Gene Network Reconstruction via Transfer Learning across Multiple Organisms

Paolo Mignone, Gianvito Pio and Michelangelo Ceci
{name.surname}@uniba.it

Dept. of Computer Science
University of Bari Aldo Moro (Italy)

1 Introduction

The task of Gene Network Reconstruction (GNR) is receiving increasing attention in the recent years. The main motivation can be found in its usefulness for biologists, who can obtain more information about the various interactions between genes, which are otherwise difficult and expensive to be identified with classical in-lab methods. The task of GNR is usually solved by adopting a graph structure, where nodes are the genes and edges correspond to gene interactions.

State of the art *link prediction* methods, as outlined in [4], can be categorized in *feature-based measures* (e.g., vector-based distances, L-norms, correlation coefficients), *local similarity based measures* (e.g., common neighbors, Jaccard index), *global similarity based measures* (e.g., Kats index [2]) and *quasi-local similarity based measures* (e.g., Local path index [3]). Focusing on machine learning approaches, it is possible to adopt link prediction algorithms to identify possibly unknown interactions, which mainly aim at identifying a binary classifier. Accordingly, several machine learning methods to solve this task have been proposed in the literature (e.g. the methods based on clustering [9], Bayesian networks [6]). However, there is no method that works optimally over all the datasets [4]. In [1, 5] , the authors show that solutions that combine different methods are able to obtain better results over multiple datasets.

2 Motivations and goals

The task of link prediction generally aims at estimating the probability of the existence of a interaction between two entities, about which there is currently no information, on the basis of information available about a set of known interactions. Therefore, a major assumption of classical machine learning algorithms is the availability of training and testing data coming from the same data distribution, about the same context and described according to the same feature space. However, in many real contexts, this assumption is easily violated. In particular, in the study of biological (but also social and technological) networks, collecting training data is very expensive. Therefore, there is the need to build models on the basis of already available data in different, but related contexts.

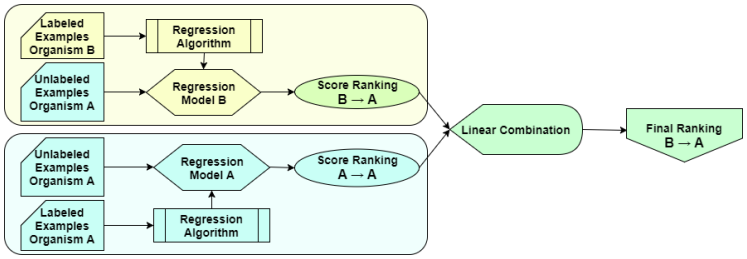

Fig. 1. Workflow of the first approach to transfer the knowledge of the source organism B on the target organism A.

At this aim, *transfer learning* strategies [8] can be exploited to leverage knowledge from a source task in order to improve the performance of a target task for which poor data are available. In [8], the authors distinguish between two main transfer learning settings: *homogeneous transfer* that considers the same feature space in source and in the target tasks, but different marginal data distribution; *heterogeneous transfer* that works on different feature spaces in the source and in the target tasks. In this context, the goal of this research project is to evaluate the possible contribution of transfer learning techniques in the reconstruction of gene networks. In particular, we aim at exploiting available information about an organism for the reconstruction of the gene network of other organisms, for which the available data is poor.

3 Possible approaches

In the first stage of this study, the starting point is the state-of-the-art framework called GENERE [1]. The peculiarities of such a method are: *i)* it is able to learn to combine the outputs of several link prediction methods *(meta learning)*; *ii)* it can learn a prediction model from only positively labeled data, also exploiting a huge amount of unlabeled data *(semi-supervised learning)*; *iii)* it analyzes different views of data iteratively *(iterative multi-view learning)*; *iv)* it predicts the probability of existence of all the unlabeled interactions observed during the training phase *(transductive setting)*. Since GENERE is limited to the transductive setting, the first contribution that can be provided regards the possibility to apply it in an inductive setting, i.e., to predict the existence of interactions that are not available (not known) during the training phase. Indeed, this allows us to learn the model on an organism and apply it to reconstruct the network of another organism. In particular, given the source organism B and the target organism A, we learn a model through GENERE on the organism B, transform it into an inductive model and apply it to reconstruct the network of the organism A. At this aim, we obtain the predictions for each organism through GENERE, consider them as a target attribute for a regression method and learn a new regression model that can be applied on unseen interactions. Finally, we apply the model learned on the organism B to predict the score of

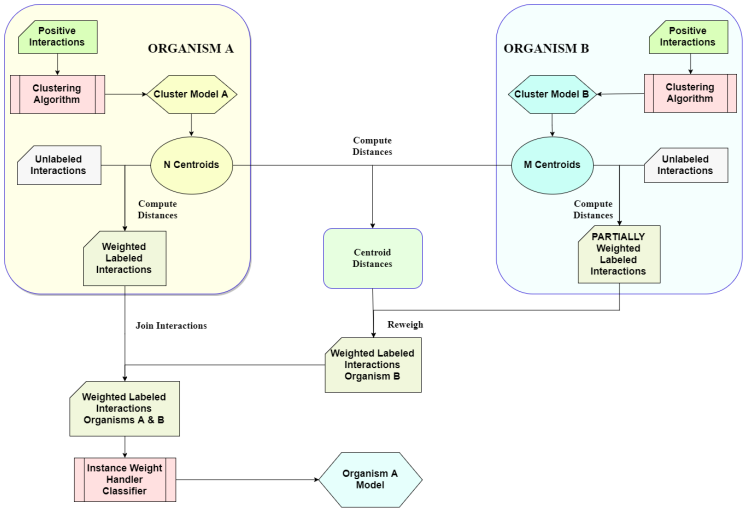

Fig. 2. Workflow of the transfer learning approach based on clustering.

unlabeled interactions of the target organism A. Obviously, this solution requires a way to weight the contribution provided by the transferred knowledge, combining linearly the scores obtained on both the source and target organisms (see Figure 1).

A more sophisticated approach, which would appear almost independent of the framework GENERE, consists in the exploitation of clustering methods to identify the semantic distance between the source and the target organisms. In particular, we can apply a clustering algorithm on the positive (i.e., validated) gene interactions on both the organisms, separately, to obtain n and m clusters respectively. In a second stage, we can weight the unlabeled interactions of the source organism B according to their distance with respect to the centroids of the organism B as well as according to the distance between the organism A and the organism B. Finally, a classification method which handles weights on the instances can be exploited to learn the final model for the target organism A. Intuitively, this approach would lead to learn a model which fully exploits the instances from the target organism A and partially exploits the instances from the source organism B, weighted according to the distance between the organisms. A sketch of such a workflow can be observed in Figure 2.

4 Preliminary experiments: Proof of Concept

These approaches can be evaluated on the three datasets of the *DREAM5 challenge*, which are: In Silico (network 1), Escherichia Coli (network 3) and Brewer's Yeast (network 4). Some preliminary experiments have been performed to evaluate the first approach (based on linear combination). In Figure 3, we can observe

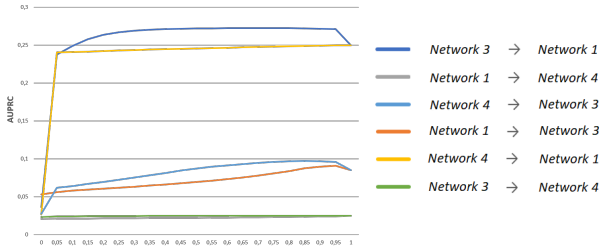

Fig. 3. Obtained AUPRC results. On the X-axis, there are the different values for the weight of the contribution of the target organism.

the results, in terms of Area Under the Precision Recall Curve (AUPRC), obtained by combining the contribution of the models pairwisely learned on the considered organisms. The adopted regression method is based on a multilayer perceptron neural network with backpropagation [7]. As we can observe from Figure 3, in some cases we have a positive transfer, i.e., the obtained result is better than that obtained by considering only the target organism. This confirms that the application of transfer learning techniques to GNR deserves a deeper investigation. We are currently performing experiments to evaluate the effectiveness of the clustering-based approach through the Apache Spark framework, in order to be able to process large-scale datasets in a distributed fashion.

References

1. M. Ceci, G. Pio, V. Kuzmanovski, and S. Džeroski. Semi-supervised multi-view learning for gene network reconstruction. *PLOS ONE*, 10(12):1–27, 12 2015.
2. L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.
3. L. Lü, C.-H. Jin, and T. Zhou. Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E*, 80:046122, Oct 2009.
4. L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.
5. D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nat Methods*, 9:796–804, 2012.
6. S. H. Shalforoushan and M. Jalali. Link prediction in social networks using bayesian networks. In *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pages 246–250, March 2015.
7. M. Ware. The weka multilayer-perceptron. *Data Mining: Practical Machine Learning Tools and Techniques*, 2000.
8. K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016.
9. M. Zhu, T. Cao, and X. Jiang. Using clustering coefficient to construct weighted networks for supervised link prediction. *Social Netw. Analys. Mining*, 4(1):215, 2014.