A Machine Learning Approach for Gene Regulatory Network Inference

Meroua Daoudi[1, 2,*], Souham Meshoul[1], and Fariza Tahi[3]

[1] Computer Science Dept, Constantine 2 University, Algeria.
[2] MISC Laboratory, Constantine 2 University, Algeria.
[3] IBISC/CNRS Lab Evry, 91000 France.

* Corresponding author. Tel.: 213 665 27 86 52; email: meroua.daoudi@univ-constantine2.dz
Manuscript submitted September 26, 2018; accepted January 12, 2019.
doi: 10.17706/ijbbb.2019.9.2.82-89

Abstract: Transcription factors are key elements in the regulation of genetic expressions. Understanding the behavior of the system is the ultimate goal behind modeling biology networks including gene regulatory networks. Prediction of regulation relationship between a transcription factor and a target gene can be viewed as a machine learning problem. Within this perspective, several algorithms have been developed to solve this problem using primordially support vector machine. In this work, we propose a semi-supervised approach to infer gene regulatory networks using both unsupervised and supervised techniques. First a set of reliable negative examples are extracted using clustering's techniques. Then we use this set for classification using SVM. We have applied the proposed method to simulated data of Escherichia Coli, the experimental results show the competitiveness of the proposed approach as the prediction accuracy of the most tested case is achieved to the best desired value.

Key words: Gene regulatory network, machine learning, semi-supervised learning, supervised learning, unsupervised learning.

1. Introduction

The goal of most recent studies of system biology is to model, simulate and identify the interaction networks of components. Gene Regulatory Networks (GRN) are essential for the control of cellular metabolism and understanding these networks plays a key role in solving many biological problems such as complex disease diagnosis and drug discovery. Using experimental methods to determine regulation between transcription factor (TF) and the target gene is costly and expensive in time [1]. In the other hand, high throughput technologies generate massive expression data. Which presents an opportunity to infer gene regulatory network using appropriate computational methods.

Inferring gene regulatory network from gene expression data is a machine learning problem. Several approaches have been proposed to learn structure and finding an accurate and reliable model of GRN [2]. Including unsupervised, supervised and semi-supervised methods. Unsupervised models can be classified generally into four categories including Boolean network, differential network, Bayesian network and co-expression network. Unsupervised methods have been widely used due the limitation of information about the regulation between transcription factors and genes. The advantage of these methods is that they don't need any farther information about the biological system but they are less efficient. Recently due to the identification of a large number of interactions between the transcription factor and target gene,

several supervised approaches have been proposed such as Sirene [3] and CompareSVM [4]. Sirene decomposes the problem of gene regulatory network into a large number of local models were each local model is associated to one TF, then SVM classifier is used to discriminate between regulated genes and non-regulated ones given a TF. CompareSVM is a tool based on SVM to compare different kernel methods for GRN inference. Supervised methods are more accurate but they depend on the amount and quality of known interaction.

Semi-supervised methods can handle large numbers of unlabeled data and take advantage from known interactions. There are two approaches in supervised learning 1) learning from unlabeled and positive data and 2) learning from only positive. A variety of approaches have been proposed, in this context Nihil Patel *et al*. in [1] propose an iterative approach to learn from unlabeled and positive data using two classification models i.e SVM and Random forest. Other approaches to extract reliable negative example from unlabeled data are proposed in [5], [6], the two approaches use K-means clustering and SVM classifier. SVM is widely used due to its efficiency in gene expression classification.

In this work, we propose a new approach to select reliable negative from unlabeled data examples based the distribution of data using K-means algorithm and SVM classifier to perform a semi-supervised learning method to infer gene regulatory network.

The rest of paper is structured as follow.in the next section we describe related works, then we present our proposed approach, in Experimental results section we discuss obtained results and we conclude the paper in conclusion section

2. Related Work

Several approaches based on supervised, unsupervised or semi-supervised paradigms have been proposed to infer gene regulatory networks from gene expression data.

The class of unsupervised approaches includes: Relevance Network (RN), Context likelihood of relatedness (CLR), Algorithm for the reconstruction of gene regulatory networks (ARACNE) and Mutual information relevance network termed as MRMR. Relevance Network [7] calculates mutual information and supposes that two genes are related if their MI is above a certain threshold. CLR [8], is an extension of relevance network algorithm to infer gene regulatory networks, it uses mutual information theory, after the calculation of mutual information between TFs and target genes, CLR applies an adaptive background correction step to eliminate false correlation and indirect influences. ARACNE [9] is based on information theoretic approach to eliminate the majority of indirect interactions inferred by co-expression methods. MRNET [10] principle suggests to select among the least redundant variables, the one that has the highest mutual information with the target gene. MRMR extends this feature selection principle to infer relationships between genes.

Regarding the class of supervised approaches, Sirene is one of the state of the art of methods and requires two types of data: gene expression and some known interaction of TF. Sirene decomposes the problem of gene regulatory network inference into a large number of binary classification problems, each subproblem is associated to one TF, and an SVM is used to predict the GRN. In [11], authors present a method for the selection of reliable negative examples, where a GRN is viewed as a graph for which a machine learning scheme can be used to infer the unknown gene regulatory connections.

Compare SVM is a tool that compares four SVM kernel functions namely linear, Gaussian, sigmoid and polynomial kernels and includes three steps. The first one is devoted to the optimization of parameters of each kernel. Once parameters are optimized, CompareSVM comparison can generate area under a curve of kernels. In the third step, the kernel with high accuracy and its optimized parameters is used in the prediction step. By another side, in [1] the authors propose a semi-supervised approach to learn from

positive and unlabeled data. An iterative approach using random Forest (RF) and SVM is applied to predict regulation of each TF. The classifier use predicted negative example from the previous iteration and the half of positive examples. The other half is used in the validation step; the model obtained at the end is then used to predict labels for the remaining unlabeled data in the testing dataset. From their side, Augustine *et al*. in [5] applied k-means on both positive and unlabeled data and consider obtained clusters without positive examples as negative training datasets. Then, an iterative procedure is applied until convergence of the algorithm. Another semi-supervised approach is proposed by Cerulo *et al*. [12] using only positive data. This method works under the assumption that all the positive examples are randomly sampled from a uniform distribution. Maetschke *et al.* in [13] present a detailed comparison of three previous fields of machine learning and show that supervised and semi-supervised approaches can be trained efficiently even when only a portion of interaction is known. Semi supervised classification seems a very promising way to achieve GRN inference. Within this context we propose in this paper a semi supervised approach that will be emphasized in the next section.

3. The Proposed Semi Supervised Approach for GRN Inference

As mentioned above, semi supervised learning can be achieved using either learning from only positive data or learning from positive and unlabeled Data. In this study, we propose a novel method to extract reliable negative example from both positive and unlabeled data to perform a semi-supervised learning for gene regulatory network inference. The key features of the proposed approach can be summarized in the following points. Kmeans algorithm is used as a first step to group genes into clusters. As k-means does not necessarily provide the optimal partition of genes we make the assumption that nearest genes to the center are correctly partitioned and we take into consideration the sensitivity of kmeans to outliers [14] and consider outliers and take them as negative examples. Negative examples are used with the half of positive examples to train the model. We run the algorithm with different values of k ranging from two to the number of positive examples we select the best obtained model.

The proposed approach can be outlined by the algorithm provided below. Let's first present some adopted notations for a better understanding of the algorithm. Let:

- C: represents a given clustering where C = {c1,c2,……..ck} where each ci denotes a cluster.
- k: number of clusters
- RNi: number of reliable negative example to be extracted
- pos : positive example in each cluster
- pi : positive examples from each cluster used in training step
- vi: positive examples from each cluster used in validation step

First, the algorithm starts with the application of K-means clustering on the positive and unlabeled dataset. For each cluster Ci We compute the number RN of extracted Reliable negatives example from each cluster witch is the number of farthest RN data from each center where:

$$RN = \frac{Npos}{k} \quad (1)$$

Npos is the half of the number of positive examples and k is the number of clusters

The positive examples in each cluster are divided into two equal parts in a random manner. The first half pi is added to the positive examples used for training and the other half vi is added to validation datasets. Then the remaining datasets of each cluster are added to the testing data.

We note that the number of negative examples TN is the sum of RN from all clusters where:

$$|TN| = k * \frac{Npos}{k} = Npos \quad (2)$$

Algorithm: Semi Supervised inference of GRN

Input : positive and unlabeled Data D
Initialize *TN, TP, V,* and *T* as empty sets
C=kmeans(D,k)
For *k* in range [2..number of positive examples]
For each cluster *ci* :
select *RNi* farthest point from each cluster
select positive examples *posi*
(pi,vi) = split positive Data into two equal parts
TP= TP ∪ *pi*
V=V ∪ *vi*
TN=TN ∪ *RNi*
T= T – (TP ∪ *V* ∪ *TN)*
modeli =TrainSVM(TP,TN)
accurayi = Validate the model with *V*
If (*accuracyi* > *accuracyi-1*)
Bestmodel = modeli
Classify *T* with *Bestmodel*
Output : regulation between TF and target genes

In this case we obtain a balanced binary classification with the number of TP is equals to the number of true negative TN.

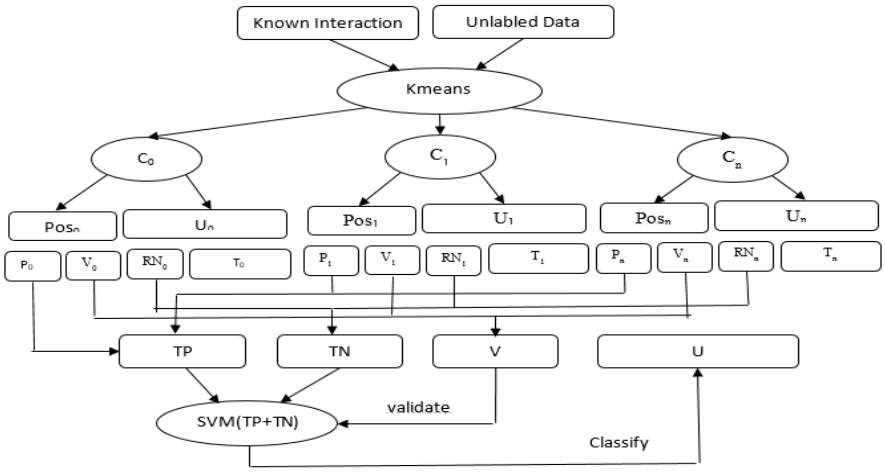

Fig. 1. Outline of the proposed approach.

RN is added to TN and they are used in the training step with the true positive TP, the remaining data from each cluster after the extraction of RN are the testing datasets T. Then, a SVM classifier is applied with the training data set to discriminate between gene known to be regulated by a TF and the others. Fig. 1 provides a graphical representation of the proposed approach and helps better understanding of its principle.

4. Results

We evaluated the proposed approach using microarray simulated Data of Ecoli and we consider the three types of data multifactorial, knockdown and knockouts with size 150 simulated by GeneNetWeaver [15]. In the multifactorial experiment a small number of genes are perturbed, a knockdown of every gene is simulated by reducing transcription rate of the corresponding gene by the half, knockout is simulated by setting the transcription rate of the corresponding gene to zero. We applied our approach to each selected TF separately and we selected the four genes having the maximum number of known interaction lexA, rcsB, rcsA , and cpxR.

Table 1. The Arrangement of Channels

TF	Known interaction	k	Multifactorial	knockdown	knockouts
LexA	54	12	1	0.92	0.96
		27	1	1	1
		39	0.92	1	0.96
cpxR	53	7	0.96	0.96	0.80
		22	1	0.80	0.84
		52	0.80	0.73	1
rcsB	21	10	0.90	0.80	1
		15	1	0.70	1
		16	0.80	0.90	0.8
rcsA	20	10	0.90	0.90	1
		15	0.80	0.90	0.90
		17	0.80	0.60	1

Where the prediction accuracy is calculated by dividing the number of correctly predicted examples in the validation data set by the total number of the examples in the validation data set.

To select the best model, we run the algorithm with different numbers of clusters that is different values of k. In Table 1, we show the obtained results with some different numbers of clusters k. The model with the best accuracy is selected and used to classify the testing datasets. The results show that the best accuracy achieved for each data is equal to 1 in most of cases.

The plots of accuracy prediction for each transcription factor including lexA , cpxR, rcsB and rcsA with all used numbers of clusters k are shown in the following figures :

The plots show that both the number k and quality of used Data has an influence at the prediction accuracy of the model.

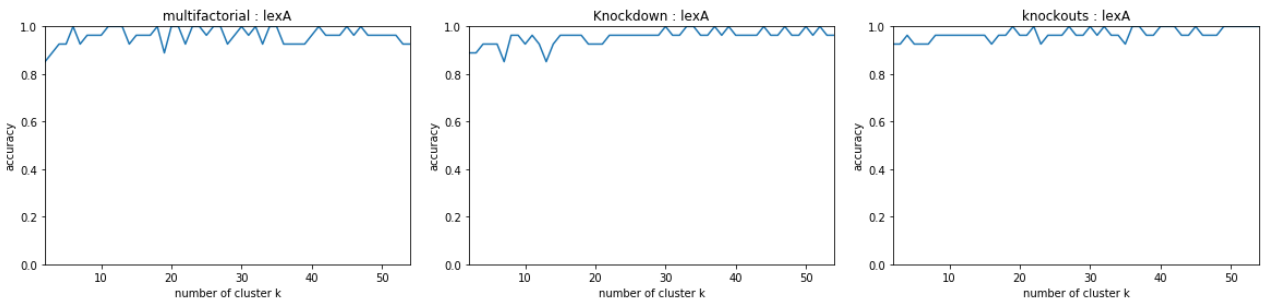

Fig. 2. Transcription factor LexA.

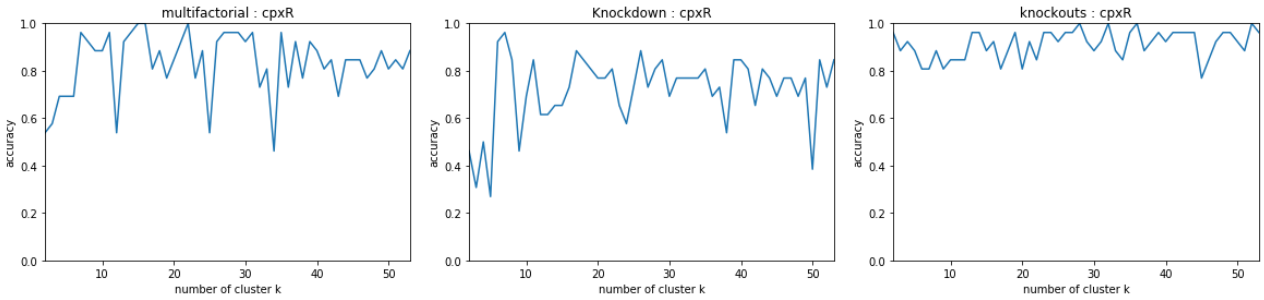

Fig. 3. Transcription factor cpxR.

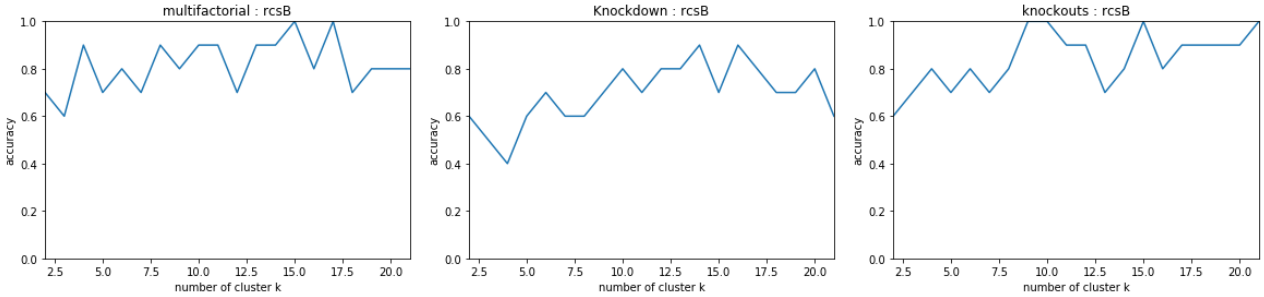

Fig. 4. Transcription factor rcsB.

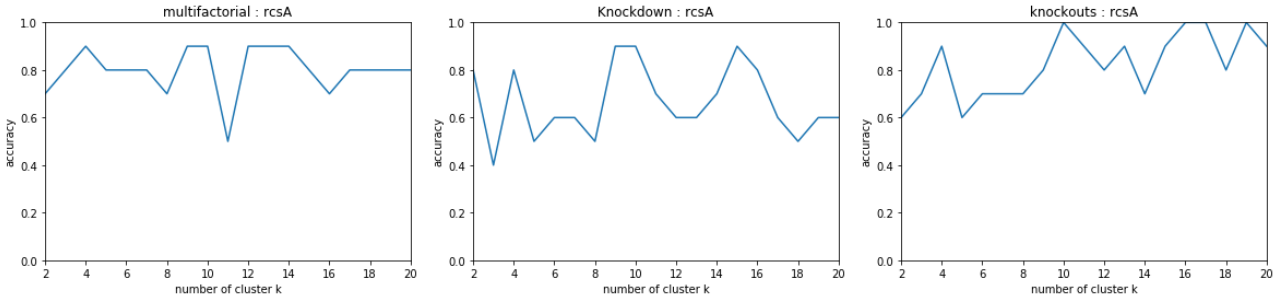

Fig. 5. Transcription factor rcsA.

5. Conclusion

Several approaches based on unsupervised, supervised and semi supervised learning paradigms have been proposed to infer gene regulatory network, unsupervised methods use only gene expression data but they less efficient. Supervised methods require gene expression data and some known regulation both positive and negative. In biology only positive example are known. Then semi supervised methods are the promising way to infer GRN. In this work we have proposed a semi-supervised approach to infer gene regulatory network from unlabeled and positive data, we have tested our approach using four transcriptions having the maximum number of iteration and very encouraging results have been obtained.

References

[1] Nihir, P., & Jason, T. L. W. (2015). Semi-supervised prediction of gene regulatory networks using machine learning algorithms. *J. Biosci*, *40(4)*, 731-740.

[2] Blagoj, R. A survey of models for inference of gene regulatory networks. *Nonlinear Analysis: Modelling and Control*, *18(4)*, 444-465.

[3] Fantine, M., & Vert, J.-P. (2008). Sirene: Supervised inference of regulatory networks. *Bioinformatics*, *4*,

76-82.

[4] Zeeshan, G., *et al*. (2014). Supervised support vector machine (SVM) inference of gene regularity network. *BMC Bioinformatics*.

[5] Jisha, A., & Jereech, A. S. (2017). Gene regulatory network: A semi supervised approach. *Proceedings of International Conference on Electronics Communication and Aerospace Technology ICECA*.

[6] Sasmita, R., *et al*. Handling unlabeled data in gene regulatory network. *Proceeding of International Conference on Frontiers of Intelligence Computing AISC 199* (pp. 113-120).

[7] Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks, functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 418–429.

[8] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, *5(1)*, e8

[9] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., & Dalla, F. R. C. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7(1)*, S7.

[10] Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, *(1)79*, 879.

[11] Ceccarelli, M., Cerulo, L. (2009): Selection of negative example in learning gene regulatory networks. *Proceedings of International Conference on Bioinformatics, Biomedicine Workshop* (pp. 56-61).

[12] Luigi, C. *et al*. (2010). Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics.*

[13] Stefan, S., & Maetschke *et al*. Supervised, semi supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, *15(2)*, 195-211.

[14] Jelili, O. *et al*. (2016). Clustering algorithms : Their application to gene expression data. *Bioinform Biol Insights*.

[15] Shaffer, T. *et al*. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 2263-2011.

Meroua Daoudi was born in 1988. She received the Dipl. master degree from Mentouri Constantine University (UMC). She is also currently pursuing the PhD at Constantine2 Abdelhamid Mehri University. Now her research areas are machine learning, bioinformatics, big data, and optimization

Souham Meshoul received the state engineer degree, MS degree and state doctorate degree from Mentouri Constantine University (UMC). Currently she is a full professor at Constantine2 Abdelhamid Mehri University where she also serves as head of the NTIC college scientific council. She has been involved in many research projects in Algeria and abroad in Europe and Kingdom of Saudi Arabia. Her research interests span several areas namely computational intelligence, optimization, data mining, big data analytics, machine learning with applications to bioinformatics, image analysis and biometrics.

Fariza Tahi is an associate professor at University of Evry, Paris-Saclay. She is a member of IBISC (Informatics, Bioinformatics and Complex Systems) laboratory. Her research work is mainly related to RNA bioinformatics. She is interested by the development of original methods for predicting the structure of RNA, and for identifying and analysing non-coding RNAs from genomic and transcriptomic sequences. She supervised the development of several algorithms and tools dedicated to RNAs, all published in international scientific journals. The different tools are made available to the scientific community through the software platform EvryRNA.