Discovery of novel pre-miRNAs: unsupervised versus supervised machine learning

G. Stegmayer, D.H. Milone

Institute for Signals, Systems and Computational Intelligence (sinc(i)), FICH-UNL, CONICET, Argentina.

Background:
The computational prediction of novel microRNAs involves identifying nucleotide sequences having the highest chance of being candidates to miRNA precursors (pre-miRNAs). This is a challenge for a machine learning algorithm because well-known pre-miRNAs are just a few in comparison to the hundreds of thousands of candidates. This is a high class-imbalance problem. The classical way of approaching it has been training a binary supervised classifier, using well-known pre-miRNAs from miRBase as positive class and artificially defining a negative class. This has two important drawbacks: i) it is extremely difficult to build a representative set of negative examples; and ii) it is well-known in machine learning that high class-imbalance has a strong influence on standard classifiers.

Results:
We state that a more adequate (and natural) way of approaching this problem would be with unsupervised learning, which does not need the definition of any class at all for training, nor it is influenced by class imbalance. Similar data are simply grouped together in a cluster. After training, positive samples are used to identify the clusters that have the best candidates: those sequences clustered together with them. We have developed an online web-demo[1] to compare the supervised vs. unsupervised approaches for pre-miRNA prediction in two model genomes. The most used and published supervised classifier, a support vector machine (SVM), is compared with two unsupervised methods: k-means (KM) and self-organizing maps (SOM). Well-known pre-miRNAs in miRBase have been used as positive class. Sequences randomly picked have been used as negative class for SVM, as unlabeled for KM and SOM. Different levels of imbalance starting from 1:1 (no imbalance) can be set for 3-fold cross-validation prediction tests in order to obtain true positive rate (*tpr*, sensitivity), true negative rate (*tnr*, specificity), accuracy (*acc*) and the geometric mean (*gm*) between *tpr* and *tnr*. At the highest imbalance cases (see Tables 1 and 2), although the *acc* of SVM seems superior, actually the most realistic measure *gm* (that takes into account both *tpr* and *tnr)* is much higher in all cases for the unsupervised models. This is an important warning on how looking only at the reported accuracy of a model can be really misleading, due to the fact that *acc* does not take into account the sensitivity of a model.

Conclusions:
In most genomes there is a very high class-imbalance between well-known pre-miRNAs and unlabeled sequences that supervised classification models cannot properly handle. We have presented comparison results in favor of unsupervised machine learning as more suited for pre-miRNA prediction. The comparative results show that unsupervised approaches are capable of maintaining good performance rates, while a supervised model quickly deteriorates, when class imbalance increases. Additionally, the unsupervised approach is more naturally suited to an end user that has good knowledge on the pre-miRNAs of the

1 http://fich.unl.edu.ar/sinc/blog/web-demo/mirna-sup-vs-unsup/

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. Stegmayer & D. H. Milone; "Discovery of novel pre-miRNAs: unsupervised versus supervised machine learning"
4th ISCB-LA Bioinformatics Conference, 2016.

genome under study, but has no knowledge regarding the definition of a negative class for training a predictor.

Table 1. Supervised vs. unsupervised approaches comparison for pre-miRNAs prediction in *Homo sapiens*

Class-imbal.	Supervised				Unsupervised							
(1:n)	SVM				KM				SOM			
	tpr	*tnr*	*acc*	*gm*	*tpr*	*tnr*	*acc*	*gm*	*tpr*	*tnr*	*acc*	*gm*
1	100.00	93.82	93.82	96.85	90.00	84.94	84.94	87.34	96.67	84.78	84.79	90.44
5	96.67	98.39	98.39	97.50	96.67	86.80	86.80	91.50	93.33	90.25	90.26	91.66
10	90.00	98.93	98.92	94.26	96.67	84.02	84.03	90.05	100.00	88.96	88.97	94.31
50	63.33	99.61	99.52	79.05	96.67	88.23	88.24	92.28	93.33	87.98	87.98	90.46
100	50.00	99.72	99.70	70.37	100.00	75.09	75.10	86.58	90.00	86.60	86.60	88.13
200	40.00	99.82	99.80	61.99	93.33	84.11	84.11	88.55	96.67	87.84	87.84	92.08

Table 2. Supervised vs. unsupervised approaches comparison for pre-miRNAs prediction in *Arabidopsis thaliana*

Class-imbal.	Supervised				Unsupervised							
(1:n)	SVM				KM				SOM			
	tpr	*tnr*	*acc*	*gm*	*tpr*	*tnr*	*acc*	*gm*	*tpr*	*tnr*	*acc*	*gm*
1	86.67	87.46	87.46	86.13	83.33	74.30	74.31	77.76	86.67	66.71	66.73	74.28
5	86.67	93.26	93.26	89.11	83.33	82.06	82.06	82.06	80.00	85.40	85.39	81.62
10	86.67	94.64	94.64	90.54	86.67	84.07	84.07	84.35	80.00	85.27	85.27	81.87
50	70.00	97.54	97.51	82.24	83.33	78.72	78.73	80.10	83.33	83.94	83.94	82.87
100	66.67	98.41	98.38	80.38	83.33	80.49	80.49	80.93	86.67	81.46	81.46	83.03
200	60.00	98.80	98.76	76.82	86.67	80.45	80.46	82.43	86.67	80.69	80.69	82.58

Supported by: CONICET, MinCyT, UNL.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. Stegmayer & D. H. Milone; "Discovery of novel pre-miRNAs: unsupervised versus supervised machine learning"
4th ISCB-LA Bioinformatics Conference, 2016.