On Exact Identification of the Structure of a Probabilistic Boolean Threshold Network from Samples

Avraham A. Melkman[1] Xiaoqing Cheng[2] Wai-Ki Ching[3] Tatsuya Akutsu[4,a)]

Abstract: We study the problem of exactly identifying the structure of a probabilistic Boolean network (PBN) from a given set of samples, where PBNs are a probabilistic model of genetic networks and neural networks. We consider PBNs consisting of Boolean threshold functions, while focusing on those functions with unit coefficients. We show that wide classes of PBNs with such threshold functions can be exactly identified from samples under reasonable constraints, including: PBNs with any number of threshold functions each of which has the same number of input variables, and PBNs consisting of pairs of threshold functions with different numbers of input variables.

1. Introduction

Identification of the network structure from observed data is an important research topic both in systems biology [1] and in neuroscience [2]. Various mathematical models have been utilized for identifying network structures. Among them, the *Boolean network* (BN) is a well-studied discrete mathematical model, which has been used in modeling gene regulatory networks [3], [4] and neural networks [5]. As a probabilistic extension of BNs, the *Probabilistic Boolean Network* (PBN) has been proposed [6]. Although extensive theoretical studies have been done on identification of BNs, almost no theoretical results had been known on identification of PBNs.

Recently, Cheng et al. studied classes of PBNs whose structures can be exactly identified from samples [7]. Although they focused on PBNs with AND/OR functions, threshold functions are popular in modeling biological networks, especially in modeling neural networks. Therefore, we study classes of PBNs with Boolean threshold functions whose structures are exactly identified from samples. In this extended abstract, we briefly present our major theoretical findings, where detailed results and their proofs are given in [8].

2. Definitions

Throughout this abstract, $\mathbf{a}$ denotes a 0-1 bit vector of length n, and $\mathbf{a}_i$ denotes the 0-1 value of its ith bit (i.e., $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$). For a Boolean variable x, a *literal* is either x or its negation $\overline{x}$.

Definition 1 A Boolean function f is a *threshold function* with integer threshold θ if there exist integers w_i such that $f(x_1, \ldots, x_n) = 1$ if and only if $\sum_{i\in\{1,\ldots,n\}} w_i \ell_i \geq \theta$, for all $(x_1, \ldots, x_n) \in \{0,1\}^n$, where ℓ_i is either x_i or $\overline{x_i}$.

Definition 2 A Boolean Threshold Network is a directed network with n nodes $x_1, \ldots, x_n$, in which node x_i has an associated Boolean threshold function $f^{(i)}$. At time step t, node x_i takes on a value $x_i(t)$ that is either 0 or 1, and $x_i(t+1)$ is determined by $x_i(t+1) = f^{(i)}(x_1(t), \ldots, x_n(t))$. A Probabilistic Boolean Threshold Network (**PBTN**) is a directed network with n nodes in which node x_i is associated with a set $F = \{f_1^{(i)}, \ldots, f_{m_i}^{(i)}\}$ of Boolean threshold functions, and with corresponding selection probabilities $c_j^{(i)}$, $\sum_{j=1}^{m_i} c_j^{(i)} = 1$. The value of node x_i at time $t+1$ is determined by $x_i(t+1) = f_j^{(i)}(x_1(t), \ldots, x_n(t))$ with probability $c_j^{(i)}$, where selection of $f_j^{(i)}$ is independent of selections at previous time steps and of selections for other nodes.

Denote $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))$. Our purpose is to identify $f_j^{(i)}$s assigned to each node x_i from a set of $(\mathbf{x}(0), \mathbf{x}(1))$s, where we do not aim to identify the probabilities $c_j^{(i)}$s. Since threshold functions assigned to each node can be identified independently of other nodes, we focus on identifying a set of threshold functions for only one output node from a given set of pairs $(\mathbf{a}, y)$, where $\mathbf{a} \in \{0,1\}^n$ and $y \in \{0,1\}$.

We assume that a class C of PBTNs is given, and that a set of samples S is generated using some PBTN $F \in C$, meaning that for each $(\mathbf{a}, y) \in S$ the value y belongs to the set $F(\mathbf{a}) = \{f_1(\mathbf{a}), \ldots, f_p(\mathbf{a})\}$.

Definition 3 A PBTN $F = \{f_1, \ldots, f_p\}$ is *consistent* with a sample $(\mathbf{a}, y)$ if $y \in F(\mathbf{a}) = \{f_1(\mathbf{a}), \ldots, f_p(\mathbf{a})\}$. If F is consistent with every sample in S, it is *consistent with S*.

We consider two models, the *Partial Information Model* (PIM), and the *Full Information Model* (FIM).

Definition 4 S identifies F from among C under PIM if F is the only PBTN in C that is consistent with all samples in S.

Definition 5 S identifies F from among C under FIM if (i) F is the only PBTN in C that is consistent with all samples in S,

[1] Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

[2] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China

[3] Department of Mathematics,The University of Hong Kong, Pokfulam Road, Hong Kong, China

[4] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan

[a)] takutsu@kuicr.kyoto-u.ac.jp

and (ii) if $(\mathbf{a}, y) \in S$ then $\mathbf{a} \times F(\mathbf{a}) \subseteq S$, i.e. all possible samples $(\mathbf{a}, f(\mathbf{a})), f \in F$ were generated.

Definition 6 A class C is identifiable from samples under PIM (resp., FIM) if for every $F \in C$, there is a set of samples that identifies F under PIM (resp., FIM).

The following theorem characterizes the identifiable classes under PIM and FIM.

Theorem 7 A class C of PBTNs is PIM-identifiable if and only if for every $F, G \in C$ there is an assignment $\mathbf{a}$ such that $F(\mathbf{a}) - G(\mathbf{a}) \neq \emptyset$. C is FIM-identifiable if and only if for every $F, G \in C$ there is an assignment $\mathbf{a}$ such that $F(\mathbf{a}) \neq G(\mathbf{a})$.

It is clear from this theorem that if C is PIM-identifiable, C is FIM-identifiable.

Example 8 Let $F = \{x_1 \geq 1, \overline{x_1} \geq 1\}$, $G = \{x_2 \geq 1, \overline{x_2} \geq 1\}$, and $C = \{F, G\}$, where $n = 4$. Then, C is not PIM-identifiable or FIM-identifiable because $F(\mathbf{a}) = G(\mathbf{a}) = \{0, 1\}$ holds for any $\mathbf{a} \in \{0,1\}^4$, i.e., y behaves as a random 0-1 value and thus we cannot discriminate between F and G. Let $F' = \{x_1 \geq 1, x_2 \geq 1\}$, $G' = \{x_3 \geq 1, x_4 \geq 1\}$, and $C' = \{F', G'\}$. Then, C' is PIM-identifiable (and also FIM-identifiable) because $G'((0,0,0,1)) - F'((0,0,0,1)) = \{0,1\} - \{0\} \neq \emptyset$ and $F'((0,1,0,0)) - G'((0,1,0,0)) = \{0,1\} - \{0\} \neq \emptyset$. It means that if we see a sample $((0,0,0,1),1)$ (resp., $((0,1,0,0),1)$), we can conclude that samples are generated by G' (resp., F').

Example 9 Let $f_1 = x_1 + x_2 \geq 1$, $f_2 = x_1 + \overline{x_2} + x_3 \geq 2$, $f_3 = x_1 + x_2 + x_3 \geq 3$, and $f_4 = x_1 \geq 1$. Let $F = \{f_1, f_2\}$, $G = \{f_2, f_4\}$, $H = \{f_2, f_3\}$, $C_1 = \{F, G\}$ and $C_2 = \{G, H\}$, where $n = 3$. Then, C_1 is identifiable from samples under FIM but not under PIM because $G(\mathbf{a}) \subseteq F(\mathbf{a})$ for all $\mathbf{a}$, whereas C_2 is identifiable from samples under both PIM and FIM because $G(\mathbf{a}') - H(\mathbf{a}') = \{1\}$ for $\mathbf{a}' = (1,1,0)$ and $H(\mathbf{a}'') - G(\mathbf{a}'') = \{0\}$ for $\mathbf{a}'' = (1,0,1)$ (see also Table 1).

Table 1 Example illustrating the difference between PIM and FIM.

x_1	0	0	0	0	1	1	1	1
x_2	0	0	1	1	0	0	1	1
x_3	0	1	0	1	0	1	0	1
f_1	0	0	1	1	1	1	1	1
f_2	0	1	0	0	1	1	0	1
f_3	0	0	0	0	0	0	0	1
f_4	0	0	0	0	1	1	1	1
$F = \{f_1, f_2\}$	0	0/1	0/1	0/1	1	1	0/1	1
$G = \{f_2, f_4\}$	0	0/1	0	0	1	1	0/1	1
$H = \{f_2, f_3\}$	0	0/1	0	0	0/1	0/1	0	1

3. Results

Before discussing identifiability, we consider the problem of deciding whether two given threshold functions are equivalent as Boolean functions.

Proposition 10 Deciding the equivalence of two Boolean threshold functions is co-NP complete.

Hereafter, we list some of our results on identifiability of PBTNs [8]. A PBTN F is called *admissible* if for each $i \in \{1, \ldots, n\}$, at most one of $x_i, \overline{x_i}$ appears in F.

Lemma 11 (Necessary Condition for PIM) A class C of admissible PBTNs is PIM-identifiable only if it does not contain F and G, such that $F \subseteq G$.

Theorem 12 Let $1 \leq \theta_1 < \theta_2 \leq K$ be two fixed thresholds, and let C be a class of admissible PBTNs satisfying the Necessary Condition for PIM, such that each $F \in C$ consists of two (not necessarily different) threshold functions with the following properties: every $f \in F$ depends on exactly K variables, has unit coefficients, and has a threshold that is either θ_1 or θ_2. Then C is PIM-identifiable.

Theorem 13 Let $1 \leq \theta_1 < \theta_m \leq K$ be two fixed thresholds, and let C be a class of admissible PBTNs satisfying the Necessary Condition for PIM, such that each $F \in C$ consists of m threshold functions with the following properties: every $f \in F$ depends on exactly K variables and has unit coefficients, and the thresholds of F are $\theta(f_1) = \theta_1$, $\theta(f_m) = \theta_m$ and $\theta_1 < \theta(f) < \theta_m$, $f \neq f_1, f_m$. Then C is PIM-identifiable if $f_1 \neq g_1$ or $f_m \neq g_m$ for all pairs $F, G \in C$.

Lemma 14 (Necessary Condition for FIM) Let C be a class of admissible PBTNs each of which consists of one or two threshold functions that have unit coefficients. If C is FIM-identifiable, it does not contain $F = \{f_1, f_2\}$ and $G = \{g_1, g_2\}$ such that $f_1 = \ell_1 \geq 1$ $f_2 = \ell_2 \geq 1$, $g_1 = \ell_1 + \ell_2 \geq 1$, $g_2 = \ell_1 + \ell_2 \geq 2$, with ℓ_1, ℓ_2 literals.

Theorem 15 Let C be a class of admissible PBTNs each of which consists of one or two threshold functions that have unit coefficients. Then C is FIM-identifiable if and only if the Necessary condition for FIM holds.

Example 16 Let $F = \{x_1 + x_2 + x_3 \geq 1, x_1 + x_2 + x_4 \geq 2\}$, $G = \{x_1 + x_2 + x_3 \geq 1, x_1 + x_2 + x_4 \geq 3\}$, and $C = \{F, G\}$. Then, C is FIM-identifiable from Theorem 15. However, C is not PIM-identifiable because $F(\mathbf{a}) \subseteq G(\mathbf{a})$ for all $\mathbf{a}$.

Finally, we discuss the sample complexity.

Theorem 17 Let C be a class of PBTNs consisting of L-tuplets of functions, each of which has at most K inputs, that satisfies the condition of PIM (resp., FIM) of Theorem 7. If, for fixed L and K, $O(\frac{1}{c} \cdot 2^{2LK} \cdot (2LK + 1 + \alpha) \cdot \log n)$ samples are generated uniformly at random, the correct PBTN can be uniquely identified at all nodes with probability no less than $1 - \frac{1}{n^\alpha}$ under PIM (resp., FIM).

References

[1] Karlebach, G. and Shamir, R.: Modelling and Analysis of Gene Regulatory Networks, *Nature Reviews Molecular Biology*, Vol. 9, 770–780 (2008).

[2] Lichtman, J. W. and Denk, W.: The Big and the Small: Challenges of Imaging the Brain's Circuits. *Science*, Vol. 334, 618–623 (2011).

[3] Kauffman, S. A.: Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets, *Journal of Theoretical Biology*, Vol. 22, 437–467 (1969).

[4] Xiao, Y.: A Tutorial on Analysis and Simulation of Boolean Gene Regulatory Network Models. *Current Genomics*, Vol. 10, 511–525 (2009).

[5] Anthony, M.: *Discrete Mathematics of Neural Networks, Selected Topics.* SIAM, Philadelphia (2001).

[6] Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W.: Probabilistic Boolean Networks: a Rule-based Uncertainty Model for Gene Regulatory Networks, *Bioinformatics*, Vol. 18, 261–274 (2001).

[7] Cheng, X., Mori, T., Qiu, Y., Ching, W-K., and Akutsu, T.: Exact Identification of the Structure of a Probabilistic Boolean Network from Samples, *IEEE/ACM Trans. Computational Biology and Bioinformatics*, Vol. 13, pp. 1107–1116 (2016).

[8] Melkman, A. M., Cheng, X., Ching, W-K., and Akutsu, T.: Identifying a Probabilistic Boolean Threshold Network from Samples, *IEEE Trans. Neural Networks and Learning Systems*, in press.