1st International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2019),
20-22 March 2019, Barcelona, Spain

Topic: Artificial Neural Networks

A neural model for the prediction of pathogenic genomic variants in Mendelian diseases

Alessio Cuzzocrea[1], **Luca Cappelletti** [1], **Giorgio Valentini**[1]
[1] AnacletoLab, Dipartimento di Informatica, Università degli Studi di Milano, Italy
E-mail: valentini@di.unimi.it

Summary: The detection of pathogenic genomic variants associated with genetic or cancer diseases represents an open problem in the context of the Genomic Medicine. In particular the detection of mutations in the non-coding regions of human genome represents a particularly challenging machine learning problem, since the number of neutral variants largely outnumber the pathogenic ones, thus resulting in highly imbalanced classification problems. We applied neural networks to the detection of pathogenic regulatory genomic variants in Mendelian diseases and we showed that leveraging imbalance-aware techniques and deep learning algorithms, we can obtain state-of-the-art results, using a less complex model than those proposed in literature for this challenging prediction task.

Keywords: Neural Networks, imbalance-aware Neural Networks, Deep Learning, Prediction of pathogenic genomic variants, Mendelian diseases.

1 Introduction

An open problem in the context of Precision Medicine is the detection of the pathogenic variants associated with genetic Mendelian diseases. Indeed for most of the about 8000 different Mendelian diseases no known causative gene is known and hence no therapy is available for affected patients [15]. Recently several studies showed that most of the pathogenic variants associated with Mendelian disorders lie in the non-coding regulatory regions of the human genome [3].

For this reason several computational methods have been proposed to disentangle the regulatory mechanisms underlying Mendelian diseases and other disorders ranging from complex genetic diseases to cancer, using mainly supervised Machine Learning-based techniques to predict the pathogenicity of genomic variants in regulatory regions of the human genome [8, 16, 6].

Unfortunately only a very small amount of positive (pathogenic) variants are available to train learning machines and in this very imbalanced context, where neutral variants (negative examples) largely outnumber positive ones, machine learning methods are severely biased toward the majority class and are not able to detect pathogenic variants with a sufficient reliability. Very recently novel imbalance-aware machine learning methods have been proposed in this context, showing that applying together ensemble and sampling techniques we can significantly improve prediction results [9, 11].

Motivated by these results and by the very recent successful application of deep neural learning methods to Genomic Medicine [14], in this work we investigate whether a neural model, by adopting imbalance-aware techniques and deep learning techniques can obtain state-of-the-art results in this challenging prediction task.

In the next sections we propose two imbalance-aware neural models able to deal with highly imbalanced genomic data, and we experimentally show that they largely outperform "vanilla" neural models, achieving state-of-the-art results in the prediction of pathogenic regulatory variants in Mendelian diseases.

2 Methods

We introduce two imbalance-aware neural methods, able to deal with highly imbalanced genomic data. The first one *MiMiS-Net* (Mini-batch Minority class Sized Neural Networks) simply enlarges the mini-batch size applied during the training of the neural network. The second one *MiBa-Net* (Mini-batch Balanced Neural Networks), inspired by [10], uses sampling techniques to balance positive and negative examples in the mini-batch.

2.1 Mini-batch Minority class Sized Neural Networks (*MiMiS-Net*)

The main idea behind this approach consists in improving the likelihood that at least one positive example will be included in each mini-batch during the training phase. We show that this can be accomplished by simply appropriately enlarging the size of the mini-batch itself. Indeed when the data are highly imbalanced, the update of the weights is likely performed with most of the mini-batches including only examples of the majority (negative) class: in this situation the neural network tends to be biased toward the negative class, since it learns only from negative examples, and hence cannot recognize positive examples.

More precisely, let N be the overall number of available examples of the training set T, n the size of the mini-batch, and p the probability that a positive example will be randomly extracted from the overall training set. If N_p is the total number of positive examples in the training set, we can estimate $p \simeq \frac{N_p}{N}$.

Let X_n be a random variable that counts how many positives have been randomly drawn from T into a mini-batch of size n. Then X_n is distributed according to a binomial distribution $B(p, n, k)$ where k is the number of successes (positive examples) across n Bernoulli experiments each one with probability of success p. Then the probability $P(X_n \geq 1)$ that we have at least on positive example in a mini-batch of size n is:

$$P(X_n \geq 1) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

We can observe that

$$\begin{aligned} P(X_n \geq 1) &= 1 - P(X_n = 0) \quad (2) \\ = 1 - \binom{n}{0} p^0 (1-p)^n &= 1 - (1-p)^n \end{aligned}$$

Hence eq. 1 can be rewritten as:

$$P(X_n \geq 1) = 1 - (1-p)^n \tag{3}$$

If we would like to estimate the size n of the mini-batch needed for having at least one positive in the mini-batch itself with probability $P(X_n \geq 1)$, we can apply a log transform to eq. 3:

$$n = \frac{\log\big(1 - P(X_n \geq 1)\big)}{\log(1-p)} \tag{4}$$

Eq. 4 shows the mini-batch size n needed for having with probability $P(X_n \geq 1)$ at least one positive example in each mini-batch. It is easy to see that n is large for large values of $P(X_n \geq 1)$ and for small values of p, i.e. when we would like to be confident that at least one example is included in the mini-batch and when the data in the training set are imbalanced (Figure 1). For a reasonable probability (say $P = 0.8$) of having at least one positive example in the mini-batch, when data are imbalanced (say $p = \frac{1}{5000}$) we need a mini-batch size of at least $n = 8046$, a size significantly larger than those usually applied for mini-batch learning.

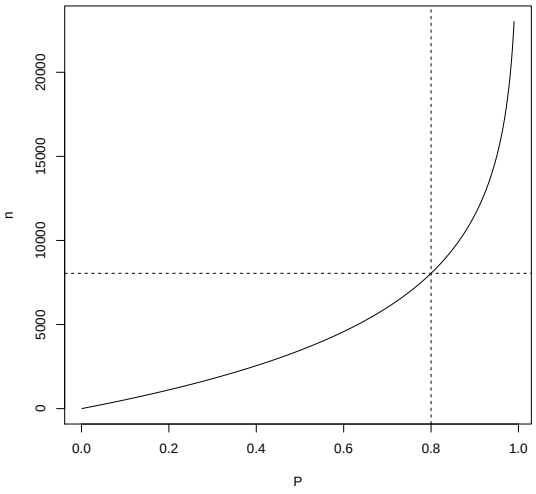


Figure 1: Plot of the size n of the mini-batch (vertical axis) for drawing with probability P (horizontal axis) at least one positive example included in it when the frequency of the positives in the training set is about $p = \frac{1}{5000}$.

2.2 Mini-batch Balanced Neural Networks (*MiBa-Net*)

Sampling procedures to deal with the imbalance of the data have just been proposed in machine learning and neural network literature [5] and have been proven successful in the context of the analysis of genomic data with ensemble methods [9, 10]. Here we propose to balance the mini-batch during the training of the neural network, in order to provide a number of positive examples (the minority examples) comparable with those of the majority (negative class). In this way at each mini-batch the weights of the network are updated taking into account in a balanced way both positive and negative examples.

The mini-batch generator samples with replacement, according to a uniform distribution, the positive examples by drawing a sample ratio $r_p \in (0, r_p^{max}]$ of the available positive examples: if $r_p < 1$ we subsample the positives, if $r_p = 1$ we have a bootstrap sample, for $r_p > 1$ we perform oversampling. Negative examples are sub-sampled without replacement according to the ratio $r_n \in (0, r_n^{max}]$ between the negatives and the positives in the mini-batch: if $r_n < 1$ we will have less negatives than positives in the mini-batch, if $r_n = 1$ positives and negatives are equally sized, and for $r_n > 1$ negatives outnumber positives in the mini-batch. As an example, if we have $N_n = 10^6$ negative examples and $N_p = 10^2$ positive examples, we have an imbalance $N_p/N_n = 1/10^4$. If we set $r_p = 1$ and $r_n = 1$ we can obtain a perfectly balanced mini-batch with 100 positives and 100 negatives. An

epoch, with this generator, is considered to be finished when all the negative samples are used. Notice that the positive samples may appear repeatedly among different batches in the same epoch, while each negative will appear only once in one specific mini-batch at each epoch.

3 Results

We evaluated the proposed methods *MiMiS-Net* and *MiBa-Net* on Mendelian data, by comparing them with a baseline "vanilla" Neural Network and with *hyperSMURF* [10], an imbalance-aware hyper-ensemble method that significantly outperformed other state-of-the-art methods such as *CADD* [8], *DeepSEA* [16], *Eigen* [6] and *GWAVA* [9] on this specific task [10].

3.1 Experimental set-up

For the experiments we used the data set of Mendelian Single Nucleotide Variants (SNV) in non-coding regions of the human genome originally collected in [12]. From this data set we used all the available manually curated 406 positive examples, and from the available 14 millions of neutral variants (negative examples) we randomly drew one million of examples, thus resulting in an imbalance $p \simeq \frac{1}{2500}$. To each SNV example are associated 26 features representing different characteristics of the genomic variants, ranging from G/C content, population-based features, to conservation scores and transcription and regulation annotations (see [12] for more details).

We trained the neural networks on all the genomic variants except those belonging to chromosome 19 (19018 examples) that have been left out for evaluating the generalization performance. In other words we performed a "chromosome aware" hold-out procedure and we did not use the examples of the test set (chromosome 19) to train the model. The main hyper-parameters of the model, i.e. different number of hidden layers (ranging from 1 to 4), the number of hidden neuron per layer (ranging from 2 to 100) have been selected by 5-fold cross-validation on the training set. We used the ReLU activation function for the hidden layers and a sigmoid for the output layer. We chose as loss function to be optimized the hinge loss with the logit function applied to the sigmoid output, and we applied both the Stochastic Gradient Descent (*SGD*) with fixed learning rate (0.01) and the *Adam* method [7] as optimization algorithms. The weight matrix of each layer have been initialized using the Glorot normal initializer [4]. Before training each feature has been standardized by subtracting its mean and dividing by its standard deviation across examples.

For evaluating the performance of the different methods we used the Area Under the Precision recall Curve (AUPRC), since it is well-known that in very imbalanced learning problems this metric is more informative than the Area Under the Receiving Operating Characteristic curve (AUROC) [2]. All the experiments and the new neural models have been implemented by deriving new Python classes from the Keras library [1] using Tensorflow as backend.

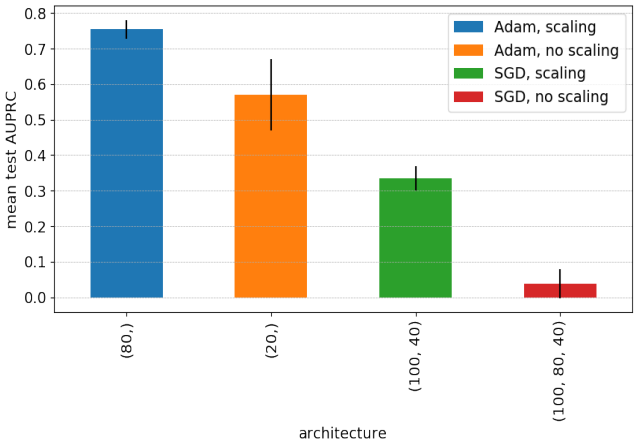


Figure 2: *MiMiS-Net* cross-validation results on the training set, using Adam and SGD optimization algorithms with and without feature normalization. In abscissa the number of hidden neurons for each layer of the selected best models is reported. The vertical lines represent the standard deviation across folds.

3.2 *MiMiS-Net* results

At first we trained and test the state-of-the-art method *hyperSMURF* on the Mendelian data set, obtaining an AUPRC = 0.911 and an AUROC = 0.999. The best "vanilla" neural model, i.e. a neural network that does not adopt any imbalance-aware learning strategy, achieved an AUPRC = 0.078 and an AUROC = 0.968. This is not so surprising since a previous work clearly showed that imbalance-unaware strategies are not able to obtain good results on this challenging learning task [10].

The proposed *MiMiS-Net* imbalance-aware method, by setting the batch size $n = 5000$, corresponding to a probability $P(X_n \geq 1) \simeq 0.85$ of drawing at least one positive example in the mini-batch in the training set (eq. 4) led to significantly better results than the vanilla Neural Network (Fig. 2). On the test set we obtained an AUPRC = 0.794 and an AUROC = 0.973, significantly lower than that obtained by *hyperSMURF* but an order of magnitude larger than that obtained by the vanilla neural model. Fig. 2 shows that *Adam* optimization achieves significantly better results than *SGD* and as expected feature standardization is necessary to improve performances. Nevertheless, looking at Fig. 3 (a), we can observe a certain overfitting of *MiMiS-Net* and for this reason we applied dropout techniques [13] to try to avoid this effect. Results show that *MiMiS-Net* with dropout reduces overfitting (Fig. 3 (b)) and achieves significantly better results on the test set (AUPRC = 0.879).

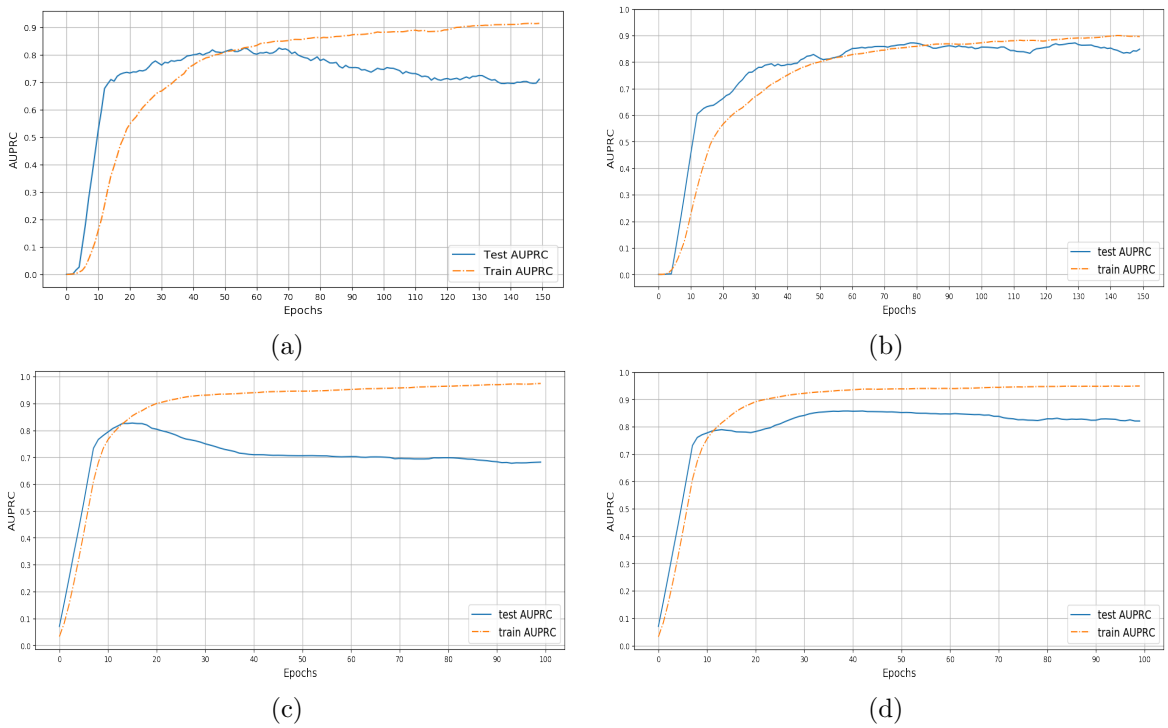


Figure 3: *MiMiS-Net* and *MiBa-Net* training and test AUPRC across epochs. Horizontal axis: epochs; vertical axis: AUPRC. Orange and blue lines represent respectively test and train AUPRC results. (a) *MiMiS-Net* (b) *MiMiS-Net* with dropout (c) *MiBa-Net* with dropout (d) *MiBa-Net* with dropout and Max norm regularization.

3.3 *MiBa-Net* results

Results with *MiBa-Net* show that also this neural imbalance-aware technique can boost pathogenic Mendelian variants detection. Indeed *MiBa-Net* with dropout obtains on the test set an AUPRC = 0.674, but with a serious overfitting towards the training set (Fig. 3 (c)). Recalling that regularization through maximization of the norm has been shown to work nicely when paired with dropout [13], we applied jointly dropout and Maxnorm regularization techniques, thus reducing overfitting (Fig. 3 (d)) and achieving a test set AUPRC = 0.835.

Even if we achieved results close to that obtained by the state-of-the-art method *hyperSMURF*, we tried to further improve performances by analyzing the correlation between the 26 features associated with the genomic variants. By systematically applying the Pearson correlation between each pair of features we individuated sets of highly correlated features, and removed accordingly 5 of them and then we retrained both *MiMiS-Net* and *MiBa-Net* with the reduced set of 21 features using dropout and regularization. Results show a further significant enhancement of the performances (Fig. 4), with AUPRC values even better than those achieved by the state-of-the-art *hyperSMURF* method.

4 Conclusion

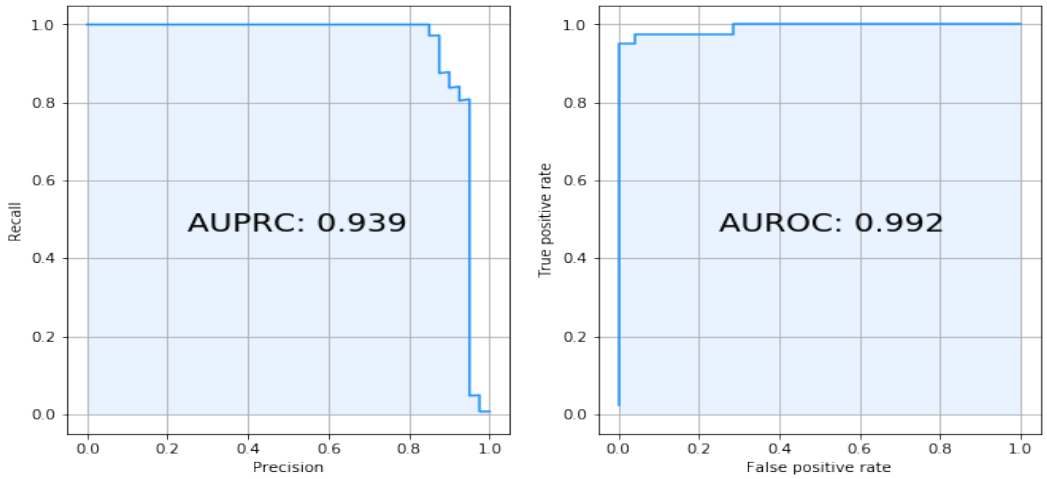
Several machine learning methods have been recently proposed in literature for the detection of pathogenic genomic variants, associated with several diseases ranging from genetic disorders to cancer. We showed that in the case of the detection of rare SNV mutations in non-coding genome causative of Mendelian diseases, imbalance-aware neural models based on mini-batch sampling techniques (*MiBa-Net*) and on the enlargement of the mini-batch (*MiMiS-Net*), we can significantly improve results obtained with imbalance-unaware "vanilla" neural models. In particular by using deep learning techniques together with imbalance-aware methods we can achieve results at least comparable with state-of-the-art results. Finally we observe that in the context of Mendelian diseases the best results have been obtained with relatively simple neural models with one or two hidden layers and some tens on hidden neurons, while state-of-the-art models used ensembles or hyper-ensemble of learning machines, characterized by a significantly larger complexity and training time.

References

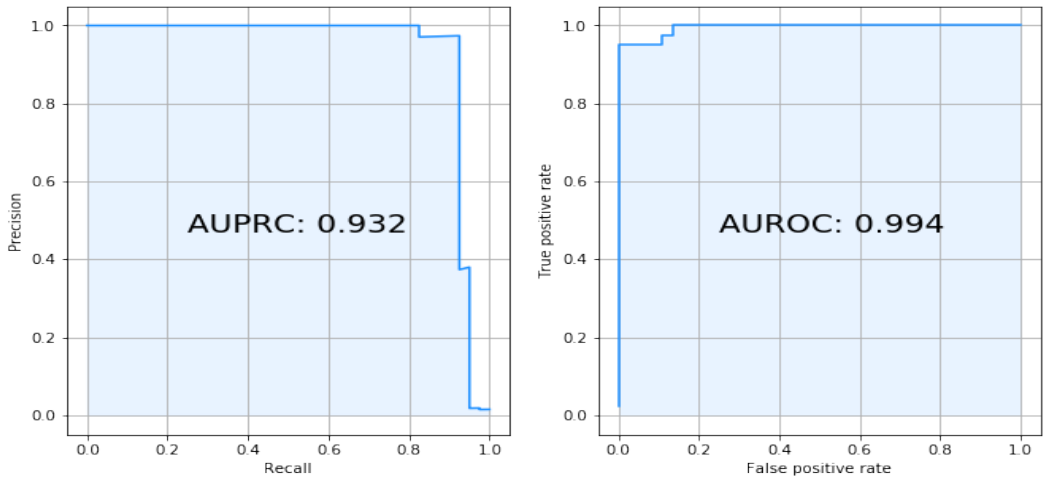
[1] Keras: the python deep learning library. `https://keras.io/`, 2018. Online; accessed 11 November 2018.

[2] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[3] S.L. Edwards, J. Beesley, J.D. French, and A.M. Dunning. Beyond gwas: illuminating the dark road from association to function. *American Journal of Human Genetics*, 93:779–797, 2013.



(a)



(b)

Figure 4: Precision Recall and ROC curves on the test set obtained with the best *MiMiS-Net* and *MiBa-Net* models using feature decorrelation, dropout and regularization techniques. (a) *MiMiS-Net* (b) *MiBa-Net*.

[4] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *JMLR W&CP: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, April 2011.

[5] H. He and E. Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[6] Ionita-Laza et al. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48(2):214–20, Feb 2016.

[7] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

[8] M. Kircher et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, Mar 2014.

[9] G. Ritchie, I. Dunham, E. Zeggini, and P. Flicek. Functional annotation of noncoding sequence variants. *Nature Methods*, 11(3):294–296, Mar 2014.

[10] M. Schubach, M. Re, PN Robinson, and G. Valentini. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Scientific Reports*, 7(2959), 2017.

[11] M. Schubach, M. Re, PN Robinson, and G. Valentini. Variant relevance prediction in extremely imbalanced training sets. *F1000Research*, 6(ISCB Comm J)(1392), 2017.

[12] D. Smedley et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *American Journal of Human Genetics*, 99(3):595–606, 2016.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[14] A. Telenti, C. Lippert, P. Chang, and M. DePristo. Deep learning of genomic variation and regulatory network data. *Human Molecular Genetics*, 27(R1):R63–R71, 2018.

[15] Y. Yang et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine*, 369:1502–1511, 2013.

[16] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, August 2015.