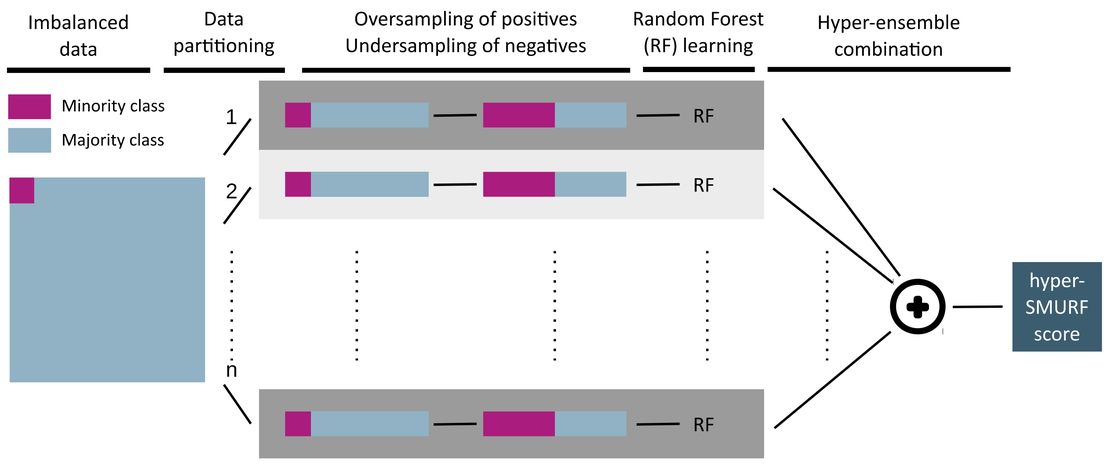
BERLIN INSTITUTE OF HEALTH
Charité & Max Delbrück Center

MOTIVATION

The imbalance between the small set of available disease-associated variants and the large set of benign genomic variants is intrinsic to genetic variant data, especially in non-coding regulatory regions of the human genome [1].

In this context machine learning (ML) for variant relevance prediction is like a chicken and egg problem – they cannot be easily found without ML, but ML cannot be applied effectively until a sufficient number of instances has been found. Unfortunately, most common ML-based methods [2,3,4] do not adopt specific imbalance-aware learning techniques to deal with imbalanced data, resulting in a poor performance with reduced sensitivity and precision [5].

HYPERSMURF



Outline of hyperSMURF

We designed the ML method *hyper SMOTE Undersampling with Random Forests (hyperSMURF)* for extremely imbalanced datasets. HyperSMURF adopts over- and undersampling techniques [6,7] to increase the count of the minority class and reduce the cardinality of the majority class. The resulting training sets are relatively small and balanced, thus avoiding a bias of learning algorithms towards the majority class and reducing runtime.

Data partitioning: The training set is partitioned into *n* partitions using all minority class instances (magenta) in every partition and an equal split of the majority class instances (blue).

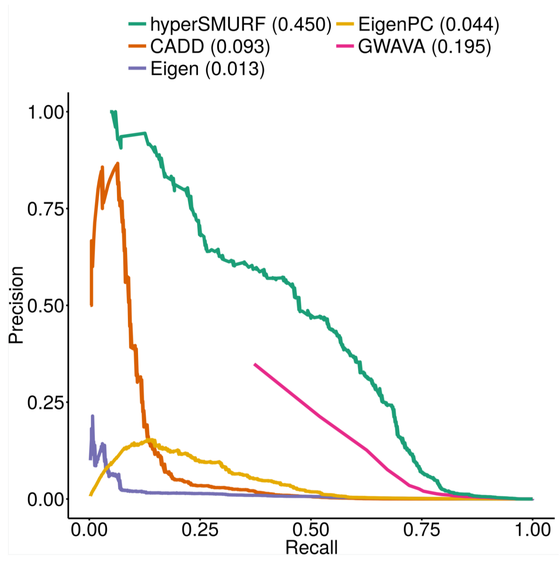
Over- and Undersampling: In every partition the minority instances are oversampled using the synthetic minority over-sampling technique (SMOTE) [6] and the majority instances are subsampled to an appropriate size. This results in balanced datasets per partition and a comprehensive coverage of the input data.

RF learning: For each partition a random forest (RF) is trained.

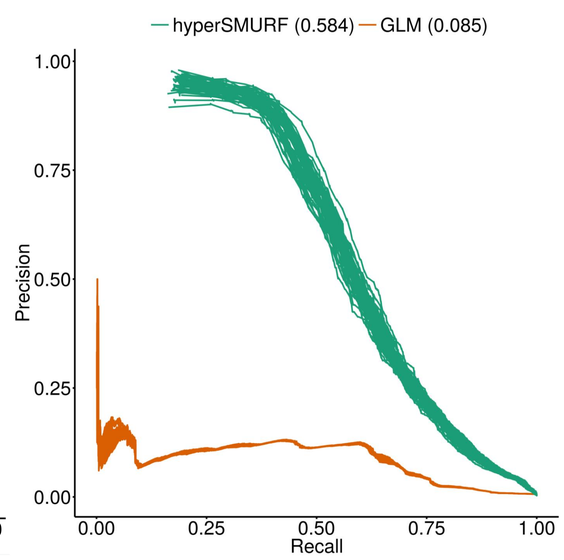
Hyper-ensemble combination: A new variant will be classified by an average vote of the resulting *n* RFs.

PERFORMANCE TESTING

	Regulatory Mendelian mutations [1]	miRNA-eQTL SNPs [8]
Positives	406	4785
Imbalance	1:36,000	1:400
Features	26 genomic attributes (e.g. conservation scores, and chromatin states)	18 genomic features (e.g. transcription factors, chromatin states, miRNA promoters)
Cross-validation (CV)	Cytoband-aware 10-fold CV [1]	50-times repeated sampling using random splits (75%/25%)
Comparison to retrained base-learners	SVM/CADD, Eigen, EigenPC and modified RF/GWAVA [2, 3, 4]	Logistic regression model (GLM) [8]



Regulatory Mendelian mutations



miRNA-eQTL SNPs

The hyperSMURF precision-recall curve (green) lies significantly above the other methods (Wilcoxon rank sum test, p-value < 10^{-9}) showing that our imbalance-aware procedure outperforms competing learning strategies on the same data set in prioritizing pathogenic variants at any given level of sensitivity.

CONCLUSION

These results strongly suggest that imbalance-aware learning strategies, like hyperSMURF, are essential for relevance prediction of disease associated non-coding variants in the human genome and we recommend their use to deal with imbalanced genomic data. HyperSMURF is implemented and freely available in R and Java. For more information see *https://hypersmurf.readthedocs.io* and *Schubach et al [9]*.

REFERENCES

[1] Smedley D, et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet. 2016;99(3):595–606.
[2] Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
[3] Ionita-Laza I, et al. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016; 48(2):214–20.
[4] Ritchie GRS, et al. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294–6.
[5] He H, et al. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.
[6] Chawla NV, et al. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57.
[7] Liu X, et al. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B Cybern. 2009;39(2):539–50.
[8] Budach, S, et al. Principles of microRNA Regulation Revealed Through Modeling microRNA Expression Quantitative Trait Loci. Genetics. 2016;203(4):1629–40.
[9] Schubach, M, et al. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. Sci Rep. 2017;7(1):2959.

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN
MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION
The Jackson Laboratory
UNIVERSITÀ DEGLI STUDI DI MILANO

Contact: max.schubach@bihealth.de

www.bihealth.org