Chapter 6
A Semi-Supervised Approach to GRN Inference Using Learning and Optimization

Meroua Daoudi
MISC Laboratory, Computer Science Department, Abdelhamid Mehri Constantine 2 University, Algeria

Souham Meshoul
IT Department, Nourah Bint Abdulrahman University, Saudi Arabia

Samia Boucherkha
Computer Science Department, Abdelhamid Mehri Constantine 2 University, Algeria

ABSTRACT

Gene regulatory network (GRN) inference is a challenging problem that lends itself to a learning task. Both positive and negative examples are needed to perform supervised and semi-supervised learning. However, GRN datasets include only positive examples and/or unlabeled ones. Recently a growing interest is being devoted to the generation of negative examples from unlabeled data. Within this context, the authors propose to generate potential negative examples from the set of unlabeled ones and keep those that lead to the best classification accuracy when used with positive examples. A new proposed genetic algorithm for fixed-size subset selection has been combined with a support vector machine model for this purpose. The authors assessed the performance of the proposed approach using simulated and experimental datasets. Using simulated datasets, the proposed approach outperforms the other methods in most cases and improves the performance metrics when using balanced data. Experimental datasets show that the proposed approach allows finding the optimal solution for each transcription factor in this study.

DOI: 10.4018/979-8-3693-3026-5.ch006

Copyright © 2024, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Understanding and modelling regulation in biological systems is a challenging task in bioinformatics as it requires identification of the relationship between the various components of these systems and also inferring potential influences that some components may have on others. Since the advancement in high throughput technologies and the generation of massive biological data, several computational biology techniques have been proposed and several models have been developed to help in knowledge discovery and better understanding of the regulations between biological components and how these regulations give rise to the functions and behaviors of biological systems. At the cell level, a Gene Regulatory Network (GRN) is a set of genes and regulators that interact with each other to govern the gene expression level. Modeling these interactions is a powerful abstraction of biological systems that can serve as a tool to understand and analyze the genes' interactions and the functions within a cell.

The first step in a GRN inference process is to identify the primary regulation between regulators known as transcription factors (TFs) and their target genes which help in understanding genetic processes and genetic modifications (Desai et al., 2017). Using experimental methods to determine regulation between TFs and target genes is costly and time-consuming (Patel & Wang, 2015). Furthermore, high throughput technologies generate massive expression data offering a great opportunity to infer GRNs using computational methods. Inferring GRNs from expression data can be cast as a machine learning problem and several approaches have been already proposed using supervised, unsupervised and semi-supervised techniques. As a first attempt, a large variety of unsupervised methods have been proposed. Then, a large number of transcription factors and their target genes is identified with experimental methods and they are considered as known regulations. New available and labeled datasets have led to the use of supervised and semi-supervised techniques to better make use of new data to infer more reliable and efficient networks. One of the most challenging tasks when inferring GRNs from gene expression data using supervised and semi-supervised learning is how to extract reliable negative examples. The challenge is due to the difficulty to verify experimentally the absence of any regulation between a TF and a target gene (Gillani et al., 2014) and to the high computational complexity of GRN inference. The performance of supervised methods depends on the quality of available data and most proposed approaches consider unknown regulations as negative, which could affect the performance of the classifier (Turki & Rajikhan, 2016).

To overcome the aforementioned limitations, we propose in this paper a generative approach to select reliable negative examples. The main idea is to generate potential negative examples using a search in the space of unlabeled examples and keep those that lead to the best classification accuracy when used with positive examples. This can be cast as an optimization process where the decision variables are related to the unlabeled examples and the objective function is related to the classification accuracy. To achieve this generation task, we propose the joint use of an optimizer and a classification model as shown in Figure 1.

The proposed approach describes a semi supervised learning model as we deal with a set of labeled and unlabeled data. More specifically, a genetic algorithm (GA) is combined with a support vector machine (SVM) model for this purpose. On the other hand, machine-learning algorithms work better in the case of balanced classes. However, the random generation of the population in the genetic algorithm could create the problem of data unbalance. To deal with the problem of unbalanced classes, we propose a new variant of genetic algorithm to select a fixed size of reliable negative examples from unlabeled data. The new proposed genetic algorithm is used with a support vector machine model to predict interactions

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/semi-supervised-approach-grn-inference/342524

Related Content

A Comparative Study of an Unsupervised Word Sense Disambiguation Approach
Wei Xiong, Min Songand Lori deVersterre (2013). *Bioinformatics: Concepts, Methodologies, Tools, and Applications (pp. 1306-1316).*
www.irma-international.org/chapter/comparative-study-unsupervised-word-sense/76119

Use of SciDBMaker as Tool for the Design of Specialized Biological Databases
Riadh Hammamiand Ismail Fliss (2013). *Bioinformatics: Concepts, Methodologies, Tools, and Applications (pp. 1755-1768).*
www.irma-international.org/chapter/use-scidbmaker-tool-design-specialized/76146

Infer Species Phylogenies Using Self-Organizing Maps
Xiaoxu Han (2010). *International Journal of Knowledge Discovery in Bioinformatics (pp. 29-49).*
www.irma-international.org/article/infer-species-phylogenies-using-self/45164

Efficient and Robust Analysis of Large Phylogenetic Datasets
Sven Rahmann, Tobias Muller, Thomas Dandekarand Matthias Wolf (2006). *Advanced Data Mining Technologies in Bioinformatics (pp. 104-117).*
www.irma-international.org/chapter/efficient-robust-analysis-large-phylogenetic/4248

Privacy Preserving Principal Component Analysis Clustering for Distributed Heterogeneous Gene Expression Datasets
Xin Li (2013). *Methods, Models, and Computation for Medical Informatics (pp. 238-271).*
www.irma-international.org/chapter/privacy-preserving-principal-component-analysis/73081