Deep neural architectures for highly imbalanced data in bioinformatics

Leandro A. Bugnon, Cristian Yones, Diego H. Milone, and Georgina Stegmayer

Research Institute for Signals, Systems and Computational Intelligence (sinc(*i*)), FICH-UNL, CONICET, Argentina.
lbugnon@sinc.unl.edu.ar

Extended abstract from *Deep neural architectures for highly imbalanced data in bioinformatics*, L. A. Bugnon, C. Yones, D. H. Milone, G. Stegmayer, (to appear in) IEEE Transactions on Neural Networks and Learning Systems (2019), doi 10.1109/TNNLS.2019.2914471

Keywords: Bioinformatics · Pre-miRNA classification · Deep neural architectures · High class imbalance.

The classification task with imbalanced data has been largely recognized as an important issue in machine learning [3, 4, 8]. Most machine learning algorithms work well with balanced datasets, but with imbalanced datasets supervised classifiers tend to be biased towards the majority class and have a very low performance on the minority one. This is of particular importance in bioinformatics, including studies such as diagnosis based on gene expression data, protein function classification, activity prediction of drug molecules and recognition of precursor microRNAs (pre-miRNAs) [5]. For example, in a real-life scenario, the number of known pre-miRNAs can be in an imbalance ratio (IR) of 1:1,000 (1 positive class sample per 1,000 negative class samples).

In [7] a deep belief neural network (deepBN) for identifying pre-miRNA sequences was proposed. This model has an unsupervised stage with hidden layers pre-trained as restricted Boltzmann machines, followed by a supervised tuning of the network. The unsupervised part was found to improve classification as deep low-level features are obtained, not being affected by class imbalance. In [6] a deep architecture of self-organizing maps (deepSOM) was proposed to overcome the problem of having very few positive class samples and a very large negative class. This model is composed of several layers of hidden SOMs, where each inner SOM discards less probable candidates to pre-miRNAs. The deepSOM model, however, has low precision because a very large number of false positive sequences remain at the last level.

In this work, we present two new variants to the deepSOM model: the deep elastic SOM (deSOM) and the deep ensemble elastic SOM (deeSOM), which overcome the mentioned issues. In deSOM the number of deep levels not only grows automatically, but also the size of each layer is expanded adaptively according to the data at each level, thus pre-miRNA neurons can be re-organized in a larger space. Several deep layers are added with this self size-adjusting method until only known pre-miRNA samples remain at the last layer. The deeSOM uses an ensemble strategy at the beginning of the network to mitigate the high class imbalance (Figure 1). Several parallel SOMs are used at the initial levels and data is split among them, preserving the positive class samples and dividing the remaining ones, thus reducing the imbalance at each SOM of the ensemble. This is, with Q_ℓ SOMs at the ℓ level, the corresponding IR will be reduced Q_1 times. This way, each SOM models just a fraction of the unlabeled space. Another particular feature of these models is that a ranked set of candidates can be obtained by checking the neurons of the next-to-last map and going back to each previous map, until a desired number of candidates is obtained. This is very relevant to choose the best candidates for further wet-lab experiments, which are expensive.

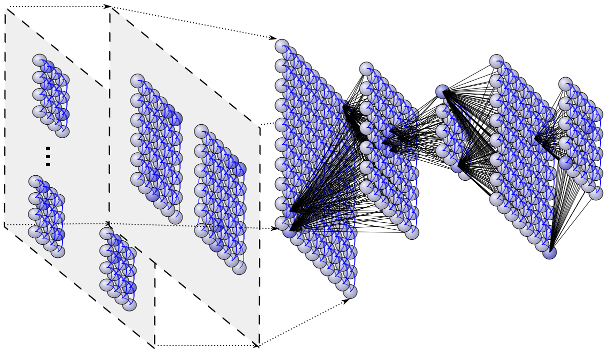

Fig. 1. The deeSOM architecture. The first layers are SOM-based ensembles.

We analyze and compare these very recent deep neural network approaches to deal with imbalanced data in the context of pre-miRNAs classification. For the comparisons we have created a number of datasets of varying IRs, ranging from 1:100 to 1:2000, using already available public data [2]. This provides a positive set with all well-known pre-miRNAs in miRBase v19 [9] and a negative set including random sequences from the genomes of a set of animals and plants[1]. For fair comparisons with state-of-the-art classifiers, we have used the 28 features originally provided in [2].

The aim of the comparisons is to analyze the classifiers robustness regarding how each deep neural model is able to manage the high imbalance by itself. We compared the deep neural architectures versus classical classifiers such as support vector machines (SVM) and multilayer perceptrons (MLP) [10]. For each model tested, a stratified 4-fold cross validation procedure has been used, giving reliable estimates of classification performance. This is assessed for each model by the harmonic mean of sensitivity and precision (F_1). In order to statistically evaluate the differences between classifiers, that is, to detect differences in methods across multiple imbalanced datasets, a Friedman rank test at significance level $\alpha = 0.05$ is carried out for F_1. After that, the Nemenyi test are used as a post-hoc test in order to show which methods are significantly different from each other according to the mean rank differences of the groups [1].

An extensive analysis of results for the models with different IR and several configurations was performed. To summarize, the statistical analysis and the global behavior of the approaches are shown in Figure 2. This figure includes the F_1 score obtained by all methods in each dataset, and for each IR. From the figure it can be easily seen

[1] Source code and data are freely available at: https://sourceforge.net/projects/sourcesinc/files/miRimbal

how all methods decrease performance as imbalance increases. However, three kind of behavior are detected: the very poor performance of SVM and MLP, the deep SOMs topologies in the middle, and the best performance of deepBN. The difference between the groups of classifiers is statistically significant. It is worth to note that SOM-based approaches get the lower computational cost for different imbalances and dataset sizes.

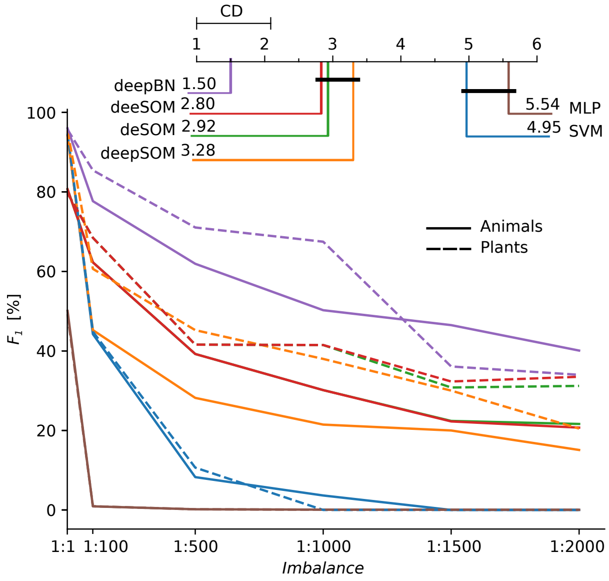

Fig. 2. F_1 score evolution for the deep versus classical approaches for different levels of IR, in animal and plant datasets. Critical difference (CD) diagram for Nemenyi tests is shown above the curves.

In conclusion, we have provided a comparative assessment of recent deep neural approaches for dealing with a highly imbalanced data problem in bioinformatics: the classification of pre-miRNAs. Moreover, two novel deep SOM topologies capable of handling large class imbalance have been presented. The models have been compared in a controlled benchmark framework and several classification tasks involving many genomes and increasing IR, much larger than commonly published IRs. The comparative results obtained have shown that the deep learning models, including unsupervised training stages, were the most capable of maintaining good performance rates, even at increasing IRs up to a very high imbalance.

References

1. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**(1), 1–30 (2006)

2. Gudy, A., Szczeniak, M., Sikora, M., Makalowska, I.: HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. BMC Bioinformatics **14**(1), 83+ (2013)
3. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**, 1263–1284 (2009)
4. Lin, M., Tang, K., Yao, X.: Dynamic sampling approach to training neural networks for multiclass imbalance classification. IEEE Transactions on Neural Networks and Learning Systems **24**(4), 647–660 (2013)
5. Stegmayer, G., Di Persia, L.E., Rubiolo, M., Gerard, M., Pividori, M., Yones, C., Bugnon, L. A., Rodriguez, T., Raad, J. and Milone, D. H.: Predicting novel microRNA: a comprehensive comparison of machine learning approaches. Briefings in Bioinformatics (2018)
6. Stegmayer, G., Yones, C., Kamenetzky, L. and Milone, D.H. : High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM. IEEE/ACM Transactions on Computational Biology and Bioinformatics **14**(6), 1316–1326 (2017)
7. Thomas, J., Thomas, S. and Sael, L.: DP-miRNA: An Improved Prediction of precursor microRNA using Deep Learning Model. IEEE Int. Conf. on Big Data and Smart Computing **1**(1), 96–99 (2017)
8. Wang, S., Yao, X.: Multiclass imbalance problems: Analysis and potential solutions. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **42**(4), 1119–1130 (Aug 2012)
9. Yones, C., Stegmayer, G., Kamenetzky, L. and Milone, D.H.: miRNAfe: a comprehensive tool for feature extraction in microRNA prediction. BioSystems **238**, 1–5 (2015)
10. Zhao, T., Zhang, N., Zhang, Y., Ren, J., Xu, P., Liu, Z., Cheng, L., Hu, Y.: A novel method to identify pre-microrna in various species knowledge base on various species. Journal of Biomedical Semantics **8**(30), 1679–1688 (2017)