1995年度日本オペレーションズ・リサーチ学会
秋季研究発表会

Extensions of Partially Defined Boolean Functions with Missing Data

ラトガース大学 (Rutgers University) BOROS Endre
01001374 京都大学 茨木俊秀 (IBARAKI Toshihide)
02601514 京都大学 *牧野和久 (MAKINO Kazuhisa)

1 Introduction

The knowledge acquisition in the form of Boolean logic has been intensively studied in the recent research (e.g., [2, 3]): given a set of data, represented as a set T of binary "true n-vectors" (or "positive examples") and a set F of "false n-vectors" (or "negative examples"), establish a Boolean function (extension) f such that f is true (resp. false) in every given true (resp. false) vector; i.e., $T \subseteq T(f)$ and $F \subseteq F(f)$, where $T(f)$ (resp. $F(f)$) denotes the set of true vectors (resp. the set of false vectors) of f. A pair of sets (T, F) is called a *partially defined Boolean function (pdBf)*.

Unfortunately, the real-world data might not be complete. For instance, we have a battery of 45 biochemical tests for carcinogenicity. However, we do not apply every test to a given chemical, since all tests cannot be checked in a laboratory or some tests are very expensive. When a test is not applied, we say that the test result is *missing*. The set of data $(\tilde{T}, \tilde{F})$, which includes the missing results, is called *Partially defined Boolean functions with missing data (pBmd)*. To cope with such situations, we introduce three types of extensions. That is, given a pBmd $(\tilde{T}, \tilde{F})$, (i) a *consistent extension* is a function f such that, for every $\tilde{a} \in \tilde{T}$ (resp. $\tilde{F}$), there is a 0-1 vector a obtained from a by fixing missing data appropriately, for which $f(a) = 1$ (resp. $f(a) = 0$) holds, (ii) a *robust extension* is a function f such that, for every $\tilde{a} \in \tilde{T}$ (resp. $\tilde{F}$), any 0-1 vector a obtained from a by fixing missing data arbitrarily satisfies $f(a) = 1$ (resp. $f(a) = 0$), and (iii) a *most robust extension* is a function f which is a robust extension of a pBmd (T', F'), where (T', F') is obtained from $(\tilde{T}, \tilde{F})$ by fixing a smallest set of missing data appropriately (the remaining missing data are assumed to take arbitrary values). All of these extensions provide the logical explanations of given pBmd $(\tilde{T}, \tilde{F})$ with varied freedom given to the missing data in $\tilde{T}$ and $\tilde{F}$. By definition, if $(\tilde{T}, \tilde{F})$ has a robust extension, it is also a most robust extension and is a consistent extension, and if $(\tilde{T}, \tilde{F})$ has a most robust extension, it is a consistent extension. In case of a most robust (and a consistent) extension, it also provides information such that some missing data must take certain values if $(\tilde{T}, \tilde{F})$ can have a consistent extension. This type of information is also useful in analyzing incomplete data sets.

In the following, we consider the problems of deciding the existence of (and constructing) these extensions for a given pBmd $(\tilde{T}, \tilde{F})$, mainly from the view point of their computational complexity.

2 Preliminaries

A *Boolean function*, or a *function* in short, is a mapping $f : \mathbb{B}^n \mapsto \mathbb{B}$, where

$$\mathbb{B} = \{0, 1\},$$

and $x \in \mathbb{B}^n$ is called a *Boolean vector* (a *vector* in short). If $f(x) = 1$ (resp. 0), then x is called a *true* (resp. *false*) vector of f. The set of all true vectors (resp. false vectors) is denoted by $T(f)$ (resp. $F(f)$). A *partially defined Boolean function (pdBf)* is defined by a pair of sets (T, F) such that $T, F \subseteq \mathbb{B}^n$. A function f is called an *extension* (or theory) of the pdBf (T, F) if $T \subseteq T(f)$ and $F \subseteq F(f)$. As a pdBf does not allow missing data, we introduce set

$$\mathbb{M} = \{0, 1, *\},$$

and interpret the asterisk components $*$ of $v \in \mathbb{M}^n$ as missing bits. For a vector $v \in \mathbb{M}^n$, let $AS(v) = \{j \mid v_j = *, j = 1, 2, ..., n\}$. Clearly, $\mathbb{B}^n \subseteq \mathbb{M}^n$, and

$v \in \mathbf{B}^n$ if and only if $AS(v) = \emptyset$. For a subset $\tilde{S} \subseteq \mathbf{M}^n$, let $AS(\tilde{S}) = \{(v,j) | v \in \tilde{S}, j \in AS(v)\}$ be the collection of all missing bits of the vectors in $\tilde{S}$. Let us consider binary assignments $\alpha \in \mathbf{B}^Q$ for subsets $Q \subseteq AS(\tilde{S})$ of the missing bits. For a vector $v \in \tilde{S}$ and an assignment $\alpha \in \mathbf{B}^Q$, let v^α be the vector obtained from v by replacing the $*$ components which belong to Q by the binary values assigned to them by α, i.e.,

$$v_j^\alpha = \begin{cases} v_j & \text{if } (v,j) \notin Q \\ \alpha(v,j) & \text{if } (v,j) \in Q. \end{cases}$$

A *pdBf with missing data* (or in short *pBmd*) is a pair $(\tilde{T}, \tilde{F})$, where $\tilde{T}, \tilde{F} \subseteq \mathbf{M}^n$. To a pBmd $(\tilde{T}, \tilde{F})$ we always associate the set $AS = AS(\tilde{T} \cup \tilde{F})$ of all missing bits. A function f is called a *robust extension* of the pBmd $(\tilde{T}, \tilde{F})$ if

$$f(a^\alpha) = 1 \text{ and } f(b^\alpha) = 0 \text{ for all } a \in \tilde{T},\ b \in \tilde{F} \text{ and for all } \alpha \in \mathbf{B}^{AS}.$$

We first consider the problem of deciding the existence of a robust extension of a given pBmd $(\tilde{T}, \tilde{F})$.

Robust Extension RE
Input: A pBmd $(\tilde{T}, \tilde{F})$, where $\tilde{T}, \tilde{F} \subseteq \mathbf{M}^n$.
Question: Does $(\tilde{T}, \tilde{F})$ have a robust extension ?

It may happen that a pBmd $(\tilde{T}, \tilde{F})$ has no robust extension, but it has if we change some (or all) $*$ bits to appropriate binary values. A function f is called a *consistent extension* of pBmd $(\tilde{T}, \tilde{F})$, if there exists an assignment $\alpha \in \mathbf{B}^{AS}$ for which $f(a^\alpha) = 1$ and $f(b^\alpha) = 0$ for all $a \in \tilde{T}$ and $b \in \tilde{F}$, respectively. In other words, by defining $\tilde{T}^\alpha = \{a^\alpha \mid a \in \tilde{T}\}$ and $\tilde{F}^\alpha = \{b^\alpha \mid b \in \tilde{F}\}$, pBmd $(\tilde{T}, \tilde{F})$ is said to have a consistent extension if the pdBf $(\tilde{T}^\alpha, \tilde{F}^\alpha)$ has an extension for some assignment $\alpha \in \mathbf{B}^{AS}$. This leads us to the following problem.

Consistent Extension CE
Input: A pBmd $(\tilde{T}, \tilde{F})$, where $\tilde{T}, \tilde{F} \subseteq \mathbf{M}^n$.
Question: Does $(\tilde{T}, \tilde{F})$ have a consistent extension ?

It may also happen that not all missing bits should be specified in order to have a robust extension. In this case we want to find a minimum subset $Q \subseteq AS$, and an assignment $\alpha \in \mathbf{B}^Q$, for which the resulting pBmd $(\tilde{T}^\alpha, \tilde{F}^\alpha)$ has a robust extension.

Most Robust Extension MRE
Input: A pBmd $(\tilde{T}, \tilde{F})$, where $\tilde{T}, \tilde{F} \subseteq \mathbf{M}^n$.
Output: NO if $(\tilde{T}, \tilde{F})$ does not have a consistent extension; otherwise a subset $Q \subseteq AS$ and an assignment $\alpha \in \mathbf{B}^Q$, for which pBmd $(\tilde{T}^\alpha, \tilde{F}^\alpha)$ has a robust extension, and for which $|Q|$ is minimum.

3 Complexity Results

Theorem 1 *Problem RE can be solved in polynomial time.* □

Theorem 2 *Problem CE is NP-complete even if every $a \in \tilde{T} \cup \tilde{F}$ satisfies $|AS(a)| \leq 2$. But it can be solved in polynomial time if all $a \in \tilde{T} \cup \tilde{F}$ satisfy $|AS(a)| \leq 1$.* □

Theorem 3 *Problem MRE is NP-hard even if every $a \in \tilde{T} \cup \tilde{F}$ satisfies $|AS(a)| \leq 2$. But it can be solved in polynomial time if all $a \in \tilde{T} \cup \tilde{F}$ satisfy $|AS(a)| \leq 1$.* □

References

[1] E. Boros, P. L. Hammer, and J. N. Hooker, Predicting cause-effect relationships from incomplete discrete observations, *SIAM Journal on Discrete Mathematics* 7 (1994) 531-543.

[2] E. Boros, T. Ibaraki and K. Makino, Error-free and best-fit extensions of partially defined Boolean functions, RUTCOR Research Report RRR 14-95, Rutgers University, 1995.

[3] Y. Crama, P. L. Hammer and T. Ibaraki, Cause-effect relationships and partially defined boolean functions, *Annals of Operations Research* 16 (1988) 299-326.