Handling Unlabeled Data in Gene Regulatory Network

Sasmita Rout, Tripti Swarnkar, Saswati Mahapatra, and Debabrata Senapati

Department of Computer Applications, ITER,
SOA University, Bhubaneswar, India
{rout_mca_sasmita,s_aswati}@yahoo.co.in,
tripti_sarap@yahoo.com,
debabratasenapati@gmail.com

Abstract. A gene is treated as a unit of heredity in a living organism. It resides on a stretch of DNA. Gene Regulatory Network (GRN) is a network of transcription dependency among genes of an organism. A GRN can be inferred from microarray data either by unsupervised or by supervised approach. It has been observed that supervised methods yields more accurate result as compared to unsupervised methods. Supervised methods require both positive and negative data for training. In Biological literature only positive example is available as Biologist are unable to state whether two genes are not interacting. A common adopted solution is to consider a random subset of unlabeled example as negative. Random selection may degrade the performance of the classifier. It is usually expected that, when labeled data are limited, the learning performance can be improved by exploiting unlabeled data. In this paper we propose a novel approach to filter out reliable and strong negative data from unlabeled data, so that a supervised model can be trained properly. We tested this method for predicting regulation in E. Coli and observed better result as compared to other unsupervised and supervised methods. This method is based on the principle of dividing the whole domain into gene clusters and then finds the best informative cluster for further classification.

Keywords: Gene, Gene Regulatory Network, Unlabeled data, SVM, K Means, Cluster, Transcription Factor.

1 Introduction

A gene is a unit of heredity of a living organism which resides on a stretch of DNA. All living organism depend on genes, as they specify all proteins and functional RNA chains. In other way a gene is "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and other functional sequence regions". Gene regulatory networks (GRN) [1] explicitly represent the causality of developmental processes. They explain exactly how genomic sequence encodes the regulation

S.C. Satapathy et al. (Eds.): Proc. of Int. Conf. on Front. of Intell. Comput., AISC 199, pp. 113–120.
DOI: 10.1007/978-3-642-35314-7_14

of expression of the sets of genes that progressively generate developmental patterns and execute the construction of multiple states of differentiation. These are inhomogeneous compositions of different kinds of sub circuits, each performing a specific kind of function. This concept is important, because it holds the key to network design principles. Better understanding of the complexity of interdependencies among gene up and down regulation helps in inferring GRN. Different model architectures to reverse engineer gene regulatory networks from gene expression data have been proposed in literature [2]. These models represent biological regulations as a network genes, proteins etc and edges represents the presence of interaction activities between such network components. Four main network models based on unsupervised method can be distinguished: such as information theory models, Boolean network models, Differential and difference equation model and Bayesian models. Information theory model correlates two genes by means of a correlation coefficient and a threshold. Two genes are predicted to interact if the correlation coefficient of their expression levels is above a threshold. For example, TD-ARACNE [3], ARACNE [4] etc. infer the network structure. Boolean network model uses a binary variable to represent the state of a gene activity and a directed graph; here edges are represented by boolean functions to represent the interactions between genes. For example REVEAL [5] infers boolean network model from gene expression data. Differential and difference equation [6] describes gene expression changes as a function of the expression level of other genes. Bayesian model makes use of Bayes rules and consider gene expressions as random variables. The major advantage is that the Bayesian framework allows combining different types of data and prior knowledge in gene networks inference [7]. Just like unsupervised method, recently different supervised methods are also used to find the gene regulatory network. But in this approach unlike unsupervised method, it requires not only gene expression data but also a list of known regulation relationship. The following table lists some of the supervised and unsupervised methods. The basic principle to predict new regulations is: if a gene X having expression profile ep(X) is known to regulate a gene Y with expression profile ep(Y), then all other couples of genes A and B, having respectively expression profiles similar to ep(X) and ep(Y) are likely to interact. Expression profiles are taken as the feature vectors in the machine learning algorithm, while the result is a binary variable representing whether two genes interact or not.

Table 1. Methods under Unsupervised and Supervised approach

Unsupervised Approach	**Supervised Approach**
Information Theory Model	Decision Tree
Boolean Networks	SVM
Ordinary Differential Equation	Neural Network

It has been observed that supervised method give more accurate result as compared to unsupervised methods. Supervised methods require both genes and their complete linkage for their training. But in Biology literature only positive data is available as Biologist only able to tell which are interacting, i.e. Biological databases lists only interacting genes, it does not provide any genes information regarding non-interacting genes, which is a great challenge in finding gene regulatory network through supervised approach.

2 Related Work

2.1 Gene Regulatory Networks

Selection of Reliable Negatives: In [8] the authors tried to predict non-coding RNA genes, where the first set of negative examples is built by maximizing the distances of negative sample points to the known positive sample points by using a distance metric built upon the RNA sequence. Such a negative set is iteratively refined by using a binary classifier based on current positive and negative examples until no further additional negative examples can be found. In [9] they proposed a method applied to gene regulatory network, which selects a reliable set of negatives by exploiting the known network topology.

Probability Estimate Correction: PosOnly method: In paper [10], the conditional probabilities produced by a model trained on the labeled and unlabeled examples differ by only a constant factor from the conditional probabilities produced by a model trained on fully labeled positive and negative examples. Such result can be used to learn a probabilistic binary classifier, such as SVM (Support Vector Machine) with Platt scaling [11], using only positive and unlabeled data.

PSEUDO-RANDOM Method: In paper [9], a gene interaction network is modeled as a directed graph $< G, E >$ where G represents the genes, and E represents the set of directed interactions between genes. Let P be the known gene-gene interactions in E, then Q = E - P the unknown regulatory links, and N=Complement(E) the edges not contained in E. The unknown gene regulatory connections Q can be inferred by a machine learning scheme trained with the set of known regulatory connections. Precisely, P is the set of known positive examples, N is the set of all unknown negative examples and Q is the set of unknown positive examples. A selection of reliable negatives approach selects, from the unlabeled set $N \cup S$ of unknown connections, a subset of reliable negative examples $S \cong N$ and $S \cap Q$ which should be as much as possible composed of negative examples, i.e. and . Such negative examples are used to improve the training phase of a classifier. The PSEUDO-RANDOM method is built over the assumption that a regulatory network has no or few cycles and that it has a tree like structure. For complex eukaryote organisms such an assumption may not be true as many complex cell functions are based on homeostasis and feedback loops.

In contrast, for simpler including Escherichia coli and Saccharomyces cerevisiae, such an assumption may be correct: there are unsupervised approaches, such as ARACNE, that prune the final network by removing 3-arc cycles [3]. This leads to an heuristic that selects as candidate negatives those given by the union of the transitive closure of the known network and its transpose.
S = TC(P) ∪ Transpose (TC(P)) ∪ Transpose(P)

SIRENE: SIRENE (Supervised Inference of Regulatory Networks) [12] is a method to infer gene regulatory networks on a genome scale from a compendium of gene expression data. SIRENE differs from other approaches in that it requires not only gene expression data, but also a list of known regulation relationships both interacting and non-interacting. The authors used Support Vector Machine algorithm for predicting gene regulatory network.

2.2 Text Mining

In traditional text classification, a classifier is built using labeled training documents of every class. In paper [13], Given a set P of documents of a particular class (called positive class) and a set U of unlabeled documents that contains documents from class P and also other types of documents , called negative class documents, the authors build a classifier to classify the documents in U into documents from P and documents not from P. The key feature of the problem is that there is no labeled negative document, which makes traditional text classification techniques inapplicable. In this paper, the author proposed an effective technique to solve the problem. It combines the Rocchio method and the SVM technique for classifier building. Experimental results show that this method outperforms existing methods significantly.

3 Proposed Model

This is a general method for extracting strong reliable negative data for training the supervised model. As it has been already discussed that, GRN can be inferred from microarray data either by unsupervised or by supervised approach. It has been observed that supervised methods yields more accurate result as compared to unsupervised methods. Supervised methods require both positive and negative data for training. In Biological literature only positive example is available as Biologist are unable to state whether two genes are not interacting. A common adopted solution is to consider a random subset of unlabeled example as negative. Random selection may degrade the performance of the classifier. It is usually expected that, when labeled data are limited, the learning performance can be improved by exploiting unlabeled data. As shown in figure 2, p is the set of known interactions and U is unknown (both interacting and non-interacting). Traditionally, while training a supervised model, a random subset of U is taken for negative data, which used to degrade the performance of the classifier as while doing random selection some positive example from Q might be taken as negative.

3.1 Data

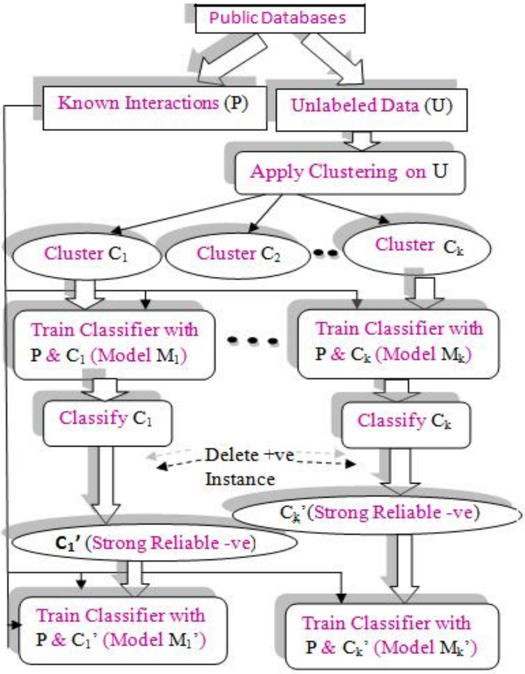

Public Databases
Known Interactions (P)
Unlabeled Data (U)
Apply Clustering on U
Cluster C_1
Cluster C_2
Cluster C_k
Train Classifier with P & C_1 (Model M_1)
Train Classifier with P & C_k (Model M_k)
Classify C_1
Classify C_k
Delete +ve Instance
C_k'(Strong Reliable -ve)
C_1' (Strong Reliable -ve)
Train Classifier with P & C_1' (Model M_1')
Train Classifier with P & C_k' (Model M_k')

Fig. 1.

In our experiment, we used the expression and regulation data of E. Coli, which is publicly available in [14]. The expression data consist of a compendium of 445 E.coli microarray expressions profiles for 4345 genes. The microarrays were collected under different experimental conditions such as growth phases, antibiotics, different media, numerous genetic perturbations and varying oxygen concentrations. The regulation data consist of 3293 experimentally confirmed regulations between 154 TF and 1164 genes, extracted from the RegulonDB database [15].

3.2 Algorithm

```
Step 1 Consider the available interacting genes as true positive ($P$) and unlabeled genes as $U$
Step 2 Apply K-Means on $U$ to build $k$ number of clusters ($C_1, C_2, ...C_k$)
Step 3 for $i = 1$ to $k$ do
Step 3.1 Train model $M_i$ with $P$ and $C_i$
Step 3.2 Classify $C_i$ itself with model $M_i$
```

Step 3.3 P=Performance of M_i
Step 3.4 Delete Positive examples from C_i if any
Step 3.5 Train classifier M_i* with P and the remaining instances of C_i i.e. C_i*
Step 3.6 $P*$=Performance of M_i*
Step 3.7 Compare P and $P*$

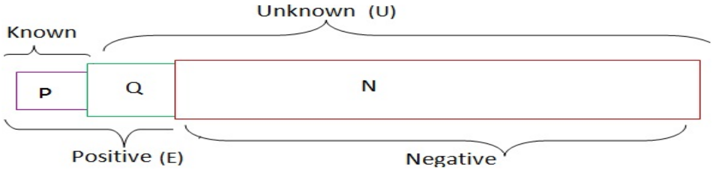


Fig. 2.

3.3 Experimental Result

The experiment is performed on those Transcription Factors (TF) having more than 50 interactions, such as crp (900), fis (1166), fnr (1218), himD (1451), rpoD (2307) etc. where each TF is associated with an unique number. We run the algorithm for each TF and observed that the performance of the classifier after removing the supposed to be positive example is better than the classifier taken earlier. It has been observed that irrespective of the number of cluster in K-means , the correct rate of almost all cluster (after removing the +ve instances) are better than the earlier model which has been shown in figure 3a. Figure 3b shows the classifiers in the ROC space. The classifiers performances are measured for both k=10 and k=15. And it has been observed that the performance is good irrespective of the number of cluster. But we have shown the results only for k=10. We have taken SVM [16] as the classifier for each cluster. The correct rate of different TF is shown in Figure 4. a and b.

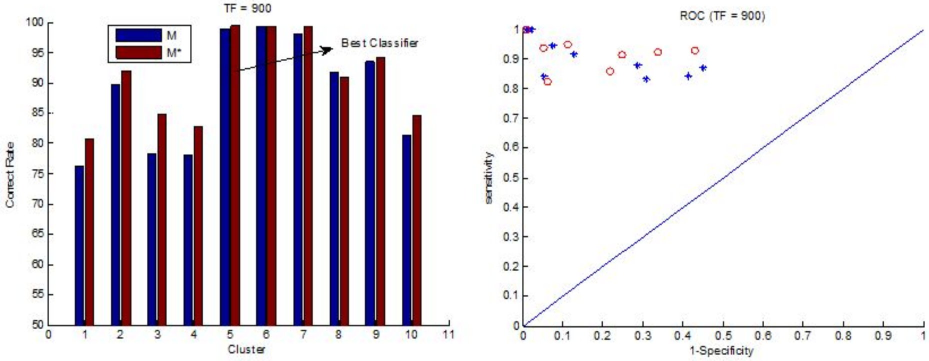
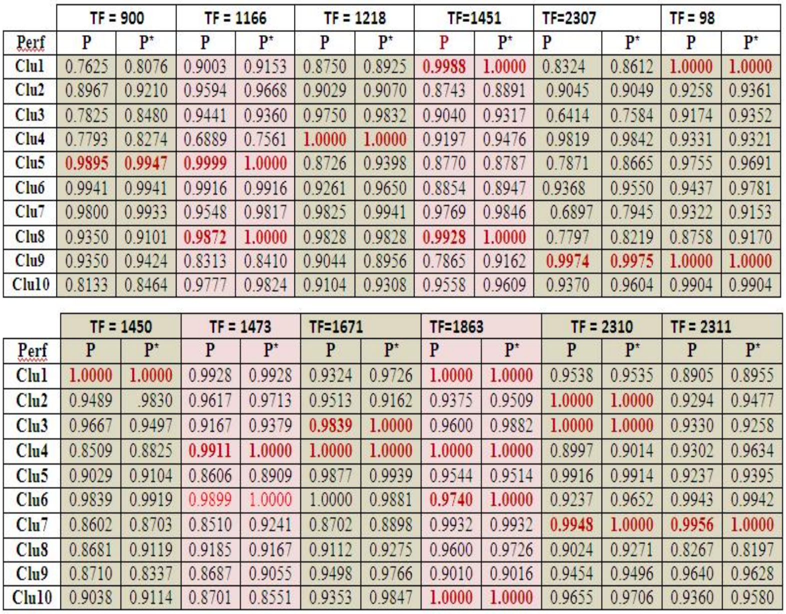


Fig. 3. a,b



	TF = 900		TF = 1166		TF = 1218		TF=1451		TF=2307		TF = 98	
Perf	P	P^*	P	P^*	P	P^*	P	P^*	P	P^*	P	P^*
Clu1	0.7625	0.8076	0.9003	0.9153	0.8750	0.8925	0.9988	1.0000	0.8324	0.8612	1.0000	1.0000
Clu2	0.8967	0.9210	0.9594	0.9668	0.9029	0.9070	0.8743	0.8891	0.9045	0.9049	0.9258	0.9361
Clu3	0.7825	0.8480	0.9441	0.9360	0.9750	0.9832	0.9040	0.9317	0.6414	0.7584	0.9174	0.9352
Clu4	0.7793	0.8274	0.6889	0.7561	1.0000	1.0000	0.9197	0.9476	0.9819	0.9842	0.9331	0.9321
Clu5	0.9895	0.9947	0.9999	1.0000	0.8726	0.9398	0.8770	0.8787	0.7871	0.8665	0.9755	0.9691
Clu6	0.9941	0.9941	0.9916	0.9916	0.9261	0.9650	0.8854	0.8947	0.9368	0.9550	0.9437	0.9781
Clu7	0.9800	0.9933	0.9548	0.9817	0.9825	0.9941	0.9769	0.9846	0.6897	0.7945	0.9322	0.9153
Clu8	0.9350	0.9101	0.9872	1.0000	0.9828	0.9828	0.9928	1.0000	0.7797	0.8219	0.8758	0.9170
Clu9	0.9350	0.9424	0.8313	0.8410	0.9044	0.8956	0.7865	0.9162	0.9974	0.9975	1.0000	1.0000
Clu10	0.8133	0.8464	0.9777	0.9824	0.9104	0.9308	0.9558	0.9609	0.9370	0.9604	0.9904	0.9904

	TF = 1450		TF = 1473		TF=1671		TF=1863		TF = 2310		TF = 2311	
Perf	P	P^*	P	P^*	P	P^*	P	P^*	P	P^*	P	P^*
Clu1	1.0000	1.0000	0.9928	0.9928	0.9324	0.9726	1.0000	1.0000	0.9538	0.9535	0.8905	0.8955
Clu2	0.9489	.9830	0.9617	0.9713	0.9513	0.9162	0.9375	0.9509	1.0000	1.0000	0.9294	0.9477
Clu3	0.9667	0.9497	0.9167	0.9379	0.9839	1.0000	0.9600	0.9882	1.0000	1.0000	0.9330	0.9258
Clu4	0.8509	0.8825	0.9911	1.0000	1.0000	1.0000	1.0000	1.0000	0.8997	0.9014	0.9302	0.9634
Clu5	0.9029	0.9104	0.8606	0.8909	0.9877	0.9939	0.9544	0.9514	0.9916	0.9914	0.9237	0.9395
Clu6	0.9839	0.9919	0.9899	1.0000	1.0000	0.9881	0.9740	1.0000	0.9237	0.9652	0.9943	0.9942
Clu7	0.8602	0.8703	0.8510	0.9241	0.8702	0.8898	0.9932	0.9932	0.9948	1.0000	0.9956	1.0000
Clu8	0.8681	0.9119	0.9185	0.9167	0.9112	0.9275	0.9600	0.9726	0.9024	0.9271	0.8267	0.8197
Clu9	0.8710	0.8337	0.8687	0.9055	0.9498	0.9766	0.9010	0.9016	0.9454	0.9496	0.9640	0.9628
Clu10	0.9038	0.9114	0.8701	0.8551	0.9353	0.9847	1.0000	1.0000	0.9655	0.9706	0.9360	0.9580

Fig. 4. a, b

4 Conclusion

Supervised methods always need a complete set of known regulatory networks i.e. gene expression data and list of known regulation relationship both interacting and non-interacting. But In Biology literature only positive examples are available, as Biologists do not have idea about the genes which are not interacting. That means only positive examples are available. So a common adopted solution is to consider all or a random subset of unlabeled example as negative, for the training of a supervised model. But the random selection of false negatives could affect the performance of the classifier, as it learns wrongly potentially positive examples as negatives. Hence learning from positive and unlabeled data is a hot topic. So instead of selecting a random subset from unlabeled data, the subset of instances can be further processed to delete the potentially positive example through clustering and classification. The instances left behind in the clusters are the strong and reliable negative instances, which can be used for training a supervised model. As supervised approach yields better result and can help in finding the functions of unknown genes, identifying pathways, finding potential target and managing patient's health based on genomic sequence.

Acknowledgments. To complete this work, we have got valuable suggestions and guidance from different experts of this field. We also thankful to all the authors of the references we have taken.

References

1. Davidson, E., Levine, M.: Gene Regulatory Network. PNAS 102(14), 4935 (2005)
2. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., Guthke, R.: Gene regulatory network inference: Data integration in dynamic models-A review. Bio Systems (2008)
3. Zoppoli, P., Morganella, S., Ceccarelli, M.: TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics (2010)
4. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics (2006)
5. Liang, S., Fuhrman, S., Somogyi, R.: Reveal, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. In: Pac. Symp. Biocomput., pp. 18–29 (1998)
6. de Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. (2002)
7. Werhli, A.V., Husmeier, D.: Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Stat. Appl. Genet. Mol. Biol. (2007)
8. Wang, C., Ding, C., Meraz, R.F., Holbrook, S.R.: PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. Bioinformatics, 2590–2596 (2006)
9. Ceccarelli, M., Cerulo, L.: Selection of negative examples in learning gene regulatory networks. In: IEEE International Conference on Bioinformatics and Biomedicine Workshop, BIBMW 2009, pp. 56–61 (2009)
10. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, pp. 213–220. ACM, New York (2008)
11. Lin, H.T., Lin, C.J., Weng, R.C.: A note on Platt's probabilistic outputs for support vector machines. Mach. Learn., 267–276 (2007)
12. Mordelet, F., Vert, J.P.: SIRENE: supervised inference of regulatory networks. Bioinformatics, 76–82 (2008)
13. Li, X., Liu, B.: Learning to Classify Texts Using Positive and Unlabeled Data. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, IJCAI 2003, Acapulco, Mexico, August 9-15, pp. 587–594 (2003)
14. Faith, J.J., et al.: Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. (2007)
15. Salgado, H., et al.: Regulondb (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and rowth conditions. Nucleic Acids Res. 34(Database issue), D394–D397 (2006)
16. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)