International Journal of Approximate Reasoning 80 (2017) 1–18

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A novel Chow–Liu algorithm and its application to gene differential analysis

Joe Suzuki

Department of Mathematics, Osaka University, Japan

ARTICLE INFO

Article history:
Received 22 October 2015
Received in revised form 29 July 2016
Accepted 1 August 2016
Available online 8 August 2016

Keywords:
Mutual information
Chow–Liu algorithm
Forest graphical model
Gene differential analysis
ANOVA

ABSTRACT

This paper proposes an estimator of mutual information for both discrete and continuous variables and applies it to the Chow–Liu algorithm to find a forest that expresses probabilistic relations among them. The state-of-the-art assumes that the continuous variables are Gaussian and that the graphical model under discrete and continuous variables is ANOVA. Consequently, it is difficult to obtain the maximum likelihood of three connected variables such that the center is Gaussian and the other two are discrete, and thus, the state-of-the-art restricts the class to the forest such that there is no Gaussian node between discrete variables. The proposed method executes in a general setting without any assumptions, preparing several meshes, computing the mutual information values, and selecting the maximum value. We prove that the number of meshes to be prepared is at most $O(\log^2 n)$ and that the estimated mutual information is no larger than zero if and only if the variables are independent for large n. Finally, we apply the proposed method to the problems of gene differential analysis and relation discovery between gene expression and SNPs (single nucleotide polymorphisms). In particular, for the latter experiment, we demonstrate that the proposed method successfully captures the relation among them but that the state-of-the-art fails because of the merits and demerits of the proposed and existing methods.

1. Introduction

We consider learning a forest structure among variables from data and its application to gene differential analysis. Although we express conditional independence (CI) relations among variables in terms of a graphical model, to reduce computational efforts,[1] we focus on forests in which no edge is directed and no loop exists in the graph.

We say that $N(\geq 2)$ discrete variables $X^{(1)}, \cdots, X^{(N)}$ are expressed by a forest if the marginal distribution is in the form

$$\prod_{i\in V} P(X^{(i)}) \cdot \prod_{\{i,j\}\in E} \frac{P(X^{(i)}, X^{(j)})}{P(X^{(i)})P(X^{(j)})}, \tag{1}$$

where $V := \{1, \cdots, N\}$ and E is a subset of $\{\{i, j\}|i, j \in V, i \neq j\}$ such that no loop exists in any path in the undirected graph G when we observe that V and E are the vertex and edge sets of G.

E-mail address: suzuki@math.sci.osaka-u.ac.jp.

[1] It is true that exact learning of BNs is NP-hard and not “scalable”; however, there are instances of exact (non-tree) BN learning for $N > 40$ for particular datasets and/or with certain constraints.

http://dx.doi.org/10.1016/j.ijar.2016.08.001
0888-613X/© 2016 Published by Elsevier Inc.

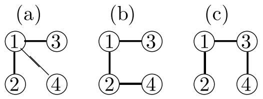

Fig. 1. Undirected Graphs.

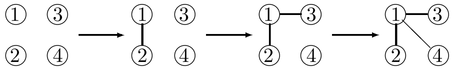

Fig. 2. The Chow–Liu Algorithm when $I(1,2) > I(1,3) > I(2,3) > I(1,4) > I(3,4) > I(2,4)$.

For example, suppose that $N = 4$. The distributions

$$P(X^{(1)})P(X^{(2)}|X^{(1)})P(X^{(3)}|X^{(1)})P(X^{(4)}|X^{(1)})$$
$$P(X^{(1)})P(X^{(2)}|X^{(1)})P(X^{(3)}|X^{(1)})P(X^{(4)}|X^{(2)})$$
$$P(X^{(1)})P(X^{(2)}|X^{(1)})P(X^{(3)}|X^{(1)})P(X^{(4)}|X^{(3)})$$

are expressed by Figs. 1 (a)(b)(c) using forests in which the edge sets are $E = \{\{1,2\},\{1,3\},\{1,4\}\}$, $E = \{\{1,2\},\{1,3\},\{2,4\}\}$, and $E = \{\{1,2\},\{1,3\},\{3,4\}\}$, respectively.

Then, let $I(j,k)$ be the mutual information between $X^{(j)}$ and $X^{(k)}$ $(j \neq k)$. Chow and Liu [2] constructed a tree rather than a forest as follows: starting with $E = \{\}$ and $\mathcal{E} = \{\{i,j\}|i \neq j\}$, repeatedly choose a pair $\{j,k\}$ such that $I(j,k)$ is the largest among $I(j',k')$ with $\{j',k'\} \in \mathcal{E}$, delete it from $\mathcal{E}$, and add it to E if adding it to E does not cause a loop to be generated. This procedure is repeated until $\mathcal{E}$ is empty. For example, suppose that $N = 4$ and that $I(1,3)$ and $I(1,3)$ are the largest and second largest mutual information values, respectively. Then, even if $\{2,3\}$ has the third largest mutual information, it cannot be joined in the edge set E because loop $\{\{1,2\},\{2,3\},\{3,1\}\}$ will be generated. If $I(1,4)$ is the fourth largest, Fig. 2 (a) is obtained through the procedure, as shown in Fig. 2.

These authors proved that the obtained E minimizes the Kullback–Leibler divergence from the original distribution $P(X^{(1)},\cdots,X^{(N)})$ among the distributions in the form of (1).

The same concept is applied to find probabilistic relations among the N variables when the true distribution is not known and when only a dataset with n observations of them is available. In particular, they considered obtaining the maximum likelihood value of (1) given n examples

$$(X^{(1)} = x_{i,1},\cdots,X^{(N)} = x_{i,N})_{i=1}^{n}$$

by replacing the mutual information $I(j,k)$ by the empirical estimator in which each probability is replaced by its relative frequency, where each example consists of N variable values, and they proved that the obtained edge set E maximizes the likelihood among the distributions in the form of (1).

In 1993, Suzuki [15] considered the problem in terms of the minimum description length (MDL) principle and showed that the obtained graph should be a forest rather than a spanning tree when a subset of the variables is independent from the others. Then, several authors revisited the same formula, such as P. Liang and N. Srebro [8], K. Panayidou [10], and Edwards et al. [4] in 2004, 2010, and 2010, respectively.

In this paper, we propose a novel estimator of mutual information and apply it to the Chow–Liu algorithm for the case in which discrete and continuous data are mixed.

Now, let us briefly present the main contributions of this paper. The state-of-the-art [4] considered estimating mutual information based on the ANOVA model when one variable is Gaussian and the other is discrete. Then, if X is Gaussian and Y, Z are discrete, although it is easy to obtain the maximum likelihoods of graphical models $Y - X$ and $X - Z$, it is difficult to obtain that of another graphical model $Y - X - Z$ by maximizing the empirical mutual information estimations of $Y - X$ and $X - Z$ [4]. Consequently, the state-of-the-art avoids such an inconvenience and restricts the class to the forests such that there is no Gaussian variable between any discrete variables, i.e., the discrete and continuous nodes are separated in the resulting forest. This paper estimates the mutual information in a general setting and solves the problem. In particular, when SNPs (single nucleotide polymorphisms) and gene expressions are contained in the set of attributes, as demonstrated in one of the genome experiments in this paper, the former and latter are categorical and continuous variables, respectively. The proposed method successfully captures the relation among multiple gene expressions and SNPs, whereas the state-of-the-art fails.

The remainder of this paper is organized as follows. Section 2 presents the background and previous works on the problem. Section 3 proposes a novel estimator of mutual information and demonstrates the advantageous properties of the estimator. In Section 4, we apply the proposed method to genome analysis using two data sets. We address gene differential analysis in Section 4.1, and we find the relations among multiple gene expressions and SNPs in Section 4.2. In Section 5, we summarize the results of this study and present directions for future work.