

## 5 L - 3

## EBL 有効性計算のための枠組

山田 誠二 辻 三郎

(大阪大学・基礎工学部)

yamada@tsuji-lab.ce.osaka-u.ac.jp

## 1. はじめに

EBL(Explanation-Based Learning)[1]における学習の本質は、「問題解決の効率化」である。よって、EBLの有効性は、学習後、問題解決がどれほど向上したかで評価できる。しかし、これまでこの効率化は、学習後のシステムに試験例を大量に与え、実験的に検証されてきた。そこで、本報告では、その計算コストを解析的に求めることを目指す。EBLの有効性は、EBLシステムと学習なしシステムの両方に、同一の例系列を与えて、それが肯定例か否かを判別するために必要とする計算コストの比で表わされる。計算コストを解析するには、問題解決も含めた全体的かつ一般的なEBL学習手続きと学習なし手続きを設定する必要がある。なお、使用言語・知識表現として、最も一般的と考えられるprologを用いる。

## 2. 標準EBL手続き: SEP

一般的なEBL学習手続き: SEPを設定する。

## (A) 入力

- 1) 目標概念:  $GC$     2) 領域理論:  $DT$   
 3) 例:  $E$ の系列    4) 操作性規範

なお、4)において、ファクトは操作可能、ルールは操作可能でないとする。

## (B) 出力

- 1) 操作可能な概念記述:  $OCD$

以上の入力から、SEPは次のように設定される。

## &lt;SEP&gt;

いま、例の系列  $TE$  を  $[E_1, \dots, E_n]$  とする。

- (1)  $i = 1$ ,  $OCD = \{\}$  に初期化。
- (2)  $i = n + 1$  なら終了。  $OCD$  が出力。
- (3) 問題解決  $PS$  手続き  $(E_i, \{E_1, \dots, E_{i-1}\})$ : 例系列  $\{E_1, \dots, E_{i-1}\}$  で学習された  $OCD$  のみを用い、SLD導出で例  $E_i$  の証明を行なう。この際、 $OCD$  は  $DT$  より優先されて用いられる。証明できれば、(7) へ。
- (4) 説明構造生成手続き  $GE(E_i)$ : 目標を例  $E_i$  と  $DT$  を用いてSLD導出で証明し、説明構造  $ES_i$  を生成。
- (5) 一般化手続き  $G(ES_i)$ :  $ES_i$  をEBG法で一般化し、操作可能な概念記述  $OCD_i$  を生成。
- (6) 知識管理手続き  $M(OCD_i)$ :  $OCD_i$  を  $OCD$  の知識ベースの最後に追加。

(7)  $i \leftarrow i + 1$  として、(2) へ。

## 3. 標準学習なし手続き: SNP

SEPと同様に、標準学習なし手続き: SNPも定義しておく。入力は、SEPと同じ。例の系列  $TE$  を  $[E_1, \dots, E_n]$  とする。

- (1)  $i = 1$  に初期化。
- (2)  $i = n + 1$  なら終了。
- (3) 問題解決手続き  $PS(E_i, DT)$ : 領域理論だけで例  $E_i$ , 目標  $T_i$  の証明をSLD導出で行なう。
- (4)  $i \leftarrow i + 1$  として、(2) へ。

また、以下の用語を用いる。

・SLD説明構造: 説明構造をSLD木で表現したもの。ファクトの節点を「操作可能節点」、ルールの節点を「操作不可能節点」という。

・SLD説明経路: SLD木において、最初に説明構造を見つけるまでに辿ったノードを要素とするリスト。

## 4. いくつかの仮定と計算コストの設定

試験例を与えずに訓練例だけでEBL有効性を計算するために、以下の仮定を設ける。

<仮定1> 例系列  $[E_1, \dots, E_n]$  のうち、 $[E_1, \dots, E_m]$  を訓練例系列とする。その要素についての説明構造は、すべて異なる。よって、学習後には、異なった  $m$  個の  $OCD$  が得られる。

<仮定2> 例系列  $[E_1, \dots, E_n]$  のうち、 $[E_{m+1}, \dots, E_n]$  を、訓練例系列による学習後、学習器に与えられる目標概念の肯定例の系列とし、試験例系列と呼ぶ。この系列は、学習された  $OCD$  それぞれについての  $g$  個の肯定例から構成される。よって、 $n = (g+1)m$  が成り立つ。

<仮定3> 一つだけの目標概念の概念記述を学習すると仮定する。ただし、概念記述は、複数得られてもかまわない。

また、計算コストとしては、prologにおいて最も一般的な  $LI$  (Logical Inference) を用いる。SLD経路の要素数(ノード数)は、述語呼出しの回数なので、それがそのまま  $LI$  となり、計算コストとなる。残念ながら現状では、バックトラック及びユニフィケーションのコストは、無視している。

## 5. 有限試験例系列の有効性計算

$ET = [E_1, \dots, E_m, E_{m+1}, \dots, E_n]$

$[E_1, \dots, E_m]$ : 訓練例系列

$[E_{m+1}, \dots, E_n]$  : 試験例系列 ( $n = (g+1)m$ )

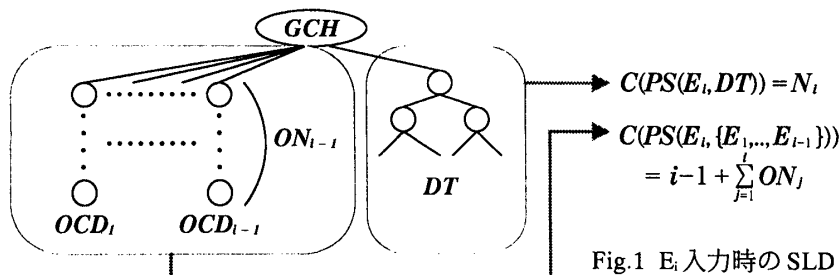
今, 上記の例系列  $TE$  が与えられたとする. Fig.1 は,  $E_i$  を証明したときの知識ベースのSLD木を表わしている.  $GCH$  は, 目標概念の頭である.  $E_i$  を証明する際, まず  $E_i \sim E_{i-1}$  で学習された  $OCD_i \sim OCD_{i-1}$  を用いて証明を行なうが, <仮定1>より, この証明はすべて失敗する. これにかかるコストが,  $C(PS(E_i, \{E_1, \dots, E_{i-1}\}))$  である. このコストにおいて,  $OCD_i \sim OCD_{i-1}$  の本体のすべての述語が調べられた後, 失敗とわかる, つまりコスト最大としている.  $OCD$  による証明が失敗すると, 次は  $DT$  による証明が行なわれ, この証明のコストが,  $C(PS(E_i, DT))$  である.

よって,  $SEP$  と  $SNP$  それぞれの計算コスト  $LC(TE)$  と  $NLC(TE)$  および EBL 有効性  $LE$  は, 以下ようになる. (式の詳細な説明は, 文献 [2] 参照.)

$$\begin{aligned}
 LC &= \sum_{i=1}^m C(PS(E_i, \{E_1, \dots, E_{i-1}\})) + GE(E_i) + G(ES_i) + M(OCD_i) \\
 &\quad + g \sum_{i=1}^m C(PS(E_i, \{E_1, \dots, E_m\})) \\
 &= \sum_{i=1}^m \{(i-1) + \sum_{j=1}^{i-1} ON_j + C(PS(E_i, DT) + G(ES_i) + M(OCD_i))\} \\
 &\quad + g \sum_{i=1}^m (i + \sum_{j=1}^i ON_j) \\
 &= \sum_{i=1}^m \{(i-1) + (m-i)ON_i + N_i + C(G(ES_i) + M(OCD_i))\} \\
 &\quad + gi + g(m-i+1)ON_i \\
 NLC &= \sum_{i=1}^m C(PS(E_i, DT)) = (g+1) \sum_{i=1}^m C(PS(E_i, DT)) = (g+1) \sum_{i=1}^m N_i \\
 LE &= 1 - \frac{LC}{NLC} \\
 &= 1 - \frac{\sum_{i=1}^m C(PS(E_i, \{E_1, \dots, E_{i-1}\})) + GE(E_i) + G(ES_i) + M(OCD_i)}{(g+1) \sum_{i=1}^m C(PS(E_i, DT))} \\
 &\quad - \frac{g \sum_{i=1}^m C(PS(E_i, \{E_1, \dots, E_m\}))}{(g+1) \sum_{i=1}^m C(PS(E_i, DT))} \\
 &= 1 - \frac{\sum_{i=1}^m \{i-1 + (m-i)ON_i + N_i + C(G(ES_i) + M(OCD_i))\}}{(g+1) \sum_{i=1}^m N_i} \\
 &\quad - \frac{\sum_{i=1}^m \{gi + g(m-i+1)ON_i\}}{(g+1) \sum_{i=1}^m N_i}
 \end{aligned}$$

ただし,

- $C(P)$  は, 手続き  $P$  の計算コスト.
- $ON_i$  は,  $E_i$  の SLD 説明構造中の操作可能節点数.
- $N_i$  は,  $E_i$  の SLD 説明経路の全節点数.



このLEを求めるには, 次のパラメータが必要である.  $ON_i, C(GE(E_i)), C(G(ES_i)), C(M(OCD_i)), N_i$ . しかし, 現状では,  $C(G(ES_i)), C(M(OCD_i))$  が設定されていないので, 残念ながら算出できない.

### 6. 無限例系列の有効性計算

これまで, 十分に多くの試験例系列が与えられた場合の EBL の有効性は, 実際に問題解決をさせた実験結果によるものであり, 非常に計算コストがかかる. しかし, ここでは  $LE$  において,  $g \rightarrow \infty$  とすることにより, 求めることができる. これを無限例系列の EBL 有効性:  $ILE$  と呼ぶ.

$$\begin{aligned}
 ILE &= \lim_{g \rightarrow \infty} LE \\
 &= 1 - \frac{\sum_{i=1}^m C(PS(E_i, \{E_1, \dots, E_m\}))}{\sum_{i=1}^m C(PS(E_i, DT))} = 1 - \frac{\sum_{i=1}^m \{i + (m-i+1)ON_i\}}{\sum_{i=1}^m N_i}
 \end{aligned}$$

$ILE$  のパラメータ:  $m, ON_i, N_i$  は,  $DT$  を用いて訓練例系列の証明を行なうだけで, 計算できる. さらに,  $GE(E_i), G(ES_i), M(OCD_i)$  の計算コストが  $ILE$  の計算に関与しないので, 実際に学習しなくても, 無限例列の EBL 有効性が計算できることである.

### 8. まとめ

本論文では, これまで実験的にしか検証されていなかった EBL の有効性を, 解析的手法で調べる方法を示した. その結果, 大量の試験例を与えなくても, 訓練例の証明を行なうだけで, 十分に多くの試験例を与えた場合の EBL 有効性:  $ILE$  を計算できるを示した. 今後, ・具体例で理論値と実験値の比較, ・ $LI$  以外の計算コストの検討, などを行う予定である.

### <参考文献>

[1] Mitchell, T.M. Keller, R.M. and Kedar-Cabelli, R. T.: Explanation-Based Generalization: A Unifying View, Machine Learning, Vol.1 No.1, pp.47-80 (1986)

[2] 山田, 辻: 論理プログラミング環境における EBL の有効性計算, 情報処理学会, 「人工知能」研究会資料, 90-AI-72 (1990・予定)

Fig.1  $E_i$  入力時の SLD 木