Learning to Combine miRNA Target Predictions: a Semi-supervised Ensemble Learning Approach
(Discussion Paper)

Gianvito Pio[1], Michelangelo Ceci[1], Domenica D'Elia[2] and Donato Malerba[1]

[1] University of Bari - Dept. of Computer Science - Via Orabona, 4 - 70125 Bari, Italy
[2] ITB-CNR, Via Amendola 122/D, 70126, Bari, Italy
{name.surname}@uniba.it, domenica.delia@ba.itb.cnr.it

Abstract. Link prediction in network data is a data mining task which is receiving significant attention due to its applicability in various domains. An example can be found in social network analysis, where the goal is to identify connections between users. Another application can be found in computational biology, where the goal is to identify previously unknown relationships among biological entities. For example, the identification of regulatory activities (links) among genes would allow biologists to discover possible *gene regulatory networks*. In the literature, several approaches for link prediction can be found, but they often fail in simultaneously considering all the possible criteria (e.g. network topology, nodes properties, autocorrelation among nodes). In this paper we present a semi-supervised data mining approach which *learns to combine* the scores returned by several link prediction algorithms. The proposed solution exploits both a small set of validated examples of links and a huge set of unlabeled links. The application we consider regards the identification of links between genes and miRNAs, which can contribute to the understanding of their roles in many biological processes. The specific application requires to learn from only positively labeled examples of links and to face with the high unbalancing between labeled and unlabeled examples. Results show a significant improvement with respect to single prediction algorithms and with respect to baseline combination.

1 Introduction

Networks have become ubiquitous in several social, economical and scientific fields, ranging from the Internet to social sciences, and including biology, epidemiology, geography, finance, and many others. Indeed, researchers in these fields have proven that systems of different nature can be represented as networks [7]. When working with networks, one of the most important data mining tasks is that of network reconstruction and, more particularly, link prediction.

In the biological context, link prediction is important to reconstruct both homogeneous networks, such as protein-protein interaction networks, and heteroge-

neous networks, such as microRNAs (miRNAs) - messenger RNAs (mRNAs) regulatory networks. We focus on this last example, where miRNAs are small non-coding RNA molecules representing one of the most interesting classes of gene regulators. Their exploitation in the clinical domain depends on the understanding of their functions. Although miRNAs are post-transcriptional regulators that inhibit translation of mRNAs, the identification of individual miRNA:mRNA interactions is not enough to catch their capacity to regulate complex gene networks. Therefore, in [8, 9] we developed the biclustering algorithm HOCCLUS2 which aims at identifying (possibly) overlapping miRNA:mRNA regulatory networks, organized in a hierarchy. However, experiments performed both on interactions validated *in vitro* and interactions predicted by automatic tools revealed that the quality of extracted biclusters is negatively affected by the poor reliability of the predictions. In this case, approaches based on link predictions produce results often uncorrelated to each other and with a poor degree of overlap. This mainly depends on the impossibility to incorporate in a single model all the interplaying variants that influence the effects of the miRNA targeting. One solution to this issue consists in the combination of the output of several prediction algorithms. In [14], some approaches were compared, i.e. majority vote, ranking aggregation and Bayesian network classification. This is one of the first attempts to exploit data mining approaches to *learn to combine* predictions and improve the reliability of predictions. However, it can be applied only when a large number examples of interactions (both positive and negative) is available.

In general, when exploiting combination approaches, some issues must be considered: *i)* Very few interactions are validated and can be considered as "stable" examples. *ii)* In specific applications, such as that of miRNA:mRNA networks, only positive examples are available. *iii)* Prediction algorithms consider similar features and their combination can lead to collinearity problems [3].

In this paper, we describe the work published in [10]. In order to face *i)* and *ii)* , we propose a semi-supervised learning algorithm, which considers both positively labeled examples of interactions and the huge set of unlabeled (unknown) instances. As for *iii)*, the collinearity problem is alleviated by considering as features the scores (outputs) obtained by several prediction algorithms (instead of original features), resorting to a solution which is similar to meta-learning algorithms. The advantage of applying machine learning techniques to the outputs of prediction algorithms consists in automatically adapting to unknown patterns of the outputs and performing more reliable predictions when these patterns occur.

2 Learning to Combine Predictions

Before formally defining the problem, we introduce some definitions. Let:

- $\mathcal{M}$ and $\mathcal{G}$ be the sets of miRNAs and mRNAs, respectively;
- $x = \langle m, g \rangle \in (\mathcal{M} \times \mathcal{G})$ be an interaction between miRNA m and mRNA g;
- $p_k(x)$ be the prediction score for the interaction x of the k-th algorithm;
- $p(x) = [p_1(x), p_2(x), \ldots, p_s(x)]$ be the vector of scores for the interaction x;
- $l(x)$ be a function returning 1 if x is labeled (validated), 0 otherwise;

– $f(x)$ be a function returning 1 if x represents a true interaction, 0 otherwise;
– $L = \{x | x \in (\mathcal{M} \times \mathcal{G}) \wedge l(x) = 1\}$ be the subset of labeled interactions;
– $U = (\mathcal{M} \times \mathcal{G}) - L$ be the subset of unlabeled interactions.

In our case, since only positive interactions are labeled, the following holds:

$$P(f(x) = 1 | l(x) = 1) = 1 \tag{1}$$

The goal is to learn a function $f'(p(x))$ which approximates the probability that $f(x) = 1$, that is $f'(p(x)) \approx P(f(x) = 1)$. As suggested in [5], it can be learned by exploiting (1) in the following steps:

$$P(l(x) = 1) = P(f(x) = 1 \wedge l(x) = 1) = P(l(x) = 1 | f(x) = 1) \cdot P(f(x) = 1) \tag{2}$$

This means that

$$f'(p(x)) \approx P(f(x) = 1) = \frac{P(l(x) = 1)}{P(l(x) = 1 | f(x) = 1)} \tag{3}$$

In this equation, both the numerator and the denominator can be estimated by a *nontraditional classifier*, whose exploitation is explained in the following.

2.1 Estimating $\mathbf{P(l(x) = 1)}$ and $\mathbf{P(l(x) = 1 | f(x) = 1)}$

In our work, the nontraditional classifier is learned through LIBSVM [2] with Platt scaling, in order to get probability estimates. However, it is noteworthy that other probabilistic classifiers can be plugged into our framework. In particular, LIBSVM is applied to solve the following problem:

Given: a set of training examples $\{(p(x), l(x))\}_x$, where $p(x)$ is the vector of prediction scores associated to the interaction x and $l(x)$ (1 if the example is labeled, 0 otherwise) represents the class for the nontraditional classifier;

Find: a probability function $g : \mathbb{R}^s \rightarrow \mathbb{R}$ which takes as its input a vector of prediction scores $p(x)$ and returns the probability that the interaction x is labeled. In this way, $g(p(x)) \approx P(l(x) = 1)$.

As for the denominator, we assume that all labeled positive examples are taken completely randomly from all positive examples. Formally:

$$P(l(x) = 1 | f(x) = 1) = P(l(\cdot) = 1 | f(\cdot) = 1) \tag{4}$$

In other words, $P(l(x) = 1 | f(x) = 1))$ is independent of the specific interaction x. This assumption is essential when learning is performed from only positive examples and is coherent with the "selected completely at random" assumption in [5]. Indeed, it allows us to use $g(p(x))$ also in the computation of $P(l(x) = 1 | f(x) = 1))$. In particular, since a possible estimator of $P(l(\cdot) = 1 | f(\cdot) = 1))$ is the average value of $g(p(x))$ for all labeled positive examples, we have:

$$P(l(x) = 1 | f(x) = 1) = P(l(\cdot) = 1 | f(\cdot) = 1) \approx \frac{1}{|L|} \sum_{x \in L} g(p(x)) \tag{5}$$

Differently from [5], we also have to deal with the problem of unbalanced class distributions when learning the nontraditional classifier $g(p(x))$. This problem is solved by resorting to a sampling solution, which is illustrated in the following.

2.2 Ensemble Learning g(·)

The sampling procedure considered here is similar to that used in bootstrap estimation [4] and in some ensemble data mining methods, such as bagging [1].

In particular, we learn K nontraditional classifiers $g_j(p(x)),\ j = 1, \ldots, K$, through LIBSVM, that are combined to obtain $g(p(x))$. Each classifier is learned from the set of examples $L \cup U^j$ $(j = 1, 2, \ldots, K)$, that is, from all the labeled examples L and a subset U^j of the unlabeled set U. The K subsets of unlabeled examples are built by randomly sampling, with replacement, n examples from U. The proportion of unlabeled examples in each U^j is $\frac{n}{|U|}$.

It is noteworthy that the K samples U^j are neither mutually exclusive nor exhaustive, i.e., they do not partition the original data set, so, for instance, even $K = 10$ samples with $n = 0.1 \cdot |U|$ do not generally cover the entire set of unlabeled examples U. Differently from data partitioning, which is affected by only one parameter K (the number of partitions), the data sampling procedure used in this work is controlled by two parameters: n and γ. The first parameter represents the number of unlabeled examples in each sample and can be reasonably chosen on the basis of the number of labeled examples, so that the unbalancing problem is mitigated. The second parameter represents the percentage of unlabeled examples we intend to take into account.

Once the K classifiers are learned, each function $g_j(p(x))$ is applied to obtain an estimate of $P(l(x) = 1)$ for all the examples in U^j. Since the same unlabeled example can belong to more than one sample, the following equation is used:

$$g(p(x)) = \underset{\{j \mid x \in U^j\}}{average}\ g_j(p(x)) \tag{6}$$

2.3 Ensemble Learning f′(·)

In order to compute $f'(p(x))$, a solution would be to apply Equation (3). However, as shown in [5], a more effective way is the computation of a weight for each example and in training a further (traditional) classifier. Specifically, we compute the probability that an unlabeled example x is a positive example as:

$$\begin{aligned} P(f(x) = 1 | l(x) = 0) &= \frac{P(l(x) = 0, f(x) = 1) P(f(x) = 1)}{P(l(x) = 0)} \\ &= \frac{[1 - P(l(x) = 1 | f(x) = 1)] P(f(x) = 1)}{1 - P(l(x) = 1)} \\ &= \frac{[1 - P(l(x) = 1 | f(x) = 1)] \cdot \frac{P(l(x)=1)}{P(l(x)=1|f(x)=1)}}{1 - P(l(x) = 1)} \end{aligned}$$

which, according to Equation (5), can be approximated to:

$$P(f(x) = 1 | l(x) = 0) \approx \frac{1 - c}{c} \cdot \frac{P(l(x) = 1)}{1 - P(l(x) = 1)}$$

where $c = \frac{1}{|L|} \sum_{x \in L} g(p(x))$ and $P(l(x) = 1)$ is approximated to $g(p(x))$.

The training set for the classifier which learns $f'(p(x))$ is then built as:

$$training\_label(x) = \begin{cases} + & if\ x \in L \\ + & if\ x \in U \wedge P(f(x) = 1 | l(x) = 0) \geq P(f(x) = 0 | l(x) = 0) \\ - & otherwise \end{cases}$$

Algorithm	N. Predictions	AUC
DIANA-microT	1,434,409	0.500
microCosm	568,103	0.519
miRanda	956,667	0.544
picTar 4-way	56,232	0.509
picTar 5-way	17,226	0.507
PITA All Targets	4,010,550	**0.640**
PITA Top Targets	208,940	0.528
TargetScan Conserved	189,078	0.563
TargetScan Non-Conserved	1,457,487	0.536
RNA22	264,633	0.509
SA-3B	$\sim$ 5 million	0.543
SAWA-3B	$\sim$ 5 million	0.543
SA	$\sim$ 5 million	**0.608**
Our Approach	$\sim$ 5 million	**0.649**

Table 1. Number of predictions and AUC values obtained by each algorithm/approach.

$$weight(x) = \begin{cases} 1.0 & if\ x \in L \\ P(f(x) = 1|l(x) = 0) & if\ x \in U \wedge P(f(x) = 1|l(x) = 0) \geq P(f(x) = 0|l(x) = 0) \\ 1 - P(f(x) = 1|l(x) = 0) & otherwise \end{cases}$$

$f'(p(x))$ is learned by applying a variant of LIBSVM (`www.csie.ntu.edu.tw/~cjlin/libsvmtools/#weights_for_data_instances`) which allows us to specify a weight for each example. In this algorithm, the weight assigned to an example represents the cost of misclassifying it, which is proportional to the confidence we have in the assigned label.

Similarly to the nontraditional classifier, also in this case, we solve the class unbalancing problem (in this case, unbalancing is between positive examples and negative examples, instead of labeled and unlabeled examples), by resorting to the same bagging procedure described in the previous section. The procedure in this case is still necessary since the number of true miRNA:mRNA interactions (positive examples) is significantly smaller than the number of remaining possible miRNA:mRNA pairs (negative examples).

3 Experiments

In order to evaluate our approach, we have considered as data sources a set of experimentally verified miRNA:mRNA interactions, i.e. miRTarBase 3.5 [6] (4,867 interactions), used as positive/labeled examples, as well as the set of miRNA target predictions in mirDIP [13] ($\sim$ 5 million predictions), considered as unlabeled examples. In the learning phase, examples are represented according to the standardized scores returned by the algorithms working on the 3' UTR region, considered by mirDIP (see Table 1). As a testing set, we used TarBase [12], containing 65,000 positive and negative verified interactions. TarBase can also be used in the training phase, but in this work we use it in the testing phase, in order to provide good estimates of the performance on a independent test set.

The main goals of the experiments are *a)* to evaluate the accuracy of predictions provided by our algorithm and *b)* to evaluate whether our algorithm can improve the identification of meaningful regulatory networks with HOCCLUS2 [8]. The complete workflow of the experiments is reported in Figure 3. As regards *a)*, we consider the Area Under the ROC Curve (AUC) [11], which evaluates the

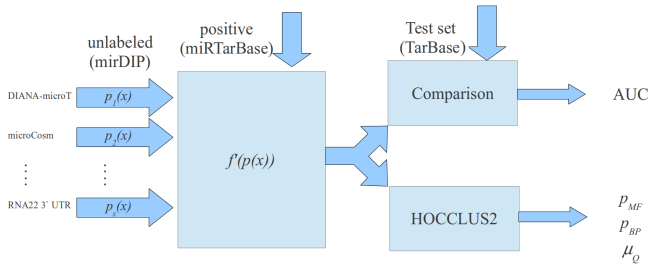

Fig. 1. Experimental settings: evaluation of AUC and of extracted regulatory networks.

classification accuracy independently of the threshold applied on the prediction score. We also compare our approach with several alternative solutions:

- *Single prediction algorithms* (included in mirDIP).
- *Score averaging* (**SA**): a simple algorithm that equally weights the contribution of each single prediction algorithm.
- *Score averaging - three best* (**SA-3B**): an algorithm that equally weights the contribution of the best three prediction algorithms (TargetScan Conserved, PITA Top Hits and picTar 5-way), according to [13].
- *Weighted score averaging - three best* (**WSA-3B**): an algorithm that weights the contribution of the best three prediction algorithms (TargetScan Conserved, PITA Top Hits and picTar 5-way). Weights are proportional to the reliability (on the basis of the F-Score) of each algorithm, according to [13].

As regards *b)*, we apply HOCCLUS2 to the datasets obtained by our approach and with SA, SA-3B and WSA-3B. Experiments are performed with different values of HOCCLUS2 parameters, i.e.: α, which is the minimum cohesiveness value a bicluster must satisfy after merging, and β, which is the minimum score an interaction must have to be considered as reliable (see [8] for details).

The quality of extracted biclusters is evaluated in terms of cohesiveness q and in terms of two p-values (p_{BP} and p_{MF}) obtained by the Student's t-test, which evaluate the biological significance with respect to Gene Ontology. Details about these measures can be found in the original paper [10].

Experiments are performed with $n = 10,000$ and $\gamma = 0.9$, which let us obtain the best results according to a preliminary study (see [10]).

In Table 1 we report AUC results for all the considered algorithms/approaches. They show that results obtained with our approach outperform those obtained with single algorithms. This confirms that combined approaches are, in general, able to outperform single algorithms. The only algorithm which is able to produce comparable results is PITA All Targets, which, however, cannot generate high-quality biclusters, due to the large amount of false positives.

Table 2 shows the quantitative results obtained for the first, the last and the best (according to p_{BP} and p_{MF} values) hierarchy level of regulatory networks extracted by HOCCLUS2. Observing the best level, it is possible to see that the proposed approach always leads to the identification of at least one level with very low p_{BP} and p_{MF} values, independently of the parameters of HOCCLUS2.

α	β	$\mathcal{N}$	level 1				max level					best level				
		(mRNA/miRNA)	#cc	p_{MF}	p_{BP}	μ_q	lev	#cc	p_{MF}	p_{BP}	μ_q	lev	#cc	p_{MF}	p_{BP}	μ_q
Predictions - SA-3B																
0.2	0.3	5698/612	700	1.000	1.000	0.49	7	183	0.000	0.000	0.24	3	210	0.000	0.000	0.31
	0.3						5	355	1.000	1.000	0.36	1	700	1.000	1.000	0.49
0.2	0.4	4735/607	619	1.000	1.000	0.52	7	144	0.006	0.001	0.24	7	144	0.006	0.001	0.24
	0.3						6	274	1.000	1.000	0.35	1	619	1.000	1.000	0.52
0.2	0.5	3337/572	599	1.000	1.000	0.58	7	101	0.315	0.146	0.23	5	108	0.257	0.112	0.26
	0.3						6	202	1.000	0.221	0.34	5	205	1.000	0.206	0.35
Predictions - WSA-3B																
0.2	0.3	6209/618	758	1.000	1.000	0.50	7	194	0.016	0.004	0.25	3	221	0.001	0.000	0.31
	0.3						6	374	1.000	1.000	0.36	1	758	1.000	1.000	0.50
0.2	0.4	5122/601	667	1.000	1.000	0.54	6	145	0.096	0.004	0.24	5	148	0.053	0.004	0.25
	0.3						5	273	1.000	1.000	0.34	1	667	1.000	1.000	0.54
0.2	0.5	3653/570	622	1.000	1.000	0.60	7	105	0.221	1.000	0.24	3	168	0.123	0.298	0.38
	0.3						6	205	0.374	1.000	0.36	2	314	0.256	1.000	0.50
Predictions - SA																
0.2	0.3	8723/599	294	0.140	0.080	0.43	7	58	0.262	0.253	0.22	1	294	0.140	0.080	0.43
	0.3						5	182	0.328	0.176	0.38	1	294	0.140	0.080	0.43
0.2	0.4	7772/620	1608	1.000	1.000	0.50	9	216	0.001	0.006	0.22	3	416	0.008	0.000	0.33
	0.3						7	604	0.000	0.000	0.34	2	830	0.000	0.000	0.42
0.2	0.5	4336/627	1038	1.000	1.000	0.58	9	96	0.399	0.364	0.22	4	148	0.286	0.261	0.31
	0.3						7	283	1.000	1.000	0.35	2	522	1.000	0.228	0.47
Predictions - Our Approach																
0.2	0.3	2379/614	888	1.000	1.000	0.69	7	143	0.000	0.000	0.24	2	444	0.000	0.000	0.52
	0.3						6	268	0.002	0.001	0.37	3	309	0.000	0.000	0.42
0.2	0.4	1626/544	591	0.404	1.000	0.77	7	84	0.015	0.001	0.24	3	152	0.000	0.000	0.39
	0.3						6	161	0.001	0.377	0.38	2	298	0.000	0.001	0.57
0.2	0.5	1245/467	417	0.361	0.244	0.83	7	53	0.000	0.000	0.24	3	105	0.000	0.000	0.43
	0.3						7	104	0.000	0.000	0.39	4	110	0.000	0.000	0.42

Table 2. Quality of biclusters obtained by HOCCLUS2. $\mathcal{N}$ is the number of biclustered mRNAs and miRNAs. #cc is the number of biclusters. *lev* represents the number of hierarchy levels number. The "best" level is that with the lowest $\frac{p_{MF}+p_{BP}}{2}$ value.

Moreover, comparing our results to SA (which is the best among the competitors), it is noteworthy that our approach always lets HOCCLUS2 extract a smaller number of biclusters, grouping less miRNAs and mRNAs. This is due to the fact that our approach is able to better filter out false positives and allows HOCCLUS2 to focus only on reliable interactions (lower FPr, for a given TPr).

As regards the biological significance, the proposed approach allowed us to analyze much wider phenomena than those observable on only experimentally validated interactions. At the same time, it allowed us to focus only on more reliable predictions than those returned by single prediction algorithms and by baseline combination approaches (a detailed analysis can be found in [10]).

The proposed system, the datasets and the obtained results are available at `www.di.uniba.it/~ceci/micFiles/systems/semisupervised_HOCCLUS2/`.

4 Conclusions

In this work we have investigated the possibility to improve the reliability of link predictions algorithms, by means of a semi-supervised ensemble-based machine learning method, which learns to combine the outputs of several prediction algorithms. As an application, we considered that of miRNA:miRNA regulatory

networks, by taking into account all the issues raised by the particular domain. The effectiveness of the proposed approach has been evaluated in terms of predictive performance and in terms of the significance of the interactions networks extracted by HOCCLUS2. Results prove that the proposed approach better filters out false positives and allows HOCCLUS2 to focus on only reliable interactions, leading to the identification of more significant interaction networks. These results give us the opportunity to use HOCCLUS2 for a comprehensive reconstruction of all the possible multiple interactions established by miRNAs when regulating the expression of gene networks, which are otherwise impossible to identify when only experimentally validated interactions are considered.

For future work, we will investigate the integration of low-level features in the learning phase, with the aim of improving the predictive performance.

Acknowledgements

We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

1. Breiman, L.: Bagging predictors. Mach. Learn. 24(2), 123–140 (1996)
2. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 27:1–27:27 (May 2011)
3. Draper, N.R., Smith, H.: Applied Regression Analysis (Wiley Series in Probability and Statistics). Wiley-Interscience, Third edn. (Apr 1998)
4. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman and Hall (1994)
5. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proc. of ACM SIGKDD. pp. 213–220. KDD '08, ACM (2008)
6. Hsu, S.D., Lin, F.M., Wu, W.Y., Liang, C., Huang, W.C., Chan, W.L., Tsai, W.T., Chen, G.Z., Lee, C.J., Chiu, C.M., Chien, C.H., Wu, M.C., Huang, C.Y., Tsou, A.P., Huang, H.D.: mirtarbase: a database curates experimentally validated microrna-target interactions. Nucl Acids Res 39, 163–169 (2011)
7. Newman, M., Watts, D.J.: The structure and dynamics of networks. Princeton University Press (2006)
8. Pio, G., Ceci, M., D'Elia, D., Loglisci, C., Malerba, D.: A novel biclustering algorithm for the discovery of meaningful biological correlations between micrornas and their target genes. BMC Bioinformatics 14(S-7), 8 (2013)
9. Pio, G., Ceci, M., Loglisci, C., D'Elia, D., Malerba, D.: Hierarchical and Overlapping Co-Clustering of mRNA: miRNA Interactions. In: ECAI. Frontiers in Artificial Intelligence and Applications, vol. 242, pp. 654–659. IOS Press (2012)
10. Pio, G., Malerba, D., D'Elia, D., Ceci, M.: Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach. BMC Bioinformatics 15(S-1), S4 (2014)
11. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Mach. Learn. 42(3), 203–231 (Mar 2001)

12. Sethupathy, P., Corda, B., Hatzigeorgiou, A.G.: TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA (New York, N.Y.) 12(2), 192–197 (Feb 2006)
13. Shirdel, E.A., Xie, W., Mak, T.W., Jurisica, I.: Navigating the micronome using multiple microrna prediction databases to identify signalling pathway-associated micrornas. PLoS ONE 6(2), e17429 (02 2011)
14. Zhang, Y., Verbeek, F.J.: Comparison and Integration of target prediction algorithms for microRNA studies. J Integr Bioinform 7(3), 127 (2010)