Extreme Learning Machine prediction under high class imbalance in bioinformatics

T. Rodriguez, L.E. Di Persia, D.H. Milone, G. Stegmayer

Abstract

Class imbalance in machine learning is when there are significantly fewer training instances of one class in comparison to another one. In bioinformatics, there is such a problem in the computational prediction of novel microRNA (miRNAs) within a full genome. The well-known precursors miRNA (pre-miRNA) are usually only a few in comparison to the hundreds of thousands of potential candidates, which makes this task a high class imbalance classification problem. It is well-known that high class imbalance usually affects any classical supervised machine learning classifier. Thus the imbalance must be explicitly considered. Extreme Learning Machine (ELM) is a supervised artificial neural network model that has gained interest in the last years because of its high learning rate and performance. In this work, we propose a novel approach to overcome the high class imbalance in pre-miRNAs prediction data in which ELMs are used for predicting good candidates to pre-miRNA, without needing balanced data sets. Real datasets were used for validation of the proposal with several class imbalance levels. The results obtained showed the superiority of the ELM approach against very recent state-of-the-art methods in the same experimental conditions.

Index Terms

Extreme learning machines, classification, high class imbalance, microRNA.

1. Introduction

The class imbalance problem has been largely recognized as an important issue in machine learning [1], [2] and, more recently, in the context of big data mining [3], [4]. The problem occurs when there are significantly fewer training instances of one class in comparison to another one. Most of the cost functions tipically used in machine learning do not work well with imbalanced data sets, where a supervised classifier can produce a model that tends to be biased towards the majority class, having a very low performance on the minority one. Although many proposals have been published on supervised classifiers for certain low levels of imbalanced data sets [5], [6], classification of high class imbalanced data where one class is significantly under-represented remains among the current challenges in the development of prediction models.

This is of particular importance in bioinformatics today, where there are large biological datasets with this type of imbalanced data. For example, in the computational prediction of microRNAs (miRNAs) [7], where there are only dozens or hundreds (it depends on the organism under study) of well-known miRNAs, versus thousand hundreds of unknown/unlabeled sequences in the rest of the genome. Many of these sequences are really negative class and among which there can be hidden candidates to novel miRNAs.

This new type of small RNA molecules, present in both animals and plants, can determine the genetic expression of cells and influence the state of the tissues [8]. Many studies have shown that miRNAs are implied, for example, in cancer progression [9] as well as in viral infection processes [10] and parasites development [11]. Given their role in promoting or inhibiting certain diseases and infections, the discovery of new miRNA precursors (pre-miRNAs) is of high interest. Existing experimental techniques have proven to be inefficient and costly for this task, thus computational methods play an important role nowadays in the identification of new miRNAs [12], [13].

For this challenging task, the earliest proposals based on machine learning for pre-miRNA identification have used simple representations to extract the main structural features of known pre-miRNAs [14], [15], [16]. After the feature extraction step, a binary

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

classifier is trained in order to classify sequences. Support vector machine (SVM) is the learning algorithm that has been most widely applied to solve this problem, using as positive sets the genuine pre-miRNA and artificially defining negative sets of hairpins [17], [18], [19], [20], [21]. Such classification models were expected to perform well in predicting novel pre-miRNAs from unseen sequences after using the well-known positive labeled examples for training. However, a recent study has stated that most of the existing machine learning approaches cannot provide reliable predictive performances on independent testing data sets because the positive training sets requiring some sort of balancing approach [22]. Given the very large number of candidates to be analyzed in a real genome (hundreds of thousands sequences) and the strong class imbalance new strategies must be proposed [23].

In this work we present a novel approach for dealing with the high imbalance problem in pre-miRNA prediction based on Extreme Learning Machine (ELM), a single layer feedforward network with random weights. This kind of classifier has shown great capability in binary task as well as multiclass problems [24]. The main hypothesis of this work is that the ELM classifier, as observed in other classification tasks, has some kind of *intrinsic* robustness to class imbalance, being able to produce good results even with severe imbalance and without need for any additional strategy. This has been tested in experiments with small and synthetic datasets [25], as well as in some biological datasets [26]. The proposed approach has been tested with two different organisms, using large and varied strongly imbalanced datasets in 10-fold cross-validation tests.

This paper is organized as follows. Section 2 explains the ELM architecture and training algorithm in detail. Section 3 presents the data sets used in this study, the experimental setup and performance measures. Section 4 shows the results obtained and their discussion. Finally, the conclusions of this work can be found in Section 5.

2. ELMs for high class imbalanced biological data

Let the training set given by N samples be defined by

$$D = \left\{(\mathbf{x}_j, t_j) : \mathbf{x}_j \in \mathbb{R}^d, t_j \in \{-1, 1\}, j = 1, \ldots, N\right\},$$

where $\mathbf{x}_j$ is a $d \times 1$ input vector and t_j is a target class label.

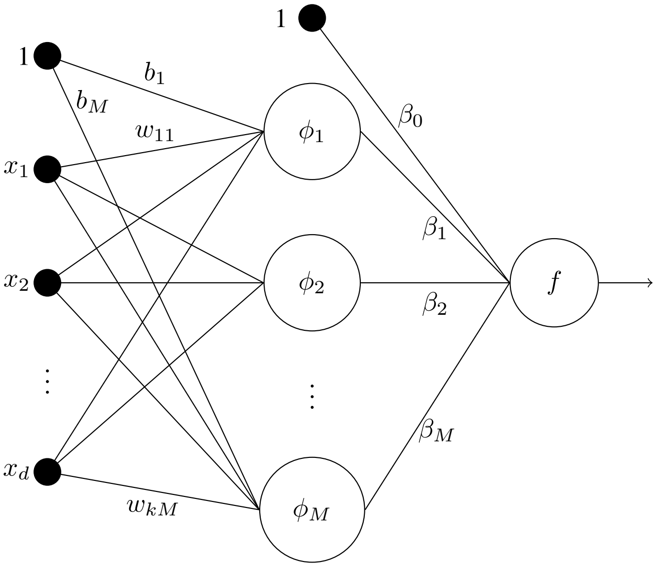

Figure 1. Single layer feedforward network. Func tions ϕ_j depend on the parameters $\{w_{ij}\}_{i=1}^{k}$ and b_j

Let us consider a single layer feedforward network with M neurons in the hidden layer, as shown in Figure 1. Its output is given by the function

$$f(\mathbf{x}_j; \boldsymbol{\theta}) = \beta_0 + \sum_{i=1}^{M} \beta_i \phi(\mathbf{x}_j, \mathbf{w}_i, b_i),$$

where $\boldsymbol{\theta} = (\mathbf{w}_i, b_i, \beta_i)$ is the parameter vector and $\phi(\mathbf{x}, \mathbf{w}, b)$ is a given activation function. If ϕ is the sigmoid function, then vectors $\{\mathbf{w}_i\}_{i=1}^{M} \subset \mathbb{R}^d$ represent the weights for the inputs in each neuron i. Besides, $b_i \in \mathbb{R}$ is the threshold for each unit. In this way, we can interpret each neuron as defining an hyperplane, with form $\mathbf{x}^T\mathbf{w}_i + b_i = 0$. Instead, if ϕ is radial function, $\mathbf{w}_i$ is considered as a centroid, while b_i weight the distance between inputs and centroids. Generally, this kind of function describes hyperspheres. In this way, a radial basis function neural network can define complex regions as the union of local hyperspheres.

In standard neural networks, $\boldsymbol{\theta}$ is found through backpropagation learning [27], an algorithm that aims minimizing the squared error

$$\mathcal{E}\left(\boldsymbol{\theta}\right) = \sum_{j=1}^{N} \left(f(\mathbf{x}_j; \boldsymbol{\theta}) - t_j\right)^2.$$

In [28], authors proposed a new strategy of learning. Unlike backpropagation scheme, the weights (or parameters) $\mathbf{w}_i$ and b_i are randomly defined. In this way, only parameters β_i are estimated. Accordingly, optimizing the vector $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2 \ldots \beta_M]^T$ is equivalent to solve a minimum square problem. This strategy significantly improves the training time at the same time that guarantees that, with a large enough number of neurons, this scheme can achieve good performance, even with a very simple architecture.

Indeed, considering this simplification we can now write the function f in matricial form as $f(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})\boldsymbol{\beta}^*$ where

$$h(\mathbf{x}) = [\phi(\mathbf{x},\mathbf{w}_1,b_1)\ldots\phi(\mathbf{x},\mathbf{w}_M,b_M)]\,.$$

Using the set D we can construct the error function

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta}) &= \sum_{j=1}^{N}\left(f(\mathbf{x}_j;\boldsymbol{\theta}) - t_j\right)^2,\\ &= \sum_{j=1}^{N}\left(\mathrm{h}(\mathbf{x}_j)\boldsymbol{\beta} - t_j\right)^2,\\ &= \|H\boldsymbol{\beta} - \mathbf{t}\|^2\,,\end{aligned}$$

where

$$H = \begin{bmatrix} \phi(\mathbf{x}_1,\mathbf{w}_1,b_1) & \ldots & \phi(\mathbf{x}_1,\mathbf{w}_M,b_M) \\ \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_N,\mathbf{w}_1,b_1) & \ldots & \phi(\mathbf{x}_N,\mathbf{w}_M,b_M) \end{bmatrix}_{N\times M}$$

and the vector $\mathbf{t}$ is

$$\mathbf{t} = [t_1,\ldots,t_N]^T\,.$$

Then we can now find the output weights

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\,\|H\boldsymbol{\beta} - \mathbf{t}\|^2\,,$$

or

$$\boldsymbol{\beta}^* = H^\dagger\mathbf{t}.$$

Here, $H^\dagger$ is the pseudo-inverse of H, usually calculated as

$$\boldsymbol{\beta}^* = \left(H^TH\right)^{-1}H^T\mathbf{t},$$

when the number of training samples N is very large with respect to the number of hidden neuron units M.

There are several choices for the activation function ϕ. In this work we use both, the gaussian activation function, defined as

$$\phi(\mathbf{x}_j,\mathbf{w}_i,b_i) \doteq e^{b_i\|\mathbf{x}_j - \mathbf{w}_i\|}$$

which will be called ELM_{RBF} in the experiments and the classical sigmoid function defined as

$$\phi(\mathbf{x}_j,\hat{\mathbf{w}}_i,\hat{b}_i) \doteq \frac{1}{1+e^{\mathbf{x_j}^T\hat{\mathbf{w}}_i+\hat{b}_i}}$$

which will be called ELM_{SIG} from now on.

Table 1. Characteristics of the high class imbalanced biological data sets used in the experiments.

Name	Positive	Negative	IR
Virus	237	839	3.54
A. thaliana	231	28,359	122.77

3. Materials and experimental methods

This section describes the datasets used, the experimental setup and the measures for performance evaluation.

3.1. High class imbalanced datasets

The characteristics of the biological data sets used in the experiments are shown in Table 1, as in [29]. For each set of data, the number of samples for each class is reported in the second and third column, respectively. These two data sets have been selected for the comparative study because they have two extreme imbalance ratios (IR): very low and very high. They include all well-known pre-miRNAs from the most studied model specie *Arabidopsis thaliana*, and also twenty nine virus. Positive and negatives sequences from the analyzed species were gathered to form complete datasets that try to imitate the corresponding pre-miRNA classification problem in real conditions For the positive class, all well-known pre-miRNAs deposited miRBase v17 [30] are used as positive samples (except those sequences lacking experimental confirmation). Negative sets were created as sequence extracted from the corresponding genomes under analysis, where sequences start positions were randomly selected and end positions were calculated so that the sequence length distribution in the resulting negative dataset is the same as in the corresponding positive one, as described in [29].

Class imbalance is reported in the fourth column of the tables, for each data set evaluated. It has been defined as the ratio of the number of negative to the number of positive samples. It can be seen from the table that two extreme imbalance situations have been taken into account, from quite low to very high class imbalance.

Selecting an informative feature set is very important for the pre-miRNA prediction problem. Most commonly used feature sets contain information about sequence, topology and structure [31]. The earliest machine learning approaches [17] proposed features, named triplets, computed from the sequence itself.

miPred [32] was the first method that proposed a representative feature set that has shown great discriminative power and that has been adopted by many other current methods [13], [29].

We have used those features as in [29]: triplets, maximal length of the amino acid string, cumulative size of internal loops found in the secondary structure, percentage of low complexity regions detected in the sequence and predicted thermodynamic and statistical properties. In this way each pre-miRNA has been represented as a feature vector with 28 real components.

3.2. Experimental setup

In this work, we have explored five different approaches. Two of them are based on Extreme Learning Machines, with the activation function previously defined and denoted as $\mathrm{ELM}_{SIG}, \mathrm{ELM}_{RBF}$ in the tables. Both depend of one parameter, the number of neurons in the hidden layer M. Also, as was used in TripletSVM [17], we have built a predictor based on Suppor Vector Machine (SVM). In this case, was used a radial basis function

$$K(x_i, x_j) = e^{-s\|x_i - x_j\|^2},$$

with slacking limit C. Such model has s, C as hyperparameters. For this model we present two variants, one trained with the original training set as ELM training and another one that includes an oversampling approach, which will be described in detail below. Finally, we have included the results with HuntMi [29] tool, an classifier based on Random Forest with the same kind of oversampling mentioned.

For SVM, a classical machine learning strategy for balancing imbalanced data sets has been evaluated as well. The synthetic minority oversampling technique (SMOTE) [33] is an approach for oversampling the positive class (in general the minority class). It is limited to the strict assumption that the local space between any two positive instances is positive. The method produces artificial samples as convex combinations of each positive sample and one of its nearest neighbors. This is repeated for all positive samples, the number of times neccesary to produce a balanced set. SMOTE is the most used technique nowadays in pre-miRNA classifiers [6]. In summary, we will compare the behaviour of five different classification methods: SVM, SVM+SMOTE, HuntMi (RF+SMOTE), ELM_{SIG}, and ELM_{RBF}.

For each training set an independent 10-fold cross validation (CV) has been performed, giving reliable estimates of the classification performance. In all classification experiments, the distributions of classes in the testing set is the same as for the entire datasets. The performance in each experiment is reported as the average values on the 10 folds using the test partitions only.

A crucial issue to achieve a good performance in any classification problem is a right choice of hyperparameters. In our experiments this selection was taken through a grid search. Once the search range was defined for each hyperparameter, for each combination of them the performance was estimated by the average of an inner 3-fold CV defined within each training set f. That is, for each one of the 10 folds, the training data was used in a 3-fold CV to select hyperparameters. Then, all training data in that fold was used to train a classifier with the selected hyperparameters. This trained classifier was used on the corresponding test set. For all classifiers, we also optimized the classification threshold μ, where if $f(\mathbf{x}; \boldsymbol{\theta}) \geq \mu$ then x is in minority class and belongs to the majority class in the other case. This was done jointly with the optimization of hyperparameters, in the inner 3-fold CV. The cost function used in the optimization was G_m (see Eq. 6 below).

When SMOTE was used, it was applied in each of the training stages, namely, on the 3-fold CV hyperparameter search and on the principal CV. In this way, we could approximate the performance of applying SMOTE in the real world. While SVM required a 2-dimensional grid search (with high computational cost) both versions of ELM just needed to select the number of hidden neurons.

Futhermore, we have analized in detail the performance of the ELM as a function of the imbalance ratio. To be able to analyze in more detail the capabilities of the proposed method in more controlled imbalance ratio situations, we built new data sets with varying IR from the original *A. thaliana*. To obtain several datasets with a certain desired imbalance ratio (IR), if the original set had n_+ positive instances then we selected $\mathrm{IR} \times n_+$ random negative instances. Whenever, $\mathrm{IR} \times n_+$ is greater than n_-, the number of majority instances, we choose $\frac{n_-}{\mathrm{IR}}$ random positive instances and keep all the negatives instances.

3.3. Model performance

The prediction quality of the model was assessed by the following classical classification measures: sensitivity (s^+), specificity (s^-), precision(p), harmonic mean of sensitivity and precision ($F1$), accuracy (Acc), geometric mean (G_m) of classification sensitivity and specificity, and geometric mean G of sensitivity and precision. These measures are defined as:

$$\begin{aligned} s^{+} &= \frac{TP}{TP+FN}, && (1)\\ s^{-} &= \frac{TN}{TN+FP}, && (2)\\ p &= \frac{TP}{TP+FP}, && (3)\\ F1 &= 2\,\frac{s^{+}\times p}{s^{+}+p}, && (4)\\ Acc &= \frac{TP+TN}{TP+TN+FP+FN}, && (5)\\ G_m &= \sqrt{s^{+}\times s^{-}}, && (6)\\ G &= \sqrt{s^{+}\times p}, && (7) \end{aligned}$$

where TP, TN, FP and FN are the number of true positive, true negative, false positive and false negative classifications, respectively.

The selection of G_m as measure performance is conventional in imbalanced classification problems, because it presents less bias to the majority class than accuracy. Also, considering the prediction problem under study, it is very important to keep high precision, thus we include $F1$ and G that particularly take this into account in the calculus.

4. Results and discussion

4.1. Comparison with other prediction methods

This section presents the results of the experiments to analyze in detail the behavior of ELMs for high class imbalance data sets in comparison to state-of-the-art pre-miRNA prediction methods.

Table 2 shows the results for the Virus and *Arabidopsis thaliana* data sets (detailed in Table 1). Average results are reported for the test data in 10-fold CV. The first column shows the classifiers. From second to last column, average s^{+}, p, s^{-}, $F1$, Acc, G_m and G are reported. This table clearly shows that very high classification rates are achieved by ELM_{RBF} in all cases. ELM_{SIG} seems to perform worse than ELM_{RBF} and in general also with respect to the other classifiers.

In the data set with low class imbalance, the virus data set, the ELM performance for recognizing pre-miRNAs is very high (more than 90%). Here, $F1$ and Acc are the highest. G_m and G, which are indexes more adequate to evaluate this imbalanced data, are also the highest. Furthermore, the best precision achieved is higher than 90% for ELM_{RBF}.

It must be highlighted that for the most imbalanced data set (*A. thaliana*), the G_m for ELM_{RBF}, is higher than 97%. The G value is the highest as well for ELM_{RBF}, in comparison to state-of-the-art methods. In this particular data set, the most interesting one from the imbalance level point of view, the highest p and the highest G are achieved by the proposed ELM_{RBF} approach. This is a very hard to achieve result for the state-of-art methods, even with SMOTE. In comparison to the first and original SVM classifier, there is more than 10% difference in precision. Our proposal is also better than the other more recent works evaluated. This supports our initial hypothesis that ELMs can be adequate and are quite suited for the large class imbalance problem of pre-miRNA prediction, without even needing any balancing scheme.

Regarding SMOTE, it seems that it produces no improvements for SVM and even more it has a negative effect, both in performance as in computational cost. This may be explained by the fact that SVM uses support vectors to define the decision boundary, and thus SMOTE will have effect only if it adds samples near this boundary (where they can became support vectors). But in a high class imbalance context, the main hypothesis of SMOTE (that the space between two near positive examples corresponds to positive class) may not be true, mainly if the boundary is complex and the two positive examples are not near enough. Therefore, the added support vectors, assumed to belong to one class, will be located in a region of the other class.

A Friedman's test on G_m, over the 10 partitions of the two datasets, gives a p-value of 0.014 (the null hypothesis can be rejected), that is, the methods have significative differences in performance. With another statistical test based in rankings [34], we obtained a critical difference (CD) of 1.23 between methods (see Figure 2). The proposed ELM_{RBF} is better positioned than ELM_{SIG} and SVM with SMOTE. In an analogous process for precision measure (p), the null hypothesis is rejected again. The mean ranking can be appreciated in Figure 3. ELM_{RBF} is more precise than HuntMi, ELM_{SIG} and SVM+SMOTE.

4.2. Performance varying imbalance ratio

In this subsection we are interested in evaluating the performance of the best classifier as the imbalance grows. Because any real dataset has a fixed imbalance, to get consistent results, it was necessary to generate synthetic datasets. We used the *A. thaliana* dataset because allows to achieve a larger range of imbalances without compromising the number of positive

Table 2. Classification results for pre-miRNA prediction in high class imbalanced data sets. Best value in bold.

Virus dataset

Classifier	s^+	p	s^-	$F1$	Acc	G_m	G
SVM [17]	96.60 (4.16)	91.32 (4.96)	97.37 (1.57)	93.83 (3.95)	97.20 (1.79)	96.96 (2.51)	93.89 (3.92)
SVM+SMOTE	95.96 (3.24)	90.33 (4.48)	97.07 (1.45)	93.02 (3.42)	96.82 (1.60)	96.50 (2.08)	93.08 (3.40)
HuntMi [29]	96.38 (2.84)	87.55 (3.11)	96.11 (1.10)	91.71 (2.21)	96.15 (1.09)	96.23 (1.51)	91.84 (2.20)
ELM_{SIG}	95.74 (4.37)	86.31 (4.79)	95.63 (1.69)	90.66 (2.92)	95.65 (1.38)	95.66 (2.12)	90.84 (2.90)
ELM_{RBF}	**96.81 (3.36)**	**92.68 (5.47)**	**97.78 (1.69)**	**94.64 (3.92)**	**97.57 (1.79)**	**97.28 (2.22)**	**94.70 (3.89)**

A. thaliana dataset

Classifier	s^+	p	s^-	$F1$	Acc	G_m	G
SVM [17]	95.65 (2.51)	45.06 (15.1)	98.87 (0.62)	59.83 (13.2)	98.84 (0.61)	97.24 (1.07)	64.76 (10.3)
SVM+SMOTE	95.22 (3.94)	44.63 (12.2)	98.91 (0.53)	59.72 (10.7)	98.88 (0.51)	97.02 (1.84)	64.53 (8.14)
HuntMi [29]	96.09 (3.21)	54.34 (12.3)	99.27 (0.35)	68.58 (9.78)	99.25 (0.34)	97.65 (1.56)	71.78 (7.83)
ELM_{SIG}	94.78 (3.27)	46.62 (8.68)	99.07 (0.31)	62.02 (7.18)	99.03 (0.29)	96.89 (1.63)	66.19 (5.67)
ELM_{RBF}	**96.30 (3.41)**	**55.50 (8.84)**	**99.33 (0.29)**	**69.89 (6.96)**	**99.31 (0.27)**	**97.79 (1.63)**	**72.81 (5.28)**

Table 3. Performance for different imbalance ratios (IR) in *A. thaliana* dataset.

$\mathbf{ELM}_{RBF}$

IR	n_+	n_-	s^+	p	s^-	$F1$	Acc	G_m	G
1	231	231	96.09	97.60	97.61	96.81	96.85	96.83	96.83
2	231	462	96.09	98.09	99.02	97.02	98.04	97.52	97.06
5	231	1155	95.87	95.22	99.00	95.48	98.48	97.42	95.51
10	231	2310	95.00	92.73	99.20	93.65	98.82	97.06	93.76
20	231	4620	95.22	88.72	99.35	91.59	99.15	97.24	91.78
50	231	11550	96.30	77.21	99.40	85.40	99.34	97.83	86.08
100	231	23100	95.65	58.00	99.25	71.59	99.22	97.43	74.14
200	141	28359	92.50	48.43	99.48	63.00	99.45	95.91	66.60
500	56	28359	82.73	49.05	99.79	59.60	99.76	90.54	62.57

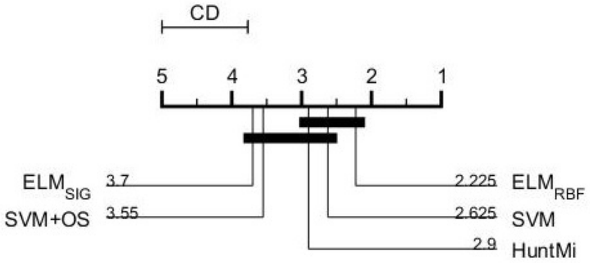

Figure 2. Critical difference diagram for G_m

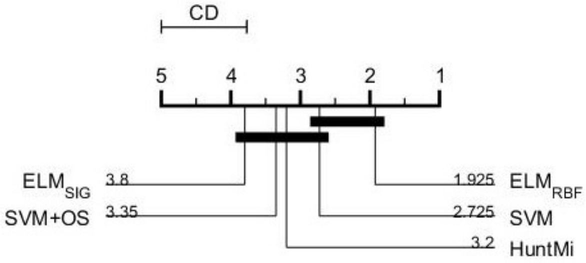

Figure 3. Critical difference diagram for precision

instances.

Table 3 shows the performance of the proposed method under different and increasing levels of imbalance. Because ELM_{RBF} achieved the best results, showing robustness for this task, we only use this version of Extreme Learning Machine. The first column of the table indicates imbalance ratio, which varies approximately exponentially from 1 to 500. The second and third column, respectively, show the number of positive and negative samples. Note that from IR=200, n_+ has decreased because the original dataset has IR=122. Analogously, third column shows the number of negative samples on each set.

It can be clearly noticed here how the classifier based on ELM is capable of handling the increasing imbalance, or better said, it is not hardly affected by it.

The G_m measure keeps a high value above 90%. The G has lower values at higher imbalances, of course, because it is affected by the precision, maintaining however a higher than 60% value. It should be noticed also how the Acc, which is the performance measure mostly reported on published works, is not a reliable performance measure in this case because it seems to show a classifier that is not affected by imbalance, when this is not true and the precision and sensitivity of the method have been affected by the imbalance level on the positive class of interest. In spite of this fact, the precission achieved by the proposed ELM classifier at the highest imbalance levels remains still at acceptable levels and close to those obtained at the previous experiment, which would have been very hard to achieve for the other methods evaluated. For example, $F1$ has fallen to around 60%, which was a level achieved at smaller IR in the other experiments, even if in this case the IR is four times the one already evaluated in the previous tables.

5. Conclusions

In this work we have presented a new and effective approach for the computational classification of miRNAs precursors in the context of a high class imbalance, and without requiring any oversampling strategy. Both, ELM$_{SIG}$ and ELM$_{RBF}$ models proposed, showed a comparable performance with state-of-art classificators. In particular, ELM$_{RBF}$ achieved good precision and the best tradeoff between sensitivity and specificity on datasets with very high imbalance ratios. Moreover, its simple architecture grants flexibility, with promising future in pre-miRNA prediction problems.

Acknowledgements

This work was supported by National Scientific and Technical Research Council [PIP 2013 117] and Agencia Nacional de Promocion Cientifica y Tecnologica (ANPCyT) [PICT 2014 2627].

References

[1] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *Natural Computation, 2008. ICNC '08. Fourth International Conference on*, vol. 4, Oct 2008, pp. 192–201.

[2] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 4, pp. 1119–1130, Aug 2012.

[3] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[4] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data." *Information Sciences*, vol. 275, pp. 314–347, 2014.

[5] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "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*, vol. 42, no. 4, pp. 463–484, July 2012.

[6] R. Blagus and L. Lusa, "Smote for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, p. 106, 2013.

[7] B. Liu, J. Li, and M. Cairns, "Identifying mirnas, targets and functions," *Briefings in Bioinformatics*, vol. 15, no. 1, pp. 1–19, 2014.

[8] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function." *Cell*, vol. 116, pp. 281–297, 2004.

[9] A. Esquela-Kerscher and F. J. Slack, "Oncomirs - microRNAs with a role in cancer," *Nature Reviews Cancer*, vol. 6, no. 1, pp. 259–269, 2006.

[10] C.-H. Lecellier, P. Dunoyer, K. Arar, J. Lehmann-Che, S. Eyquem, C. Himber, A. Saib, and O. Voinnet, "A cellular MicroRNA mediates antiviral defense in human cells," *Science*, vol. 308, no. 5721, pp. 557–560, 2005.

[11] M. Rosenzvit, M. Cucher, L. Kamenetzky, N. Macchiaroli, L. Prada, and F. Camicia, *MicroRNAs in Endoparasites*. Nova Science Publishers, 2013.

[12] L. Li, J. Xu, D. Yang, X. Tan, and H. Wang, "Computational approaches for microRNA studies: a review," *Mamm Genome*, vol. 21, no. 1, pp. 1–12, 2010.

[13] Ivani de ON Lopes and Alexander Schliep and Andre de Carvalho, "The discriminant power of RNA features for pre-miRNA recognition," *BMC Bioinformatics*, vol. 15, no. 1, pp. 124+, 2014.

[14] S. A. Helvik, O. Snove, and P. Saetrom, "Reliable prediction of Drosha processing sites improves microRNA gene prediction." *Bioinformatics*, vol. 23, no. 2, pp. 142–149, 2007.

[15] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features," *Nucleic Acids Research*, vol. 35, no. 1, pp. W339–W344, 2007.

[16] K. Gkirtzou, I. Tsamardinos, P. Tsakalides, and P. Poirazi, "MatureBayes: A probabilistic algorithm for identifying the mature miRNA within novel precursors," *PLOS one*, vol. 5, no. 8, p. e11843, 2010.

[17] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinformatics*, vol. 6, no. 1, p. 310, 2005.

[18] J. Hertel and P. F. Stadler, "Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data," *Bioinformatics*, vol. 22, no. 14, pp. e197–e202, 2006.

[19] T. H. Huang, B. Fan, M. Rothschild, Z. L. Hu, K. Li, and S. H. Zhao, "MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans," *BMC Bioinformatics*, vol. 8, no. 1, pp. 341+, 2007.

[20] J. Ding, S. Zhou, and J. Guan, "MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features," *BMC Bioinformatics*, vol. 11, no. 11, p. S11, 2010.

[21] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, and S. Mavroudi, "YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 5, pp. 1183–1192, 2015.

[22] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and Promising Identification of Human MicroRNAs by Incorporating a High-quality Negative Set," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 11, no. 1, pp. 192–201, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2013.146

[23] S. Dua and P. Chowriappa, Eds., *Data Mining for bioinformatics*. CRC Press, 2012.

[24] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 4, no. 3, pp. 485–495, 2007.

[25] H. Yu, C. Sun, X. Yang, W. Yang, J. Shen, and Y. Qi, "ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data," *Knowledge-Based Systems*, vol. 92, pp. 55–70, 2016.

[26] K. Cheng, Q. Chen, X. Yang, S. Gao, and H. Yu, "Classification of imbalanced bioinformatics data by using boundary movement-based elm," *Bio-medical materials and engineering*, vol. 26, no. s1, pp. S1855–S1862, 2015.

[27] R. Hecht-Nielsen *et al.*, "Theory of the backpropagation neural network." *Neural Networks*, vol. 1, no. Supplement-1, pp. 445–448, 1988.

[28] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 2. IEEE, 2004, pp. 985–990.

[29] A. Gudy, M. Szczeniak, M. Sikora, and I. Makalowska, "HuntMi: an efficient and taxon-specific approach in pre-miRNA identification," *BMC Bioinformatics*, vol. 14, no. 1, pp. 83+, 2013. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-14-83

[30] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, pp. 152–157, 2011.

[31] C. Yones, G. Stegmayer, L. Kamenetzky, and D. Milone, "miRNAfe: a comprehensive tool for feature extraction in microRNA prediction," *BioSystems*, vol. 238, pp. 1–5, 2015.

[32] R. Batuwita and V. Palade, "*microPred*: effective classification of pre-mirnas for human mirna gene prediction," *Bioinformatics*, vol. 25, no. 8, pp. 989–995, 2009.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[34] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.