INSTITUTO
DE INFORMÁTICA
UFRGS

Combining Performance and Diversity Measures for Optimizing Classification Ensembles via a Genetic Algorithm in the miRNA-Target Prediction Problem

Gabriel Marangoni Moita

December 2018

Outline

1. Introduction

Outline

1. Introduction
2. Biological Background

Outline

1. Introduction
2. Biological Background
3. Computational Background

Outline

1. Introduction
2. Biological Background
3. Computational Background
4. Related Work

Outline

1. Introduction
2. Biological Background
3. Computational Background
4. Related Work
5. Proposed Solution

Outline

1. Introduction
2. Biological Background
3. Computational Background
4. Related Work
5. Proposed Solution
6. Experimental Results

Outline

1. Introduction
2. Biological Background
3. Computational Background
4. Related Work
5. Proposed Solution
6. Experimental Results
7. Conclusion

Introduction

- miRNA-Target Prediction Problem is a difficult challenge in the molecular biology area, with millions of possible miRNA-mRNA combinations.

Introduction

- miRNA-Target Prediction Problem is a difficult challenge in the molecular biology area, with millions of possible miRNA-mRNA combinations.
- Machine Learning (ML) algorithms were proven promising to train predictive models and better understand miRNA-mRNA interactions and its influence in the metabolism.

Introduction

- miRNA-Target Prediction Problem is a difficult challenge in the molecular biology area, with millions of possible miRNA-mRNA combinations.
- Machine Learning (ML) algorithms were proven promising to train predictive models and better understand miRNA-mRNA interactions and its influence in the metabolism.
- Datasets contains a relatively low number of combinations, and they are mainly functional pairs, further increasing the difficulty.

Introduction

- miRNA-Target Prediction Problem is a difficult challenge in the molecular biology area, with millions of possible miRNA-mRNA combinations.
- Machine Learning (ML) algorithms were proven promising to train predictive models and better understand miRNA-mRNA interactions and its influence in the metabolism.
- Datasets contains a relatively low number of combinations, and they are mainly functional pairs, further increasing the difficulty.
- Genetic Algorithms are already used in this area of study to learn the best heterogeneous ensemble starting from a set of possible classifiers.

Introduction

- miRNA-Target Prediction Problem is a difficult challenge in the molecular biology area, with millions of possible miRNA-mRNA combinations.
- Machine Learning (ML) algorithms were proven promising to train predictive models and better understand miRNA-mRNA interactions and its influence in the metabolism.
- Datasets contains a relatively low number of combinations, and they are mainly functional pairs, further increasing the difficulty.
- Genetic Algorithms are already used in this area of study to learn the best heterogeneous ensemble starting from a set of possible classifiers.
- Combining performance and diversity measures achieves better results in other areas. If we use this idea here, will we achieve better results?

Central Dogma of Molecular Biology

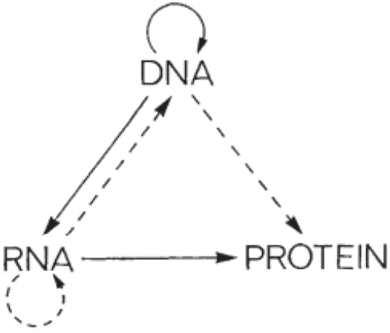


Figure 2.1: Molecular biology information flow. Solid arrows show general transfers; dotted arrows show special transfers. The absent arrows are the undetected transfers specified by the central dogma.

Source: (CRICK, 1970).

Genetic Expression

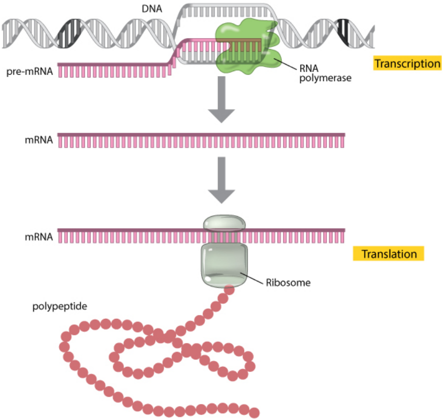


Figure 2.2: Genetic expression process.

Source: (CLANCY; BROWN, 2008).

MicroRNA-mRNA Interaction

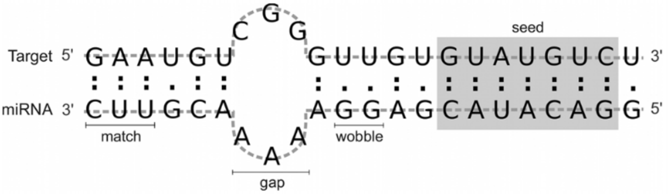

seed
Target 5' G A A U G U C G G G U U G U G U A U G U C U 3'
miRNA 3' C U U G C A A A A A G G A G C A U A C A G G 5'
match
wobble
gap

Figure 2.3: Example of miRNA-target alignment. Nucleotides matches are shown by colons and G:U wobble pairs are represented by dots. There can be gaps.

Source: (MENDOZA *et al.*, 2013).

Supervised Learning

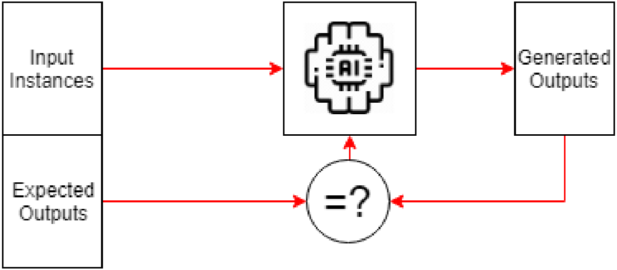


Figure 3.1: General Supervised Learning.

Source: Elaborated by the Author.

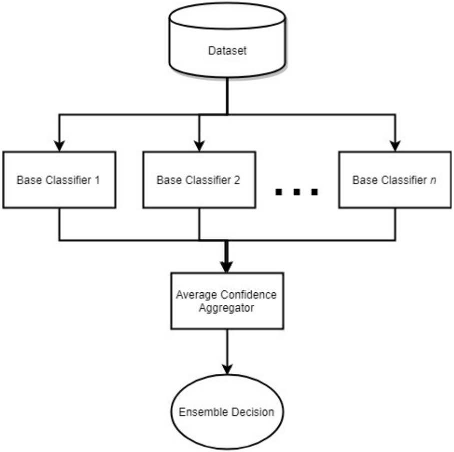


Figure 3.2: Schematic of the Voting Classifier method.

Source: Author

Genetic Algorithms (GA)

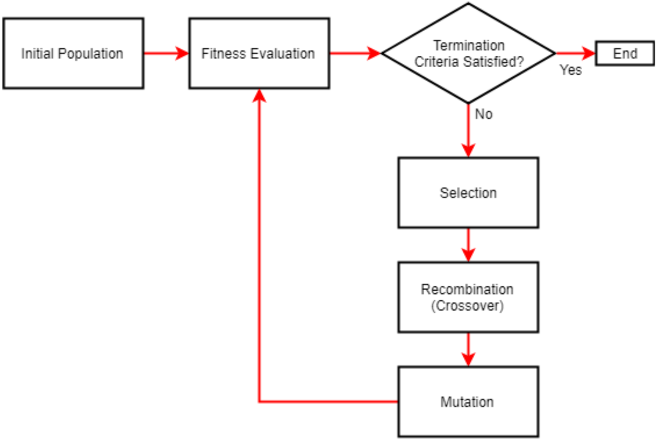


Figure 3.3: Genetic Algorithm Loop.

Source: Elaborated by the Author.

Model Evaluation
K-fold Cross-Validation

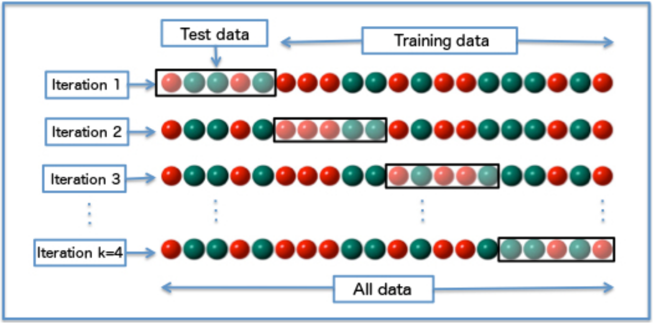
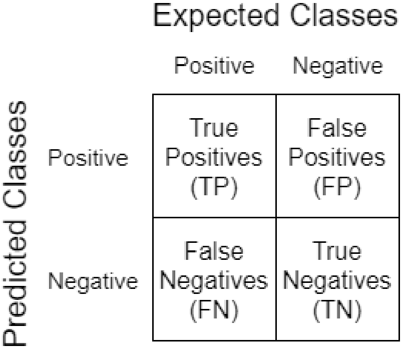


Figure 3.4: Diagram of a k-fold Cross Validation, with $k = 4$.

Source: (WIKIPEDIA, 2018).



		Expected Classes	
		Positive	Negative
Predicted Classes	Positive	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Figure 3.5: 2x2 Confusion Matrix

Source: Elaborated by the Author.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1Score = \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

$$TPR = Recall = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity \tag{7}$$

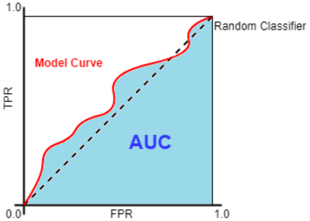


Figure 3.6: Area Under ROC Curve for a Model. A real model curve won't be curved, it will look like a ladder.

Source: Elaborated by the Author.

$$E = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{(L - \lceil L/2 \rceil)}\min\{l(z_j), L - l(z_j)\} \qquad (8)$$

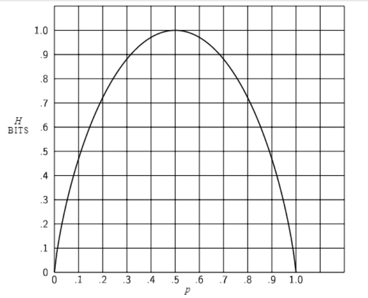


Figure 3.7: Entropy H in the case of two possibilities with probabilities p and $(1 - p)$.

Source: (SHANNON, 1948).

Homogeneous Ensemble
Mendoza *et al.*, 2013

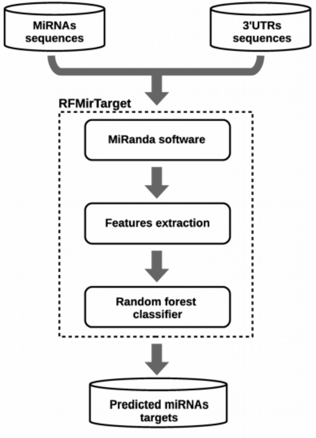


Figure 4.1: Mendoza *et al.* (2013) proposed framework, RFMirTarget.

Source: (MENDOZA *et al.*, 2013).

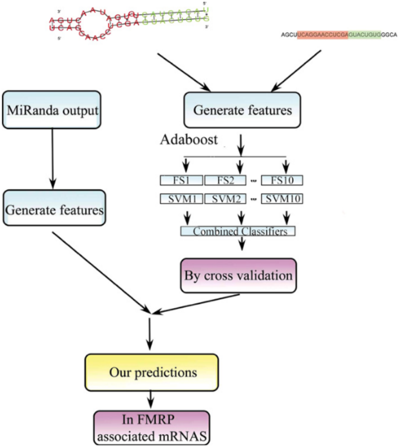


Figure 4.2: Yan *et al.* (2007) proposed workflow.

Source: (YAN *et al.*, 2007)

Combining Performance and Diversity Measures for Optimizing Classification Ensembles via a Genetic Algorithm in the miRNA-Target Prediction Problem

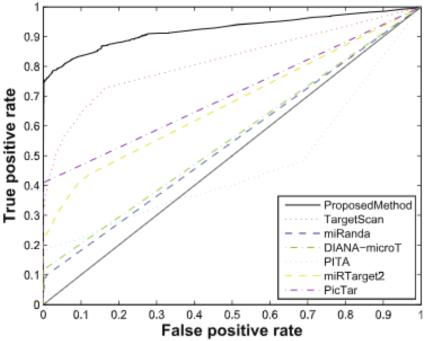


Figure 4.3: Comparison of Yu *et al.* (2014) proposed ensemble method against the tools used in the ensemble. It outperforms them by 52.5% in terms of AUC Score.

Source: (YU *et al.*, 2014)

Heterogeneous Ensemble
Le *et al.*, 2015

Combining Performance and Diversity Measures for Optimizing Classification Ensembles via a Genetic Algorithm in the miRNA-Target Prediction Problem

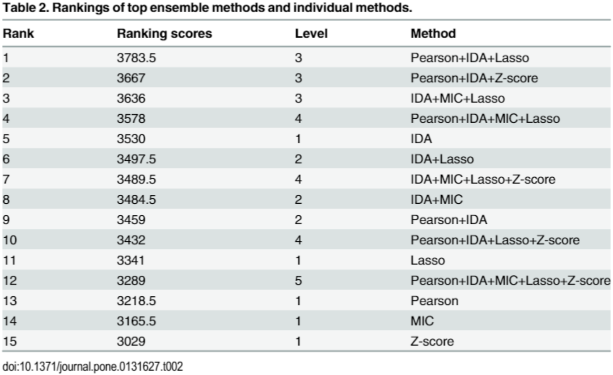


Table 2. Rankings of top ensemble methods and individual methods.

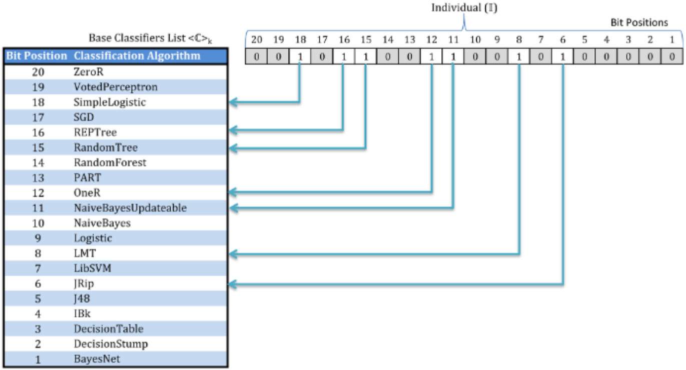
Rank	Ranking scores	Level	Method
1	3783.5	3	Pearson+IDA+Lasso
2	3667	3	Pearson+IDA+Z-score
3	3636	3	IDA+MIC+Lasso
4	3578	4	Pearson+IDA+MIC+Lasso
5	3530	1	IDA
6	3497.5	2	IDA+Lasso
7	3489.5	4	IDA+MIC+Lasso+Z-score
8	3484.5	2	IDA+MIC
9	3459	2	Pearson+IDA
10	3432	4	Pearson+IDA+Lasso+Z-score
11	3341	1	Lasso
12	3289	5	Pearson+IDA+MIC+Lasso+Z-score
13	3218.5	1	Pearson
14	3165.5	1	MIC
15	3029	1	Z-score

doi:10.1371/journal.pone.0131627.t002

Figure 4.4: Le *et al.* (2015) ranking of different ensemble compositions against the individual methods.

Source: (LE *et al.*, 2015)

GA with Heterogeneous Ensemble
Haque *et al.*, 2016



20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
0	0	1	0	1	1	0	0	1	1	0	0	1	0	1	0	0	0	0	0

Bit Position	Classification Algorithm
20	ZeroR
19	VotedPerceptron
18	SimpleLogistic
17	SGD
16	REPTree
15	RandomTree
14	RandomForest
13	PART
12	OneR
11	NaiveBayesUpdateable
10	NaiveBayes
9	Logistic
8	LMT
7	LibSVM
6	JRip
5	J48
4	IBk
3	DecisionTable
2	DecisionStump
1	BayesNet

Figure 4.5: Haque *et al.* (2016) genotype ensemble representation example.

Source: (HAQUE *et al.*, 2015)

GA with Heterogeneous Ensemble Haque *et al.*, 2016

Combining Performance and Diversity Measures for Optimizing Classification Ensembles via a Genetic Algorithm in the miRNA-Target Prediction Problem

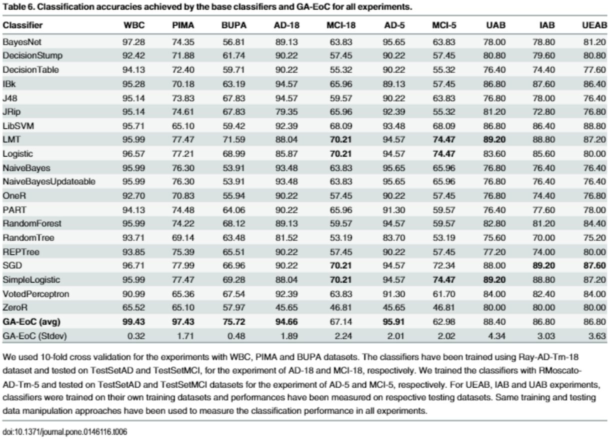


Table 6. Classification accuracies achieved by the base classifiers and GA-EoC for all experiments.

Classifier	WBC	PIMA	BUPA	AD-18	MCI-18	AD-5	MCI-5	UAB	IAB	UEAB
BayesNet	97.28	74.35	56.81	89.13	63.83	95.65	63.83	78.00	78.80	81.20
DecisionStump	92.42	71.88	61.74	90.22	57.45	90.22	57.45	80.80	79.60	80.80
DecisionTable	94.13	72.40	59.71	90.22	55.32	90.22	55.32	76.40	74.40	77.60
IBk	95.28	70.18	63.19	94.57	65.96	89.13	57.45	86.80	87.60	86.40
J48	95.14	73.83	67.83	94.57	59.57	90.22	63.83	76.80	78.00	76.40
JRip	95.14	74.61	67.83	79.35	65.96	92.39	55.32	81.20	72.80	76.80
LibSVM	95.71	65.10	59.42	92.39	68.09	93.48	68.09	86.80	86.40	88.80
LMT	95.99	77.47	71.59	88.04	**70.21**	94.57	**74.47**	**89.20**	88.80	87.20
Logistic	96.57	77.21	68.99	85.87	**70.21**	94.57	**74.47**	83.60	85.60	80.00
NaiveBayes	95.99	76.30	53.91	93.48	63.83	95.65	65.96	76.80	76.40	76.40
NaiveBayesUpdateable	95.99	76.30	53.91	93.48	63.83	95.65	65.96	76.80	76.40	76.40
OneR	92.70	70.83	55.94	90.22	57.45	90.22	57.45	76.80	74.40	76.80
PART	94.13	74.48	64.06	90.22	65.96	91.30	59.57	76.40	77.60	78.00
RandomForest	95.99	74.22	68.12	89.13	59.57	94.57	59.57	82.80	81.20	84.40
RandomTree	93.71	69.14	63.48	81.52	53.19	83.70	53.19	75.60	70.00	75.20
REPTree	93.85	75.39	65.51	90.22	57.45	90.22	57.45	77.20	74.00	80.00
SGD	96.71	77.99	66.96	90.22	**70.21**	94.57	72.34	88.00	**89.20**	**87.60**
SimpleLogistic	95.99	77.47	69.28	88.04	**70.21**	94.57	**74.47**	**89.20**	88.80	87.20
VotedPerceptron	90.99	65.36	67.54	92.39	63.83	91.30	61.70	84.00	82.40	84.00
ZeroR	65.52	65.10	57.97	45.65	46.81	45.65	46.81	80.00	80.00	80.00
GA-EoC (avg)	**99.43**	**97.43**	**75.72**	**94.66**	67.14	**95.91**	62.98	88.40	86.80	86.80
GA-EoC (Stdev)	0.32	1.71	0.48	1.89	2.24	2.01	2.02	4.34	3.03	3.63

We used 10-fold cross validation for the experiments with WBC, PIMA and BUPA datasets. The classifiers have been trained using Ray-AD-Tm-18 dataset and tested on TestSetAD and TestSetMCI, for the experiment of AD-18 and MCI-18, respectively. We trained the classifiers with RMoscato-AD-Tm-5 and tested on TestSetAD and TestSetMCI datasets for the experiment of AD-5 and MCI-5, respectively. For UEAB, IAB and UAB experiments, classifiers were trained on their own training datasets and performances have been measured on respective testing datasets. Same training and testing data manipulation approaches have been used to measure the classification performance in all experiments.

doi:10.1371/journal.pone.0146116.t006

Figure 4.6: Haque *et al.* (2016)'s GA-EoC comparison against the individual algorithms.

Source: (HAQUE *et al.*, 2015)

GA with Heterogeneous Ensemble Haque *et al.*, 2016

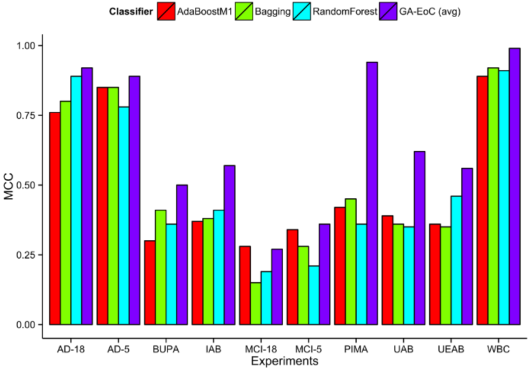


Fig 7. Comparison of MCC scores achieved by GA-EoC and other ensemble of classifiers (AdaBoostM1, Bagging and Boosting) for all experiments.

doi:10.1371/journal.pone.0146116.g007

Figure 4.7: Haque *et al.* (2016)'s GA-EoC comparison against other ensemble methods.

Source: (HAQUE *et al.*, 2015)

GA with Heterogeneous Ensemble Mousavi, Eftekhari, Haghighi, 2015

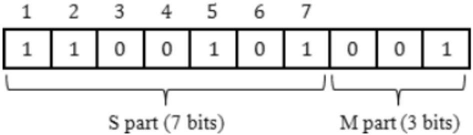


Fig. 3. Chromosome representation for a problem with seven classifiers and eight values for M.

Figure 4.8: Mousavi, Eftekhari, Haghighi (2015) genotype ensemble representation example. Each chromosome encodes a subset of classifiers and one value of M.

Source: (MOUSAVI; EFTEKHARI; HAGHIGHI, 2015)

GA with Heterogeneous Ensemble Mousavi, Eftekhari, Haghighi, 2015

Combining Performance and Diversity Measures for Optimizing Classification Ensembles via a Genetic Algorithm in the miRNA-Target Prediction Problem

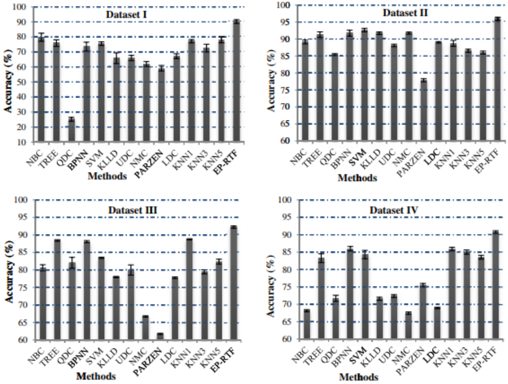


Fig. 5. Comparison of average accuracy with the corresponding standard deviation over the four human miRNAs target prediction datasets of the 14 methods.

Figure 4.9: Mousavi, Eftekhari, Haghighi (2015)'s EP-RTF comparison against the individual algorithms.

Source: (MOUSAVI; EFTEKHARI; HAGHIGHI, 2015)

GA with Heterogeneous Ensemble Mousavi, Eftekhari, Haghighi, 2015

Combining Performance and Diversity Measures for Optimizing Classification Ensembles via a Genetic Algorithm in the miRNA-Target Prediction Problem

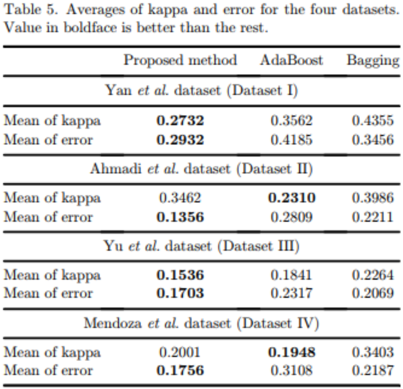


Table 5. Averages of kappa and error for the four datasets. Value in boldface is better than the rest.

	Proposed method	AdaBoost	Bagging
Yan *et al.* dataset (Dataset I)			
Mean of kappa	**0.2732**	0.3562	0.4355
Mean of error	**0.2932**	0.4185	0.3456
Ahmadi *et al.* dataset (Dataset II)			
Mean of kappa	0.3462	**0.2310**	0.3986
Mean of error	**0.1356**	0.2809	0.2211
Yu *et al.* dataset (Dataset III)			
Mean of kappa	**0.1536**	0.1841	0.2264
Mean of error	**0.1703**	0.2317	0.2069
Mendoza *et al.* dataset (Dataset IV)			
Mean of kappa	0.2001	**0.1948**	0.3403
Mean of error	**0.1756**	0.3108	0.2187

Figure 4.10: Mousavi, Eftekhari, Haghighi (2015)'s EP-RTF comparison against other ensemble methods.

Source: (MOUSAVI; EFTEKHARI; HAGHIGHI, 2015)

	Positive/Functional	Negative/Non-Functional	Total
miRTarBase v6.1	6958 *(96.1%)*	283 *(3.9%)*	7241
DIANA-TarBase v7.0	**5619 *(74.3%)***	**1944 *(25.7%)***	**7563**

Table 5.1: Positive and negative examples in miRTarBase v6.1 and DIANA-TarBase v7.0, the latter being used in the current work.

Source: Elaborated by the Author.

miRNA and mRNA Sequences

- miRNA sequences were gathered from miRBase.

miRNA and mRNA Sequences

- miRNA sequences were gathered from miRBase.
- mRNA sequences were gathered from the BioMart portal.
  - In this source, the same mRNA can have different versions.

Final Valid Instances

	miRTarBase	Combinations	**Valid Combinations**	Different miRNA
Positive/Functional	5,619 *(74.3%)*	23,019 *(74,6%)*	**7,100 *(76.9%)***	2441
Negative/Non-Functional	1,944 *(25.7%)*	7,855 *(25.4%)*	**2,131 *(23.1%)***	746
Total	7,563	30,874	**9,231**	-[1]

Table 5.2: Progression of positive and negative examples.

Source: Elaborated by the Author.

[1] - There is an intersection between miRNA in Positive and Negative classes.

#	Feature Name	#	Feature Name
1	Alignment Score (by miRanda)	18	Position 10
2	Alighment Length	19	Position 11
3	Minimum free energy of the alignment	20	Position 12
4	G:C's absolute frequency in the alignment	21	Position 13
5	A:U's absolute frequency in the alignment	22	Position 14
6	G:U's absolute frequency in the alignment	23	Position 15
7	Number of gaps in the alignment	24	Position 16
8	Number of mismatches in the alignment	25	Position 17
9	Position 1	26	Position 18
10	Position 2	27	Position 19
11	Position 3	28	Position 20
12	Position 4	29	Minimum free energy of the seed
13	Position 5	30	G:C's absolute frequency in the seed
14	Position 6	31	A:U's absolute frequency in the seed
15	Position 7	32	G:U's absolute frequency in the seed
16	Position 8	33	Number of gaps in the seed
17	Position 9	34	Number of mismatches in the seed

Table 5.3: Features used in this work, based on Mendoza *et al.* (2013)'s features for RFMirTarget.

Source: Elaborated by the Author.

Genetic Algorithm

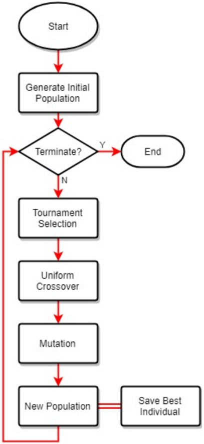
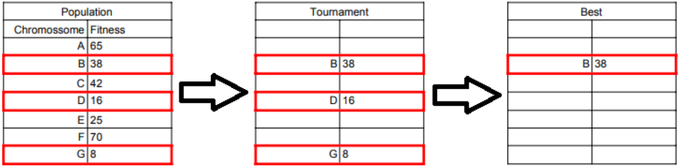


Figure 5.1: Genetic Algorithm high level execution pipeline.

Source: Elaborated by the Author.

Genetic Algorithm
Tournament Selection

- Select the fittest with probability p, the second fittest with probability $p \cdot (1 - p)$, the third fittest with probability $p \cdot ((1 - p)^2)$, and so on. When $p = 1$, it is called a *deterministic* tournament selection.



Population	
Chromossome	Fitness
A	65
B	38
C	42
D	16
E	25
F	70
G	8

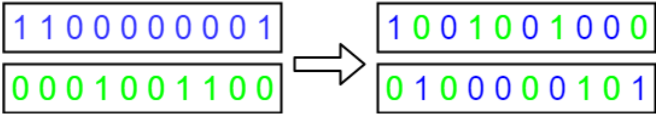
Tournament	
B	38
D	16
G	8

Best	
B	38

Figure 5.2: Tournament Selection example, with $size = 3$ and $p = 1$.

Source: Elaborated by the Author.

Genetic Algorithm
Uniform Crossover



1 1 0 0 0 0 0 0 0 1

0 0 0 1 0 0 1 1 0 0

1 0 0 1 0 0 1 0 0 0

0 1 0 0 0 0 0 1 0 1

Figure 5.3: Uniform Crossover example.

Source: Elaborated by the Author.

Genetic Algorithm
Flip-Bit Mutation

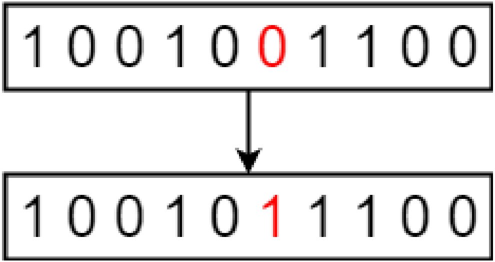


Figure 5.4: Flip Bit example.

Source: Elaborated by the Author.

Genetic Algorithm

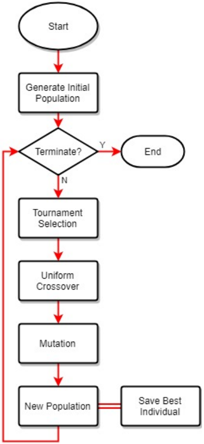


Figure 5.1: Genetic Algorithm high level execution pipeline.

Source: Elaborated by the Author.

Genetic Algorithm's Hyperparameters

Hyperparameter Name	Hyperparameter Value
Population Size	55
Crossover Rate	60%
Mutation Rate	1%
Elitism Size	1
Tournament Size	Population Size$/10 = 5$
Generations Limit	10

Table 5.4: Adopted GA's hyperparameters configuration.

Source: Elaborated by the Author.

Population Size

$$population\_size = min\left((5 \times k), \left(\frac{2^k}{2}\right)\right) \tag{9}$$

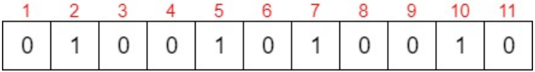
Ensemble's Base Classifiers

#	Classifier Name	Classifier *scikit-learn* Call
1	Gaussian Naïve Bayes	*GaussianNB()*
2	Decision Tree (Gini index, $max\_depth = 5$)	*DecisionTreeClassifier(max_depth=5, criterion='gini')*
3	Decision Tree (Entropy, $max\_depth = 5$)	*DecisionTreeClassifier(max_depth=5, criterion='entropy')*
4	Random Forest (Gini index, $max\_depth = 5$)	*RandomForestClassifier(max_depth=5, criterion='gini')*
5	Random Forest (Gini index, $max\_depth = 5$)	*RandomForestClassifier(max_depth=5, criterion='entropy')*
6	Quadratic Discriminant Analysis	*QuadraticDiscriminantAnalysis()*
7	Support Vector Machine	*SVC(kernel='rbf', probability=True)*
8	K-Nearest Neighbors ($K = 3$)	*KNeighborsClassifier(n_neighbors=3)*
9	K-Nearest Neighbors ($K = 5$)	*KNeighborsClassifier(n_neighbors=5)*
10	K-Nearest Neighbors ($K = 7$)	*KNeighborsClassifier(n_neighbors=7)*
11	Logistic Regression	*LogisticRegression()*

Table 5.5: Classifiers used in the ensemble.

Source: Elaborated by the Author.

Genetic Algorithm's Chromosome



1	2	3	4	5	6	7	8	9	10	11
0	1	0	0	1	0	1	0	0	1	0

Figure 5.5: Example chromosome representing an ensemble that uses classifiers $\#2$, $\#5$, $\#7$, and $\#10$ from Table 3.1.

Source: Elaborated by the Author.

Ensemble Construction and Evaluation

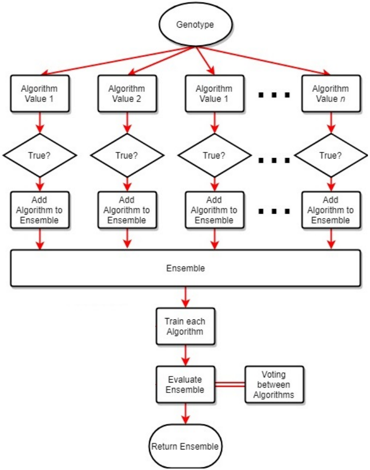


Figure 5.6: Ensemble construction and evaluation.

Source: Elaborated by the Author.

Ensemble's Fitness

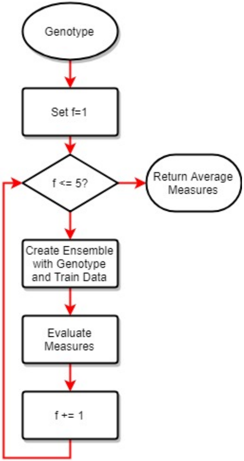


Figure 5.7: Stratified K-Fold Cross Validation ensemble evaluation.

Source: Elaborated by the Author.

Ensemble's Fitness

$$fitness(x) = \beta \times Performance(x) + (1 - \beta) \times Diversity(x) \tag{10}$$

Discarded Variations

- Rotation Forest and Bagging.

Discarded Variations

- Rotation Forest and Bagging.
- Downsampling and Oversampling.

Discarded Variations

- Rotation Forest and Bagging.
- Downsampling and Oversampling.
- Some classifiers:
  - Neural Networks.
  - Support Vector Machine with Sigmoid Kernel.
  - Stochastic Gradient Descent.

Discarded Variations

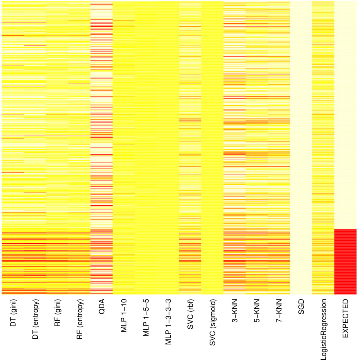


Figure 5.8: Heatmap of classifiers predicted probabilities for the positive class for instances in a test dataset. Probabilities closer to 0.0 are shown in red, whereas probabilities closer to 1.0 are represented in light yellow.

Source: Elaborated by the Author.

- The following GA Fitness functions were tested:

- Pure Accuracy
- 75% Accuracy + 25% Diversity
- 50% Accuracy + 50% Diversity
- 25% Accuracy + 75% Diversity

- Pure AUC Score
- 75% AUC Score + 25% Diversity
- 50% AUC Score + 50% Diversity
- 25% AUC Score + 75% Diversity

- Pure F1 Measure
- 75% F1 Measure + 25% Diversity
- 50% F1 Measure + 50% Diversity
- 25% F1 Measure + 75% Diversity

- Pure MCC
- 75% MCC + 25% Diversity
- 50% MCC + 50% Diversity
- 25% MCC + 75% Diversity

Experimental Methodology

- The following GA Fitness functions were tested:

- Pure Accuracy
- 75% Accuracy + 25% Diversity
- 50% Accuracy + 50% Diversity
- 25% Accuracy + 75% Diversity

- Pure AUC Score
- 75% AUC Score + 25% Diversity
- 50% AUC Score + 50% Diversity
- 25% AUC Score + 75% Diversity

- Pure F1 Measure
- 75% F1 Measure + 25% Diversity
- 50% F1 Measure + 50% Diversity
- 25% F1 Measure + 75% Diversity

- Pure MCC
- 75% MCC + 25% Diversity
- 50% MCC + 50% Diversity
- 25% MCC + 75% Diversity

- Accuracy and Diversity are results of a 5-Fold Stratified Cross-Validation.

Experimental Methodology

- The following GA Fitness functions were tested:

- Pure Accuracy
- 75% Accuracy + 25% Diversity
- 50% Accuracy + 50% Diversity
- 25% Accuracy + 75% Diversity

- Pure AUC Score
- 75% AUC Score + 25% Diversity
- 50% AUC Score + 50% Diversity
- 25% AUC Score + 75% Diversity

- Pure F1 Measure
- 75% F1 Measure + 25% Diversity
- 50% F1 Measure + 50% Diversity
- 25% F1 Measure + 75% Diversity

- Pure MCC
- 75% MCC + 25% Diversity
- 50% MCC + 50% Diversity
- 25% MCC + 75% Diversity

- Accuracy and Diversity are results of a 5-Fold Stratified Cross-Validation.
- Each fitness function was tested 10 times.

Genetic Algorithm Learning Curves

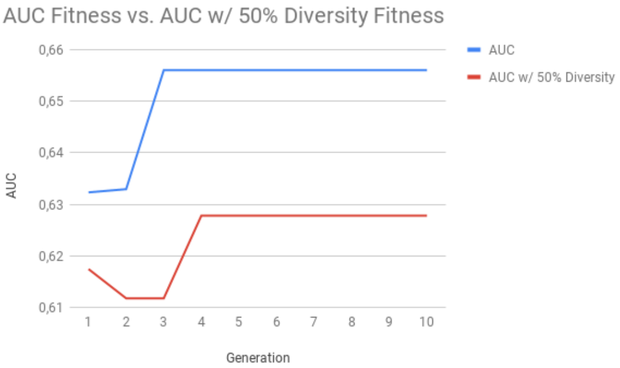
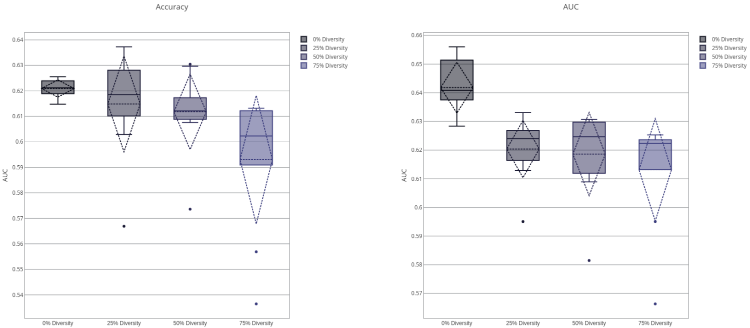


Figure 6.1: Learning curves for fitness curves using pure AUC and AUC combined with 50% of diversity. Performed is compared by means of AUC Score.

Source: Elaborated by the Author.

Different Diversity Proportions



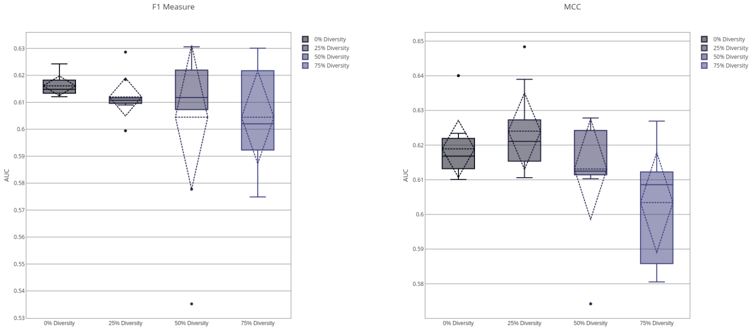
(a) Accuracy

(b) AUC

Figure 6.2a: Boxplots comparing the performance in terms of AUC Score. Results are extracted from 10 repetitions of the proposed solution.

Source: Elaborated by the Author.

Different Diversity Proportions



(c) F1 Measure

(d) MCC

Figure 6.2b: Boxplots comparing the performance in terms of AUC Score. Results are extracted from 10 repetitions of the proposed solution.

Source: Elaborated by the Author.

Individual Classifiers & Full Ensemble

Approach	AUC Score
Gaussian Naïve Bayes	0.5943 ± 0.0106
Decision Tree (Gini index, $max\_depth = 5$)	0.6182 ± 0.0111
Decision Tree (Entropy, $max\_depth = 5$)	**0.6220 ± 0.0075**
Random Forest (Gini index, $max\_depth = 5$)	0.6110 ± 0.0096
Random Forest (Entropy, $max\_depth = 5$)	0.6108 ± 0.0082
Quadratic Discriminant Analysis	0.5868 ± 0.0382
Support Vector Machine	0.6054 ± 0.0135
K-Nearest Neighbors ($K = 3$)	0.6084 ± 0.0195
K-Nearest Neighbors ($K = 5$)	0.6058 ± 0.0198
K-Nearest Neighbors ($K = 7$)	0.5961 ± 0.0165
Logistic Regression	0.5365 ± 0.0062
Full Ensemble	0.6107 ± 0.0110
Genetic Algorithm (AUC w/o Diversity)	**0.6418 ± 0.0094**
Genetic Algorithm (AUC w/ 25% Diversity)	0.6204 ± 0.0105
Genetic Algorithm (AUC w/ 50% Diversity)	0.6186 ± 0.0153
Genetic Algorithm (AUC w/ 75% Diversity)	0.6131 ± 0.0188

Table 6.1: Average AUC Score for each approach.

Source: Elaborated by the Author.

Conclusion

- Using an Genetic Algorithm to build an heterogeneous ensemble for the miRNA-target prediction problem produces better ensembles than simply using all possible classifiers, or using them alone.

Conclusion

- Using an Genetic Algorithm to build an heterogeneous ensemble for the miRNA-target prediction problem produces better ensembles than simply using all possible classifiers, or using them alone.
- We can't conclude from the results obtained if adding a diversity measure to the GA fitness function increases its performance of not, due to:

Conclusion

- Using an Genetic Algorithm to build an heterogeneous ensemble for the miRNA-target prediction problem produces better ensembles than simply using all possible classifiers, or using them alone.
- We can't conclude from the results obtained if adding a diversity measure to the GA fitness function increases its performance of not, due to:
  - **Class unbalancing in the dataset**.

Conclusion

- Using an Genetic Algorithm to build an heterogeneous ensemble for the miRNA-target prediction problem produces better ensembles than simply using all possible classifiers, or using them alone.
- We can't conclude from the results obtained if adding a diversity measure to the GA fitness function increases its performance of not, due to:
  - **Class unbalancing in the dataset**.
  - Relatively small sample size, presented a big challenge to the proposed technique.

Conclusion

- Using an Genetic Algorithm to build an heterogeneous ensemble for the miRNA-target prediction problem produces better ensembles than simply using all possible classifiers, or using them alone.
- We can't conclude from the results obtained if adding a diversity measure to the GA fitness function increases its performance of not, due to:
  - **Class unbalancing in the dataset**.
  - Relatively small sample size, presented a big challenge to the proposed technique.
- Higher variance when using the entropy in the GA fitness function leaded to some ensembles that are better than those generated with a pure performance fitness, but also generated worse ones.

Future Work

- A larger and more balanced miRNA-mRNA dataset is strongly required for a new execution of the proposed solution to achieve more certain conclusions.

Future Work

- A larger and more balanced miRNA-mRNA dataset is strongly required for a new execution of the proposed solution to achieve more certain conclusions.
- Explore newer state-of-art multi-objective GA approaches, such as NSGA-II.

Questions?

Gabriel Marangoni Moita

gmmoita@inf.ufrgs.br

.inf INSTITUTO DE INFORMÁTICA UFRGS