Neural Networks 43 (2013) 63–71

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Causal gene identification using combinatorial V-structure search

Ruichu Cai [a,b,*], Zhenjie Zhang [c], Zhifeng Hao [a]

[a] Faculty of Computer Science, Guangdong University of Technology, Guangzhou, PR China
[b] State Key Laboratory for Novel Software Technology, Nanjing University, PR China
[c] Advanced Digital Sciences Center, Illinois at Singapore Pte. Ltd., Singapore

ARTICLE INFO

Article history:
Received 15 May 2012
Received in revised form 23 January 2013
Accepted 31 January 2013

Keywords:
Causal gene
V-Structure
Gene expression data
Causality

ABSTRACT

With the advances of biomedical techniques in the last decade, the costs of human genomic sequencing and genomic activity monitoring are coming down rapidly. To support the huge genome-based business in the near future, researchers are eager to find killer applications based on human genome information. Causal gene identification is one of the most promising applications, which may help the potential patients to estimate the risk of certain genetic diseases and locate the target gene for further genetic therapy. Unfortunately, existing pattern recognition techniques, such as Bayesian networks, cannot be directly applied to find the accurate causal relationship between genes and diseases. This is mainly due to the insufficient number of samples and the extremely high dimensionality of the gene space. In this paper, we present the *first* practical solution to causal gene identification, utilizing a new combinatorial formulation over *V-Structures* commonly used in conventional Bayesian networks, by exploring the combinations of significant V-Structures. We prove the NP-hardness of the combinatorial search problem under a general settings on the significance measure on the V-Structures, and present a greedy algorithm to find sub-optimal results. Extensive experiments show that our proposal is both scalable and effective, particularly with interesting findings on the causal genes over real human genome data.

1. Introduction

With the advances of biomedical techniques in the last decade, such as microarray (Bassett, Eisen, & Boguski, 1999), the cost of gene activity monitoring is coming down to several hundreds.[1] In the near future, it is likely that microarrays will be used to test the gene activities for every person for disease diagnosis or gene therapy, forming a business market worth billions of dollars. Before the arrival of the new genomic age, biomedical researchers are now eager to look for killer applications in the huge genomic business. Causal gene identification is one of the most promising applications (Noble, 2008), which aims to help potential patients to accurately estimate their risk with respect to certain diseases.

Unfortunately, identification of the causal genes related to genetic diseases is by no means an easy task in the biomedical domain (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009). While traditional biological and pathological methods fail to effectively and efficiently discover the causal genes, computer scientists and statisticians are trying to apply machine learning and data mining techniques to tackle the problem, e.g. Cai, Hao, Yang, and Wen (2009); Cai, Tung, Zhang, and Hao (2011), Kim, Wuchty, and Przytycka (2011) and Schadt et al. (2005). Given the gene expression data from humans with/without certain genetic diseases, algorithms are designed to automatically find out significant genes causing these diseases.

In statistics and learning communities, Bayesian networks (BN) are a common tool used to analyze the correlation and causality relationships between variables. By running statistical significance tests on variable combinations, it is possible to construct a probabilistic graphical model to simulate and evaluate the impact of certain variables over others (Cai, Zhang, & Hao, 2011). However, existing BN methods suffer from three major drawbacks on the causal gene identification problem. Firstly, complete BN construction needs an exponential number of samples to support accurate estimation in statistical tests. Secondly, most of the BN learning methods focus only on building a probabilistic model with high likelihood, instead of finding the exact causality relationship. This potentially leads to a large number of false positive causality connections between genes and diseases, even when the probabilistic model achieves a high likelihood. Therefore, although BN structure learning methods are capable of finding partial causal genes, these methods tend to output more noisy results with diminishing accuracy and low robustness, due to the low signal–noise ratio, the high

* Corresponding author at: Faculty of Computer Science, Guangdong University of Technology, Guangzhou, PR China. Tel.: +86 015800030523; fax: +86 20 39323163.
E-mail address: cairuichu@gmail.com (R. Cai).

[1] http://www.cincinnatichildrens.org/research/cores/gene-expression/fees/.

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2013.01.025

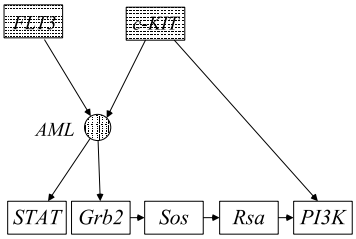

Fig. 1. Pathway related to *Acute Myeloid Leukemia*.

dimensionality in gene expression data and the limited number of samples.

In Fig. 1, we present an example of a known pathway related to *Acute Myeloid Leukemia* (AML). All the genes' expression levels are highly correlated to the variable AML, i.e. the disease status of the patient. However, the gene expression level of STAT, Grb2, Sos, Rsa and P13K does not affect the state of the disease, while FLT3 and c-KIT are the only direct causal genes of AML. It means that these two genes should be the target genes in any gene therapy on AML. In the traditional gene selection or Bayesian Network methods, it is difficult to distinguish FLT3 and c-KIT from the other related genes. This is due to the existence of probabilistic equivalent models different from the true pathway in the figure, which also fully fit the observations, but probably with completely different structures.

In this paper, we present a new solution to tackle the causal gene identification problem, based on a combinatorial formulation over *V-Structures* in Bayesian network. V-Structure is a special type of local probabilistic model involving only three variables X, Y, Z in form of $X \rightarrow Y \leftarrow Z$. Different from other local structures, V-Structures are supposed to be more robust and discriminating in causality identification problems, since it is not statistically equivalent to any other structures involving the same variables. On the other hand, the computation and verification of V-Structures is relatively cheap, compared to the complete construction of Bayesian network. V-structure thus plays an important role in conventional Inductive Causality methods commonly used in causal discovery methods (Pearl, 2009).

However, in the gene expression data, a large number of false V-Structures may be discovered, since it is hard to obtain a significant statistic test on small sample high dimensional data. Fortunately, we observe that the falsely discovered V-Structure can be detected for it is usually in conflict with true ones. Thus, we transform the causal gene selection problem into another optimization problem, targeting to identify a group of most significant V-Structures with maximal coverage and without conflict. Although the optimization formulation is proven NP-hard, we present a greedy algorithm effective on finding sub-optimal causality results with high accuracy. Our tests on synthetic datasets verify the effectiveness of our algorithm. Experiments on real gene expression data reveal interesting results on causal genes related to *Prostate Cancer* and *Leukemia*.

The outline of the paper is listed as follows. First, we will review some existing studies on disease-related gene discovery and causality inference in Section 2. Second, we will discuss the problem definition and preliminary knowledge in Section 3. Third, we will show our theoretical analysis and details of the algorithm in Section 4. We will then present experimental results in Section 5 and finally conclude this paper in Section 6.

2. Related work

Feature selection is the most commonly used tool for the disease-related gene discovery method (Saeys, Inza, & Larrañaga, 2007). Generally speaking, feature selection methods can be classified into three categories: filters, wrappers and embedded methods. Filters employ only intrinsic properties of the feature without considering its interaction with the classifier. In wrapper methods, a classifier is usually built and employed as the evaluation criterion. If the feature selection criterion is derived from the intrinsic properties of a classifier, the corresponding method belongs to the embedded methods category. False discovery (Reiner, Yekutieli, & Benjamini, 2003) and feature set redundancy (Yu & Liu, 2004) are two problems we need to consider for all feature selection problems.

A causality Bayesian network is part of the theoretical background of this work. A causality Bayesian network is a special case of a Bayesian network, whose edge direction presents the causality relations among the nodes (Pearl, 2009). The Causality Bayesian network is different from the Bayesian network used in the regulatory network reconstruction problem, such as Friedman, Linial, Nachman, and Pe'er (2000) and Kim, Imoto, and Miyano (2004).

Structure learning of a Bayesian network is closely related to the algorithmic background of this work, e.g. the well-known PC algorithm (Kalisch & Bühlmann, 2007; Spirtes, Glymour, & Scheines, 2001) and Markov Blanket discovery methods (Zhu, Ong, & Dash, 2007). These methods provide the skeleton of causal structures, i.e. parent–child pairs and Markov Blanket. However, these methods usually cannot distinguish causes from consequences, which mostly relies on other techniques to conduct exact causal discovery.

Pearl is the founder of the causality analysis theory (Pearl, 2009). Most causality inference works simply assume the acquisition of a sufficiently large sample set (Aliferis, Statnikov, Tsamardinos, Mani, & Koutsoukos, 2010a, 2010b), or expensive intervention experiments (He & Geng, 2008). Though there are some works aiming to solve the inference problem when a small number of samples are available (Bromberg & Margaritis, 2009), the exact sample sizes used in their empirical studies remain significantly larger than the scale of gene expression data. To the best of our knowledge, there does not exist a provable method to run robust causal inference on the real gene expression data. In this paper, we present the *first* practical algorithm to tackle the problems of small sample size and high dimensionality in gene data.

Another concept that is closely related to our work is Granger's causality (Lozano, Abe, Liu, & Rosset, 2009; Mukhopadhyay & Chatterjee, 2007), which uses Granger's causality theory to infer the gene regulatory networks from the time series gene expression data. Granger's work differs from traditional causality inference techniques in two aspects. Firstly, compared with the conventional definition of causality, Granger's causality is more likely a regression method and does not reflect the true causality mechanism. Secondly, the temporal information is essential for Granger's causality inference, which is hard to collect in the disease–gene relationship analysis context.

3. Preliminaries

Assume that all samples from the problem domain contain information on m different genes, i.e. $G = \{g_1, g_2 \ldots, g_m\}$, and the disease state of the sample y. Let $D = \{x_1, x_2, \ldots, x_n\}$ denote the complete sample set. Each sample x_i is denoted by a vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{im}, y_i)$, where x_{ij} indicates the expression level of the sample x_i on gene g_j. And y_i is the disease state associated with the sample x_i.

In particular, if $\mathcal{P}$ is the distribution defined on all the genes' expression level and the state of the disease, i.e. $V = G \bigcup \{y\}$, we assume that there exists a Bayesian network *BN* faithful to the distribution $\mathcal{P}$. A Bayesian network includes a directed acyclic graph which indicates conditional (in)dependent relationships among the variables, and conditional probability functions which simulate conditional probability distribution of each variable given the parent nodes. Following the common assumption of existing studies, we only consider a problem domain with the *Faithfulness Condition* (Koller & Friedman, 2009) as listed below.