既存知識と事例に基づく融合型学習手法 (II)*

1C－5

金田 重郎† フセイン アルモアリム‡ 石井 恵† 秋葉 泰弘†
NTTコニュニケーション科学研究所† サウジアラビア国立石油鉱物大学‡

1 始めに

事例からの学習は，知識獲得の有力な手段とされているものの，現実には，収集可能な事例が少なく，十分な性能をもつ知識を獲得できない事が多い．著者らは，この問題の解決を狙いとして，既存知識と事例(実事例)を融合して，最終的な知識を得る手法を提案している [1].

この従来手法は，既存知識から逆に事例(仮事例)を生成し，この仮事例と実事例を学習アルゴリズムに投入する点に特徴がある．但し，この従来手法は，仮事例を実際に生成しているため，その個数が少ない場合には，既存知識が完全に反映されない恐れがある．

一方，ID3やCN2等のDivide-and-Conquer型の決定木生成アルゴリズムでは，決定木生成の各ノードにおいて，ある種のスコアを計算する．そこで，本論文では，仮事例を生成する事なく，あたかも膨大な個数の仮事例が生成された場合と同等の情報を，このスコア計算に反映させる手法を提案する [2].

2 仮事例生成を用いた融合型学習

2.1 融合型学習の原理

既存知識と実事例に，以下の情報が付与されているとする．

- 既存知識の各構成要素(ルール，決定木の葉)には，その要素の生起する確率を表す情報 p を与える．この値は，知識作成者が与える．それが困難な場合には，全構成要素等確率とする．
- 信頼度 (RR) が与えられる．このパラメータは，既存知識と実事例の相対的な重要性を示す．RR= 1.0では，既存知識は実事例と同等である．小さな RR 値は，実事例をより信頼する事を意味する．

*A Revision Learner for Expert-Knowledge and Real Examples,
Shigeo KANEDA †, Hussein Almuallim ‡,
Megumi ISHII †, Yasuhiro AKIBA †
†)NTT Communication Science Laboratories, 1-2356, Take, Yokosuka-shi, Kanagawa-ken, 238-03, JAPAN
‡) the Dept. of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261,Saudi Arabia

仮事例生成の原理を以下に述べる．最適な RR の値は，クロスバリデーションを用いて定める．

1. 既存知識からそれを満たす事例をランダムに m 個生成する．この生成された事例を仮事例 と呼ぶ．
2. 実事例を構成する各事例の重みを 1 とする．そして，Step1で生成された事例の各々に $RR \times |\text{実事例}|/m$ の重みを与える．
3. 仮事例と実事例を，最終知識を生成するための学習アルゴリズムに投入する．

3 仮事例生成を用いない融合型学習

上記手法において，既存知識の情報を完全に反映させるには，多くの仮事例を必要とする恐れがある．そこで，仮事例生成を用いないで，学習アルゴリズムに，既存知識を持ち込む．

ここで，以下の *conjunction* 形式を考える．

$$\{(x_i, v_j):\ x_i \text{ は属性であり}, v_j \text{ は属性値．}\}$$

conjunction P を考える．事例 e は $\forall(x_i, v_j) \in P$ において，属性値 x_i が v_j の時，conjunction P を満足すると定義する．

決定木生成のどのノードでも，ルートから現在のノードまでのパスは，ひとつの conjunction を構成する．この conjunction を，*LearnedTreeP* と呼ぶ．ノードに到着している実事例は，*LearnedTreeP* を満たす．必要なのは，仮事例でこの条件を満たすものの個数である．このタスクを実行する深さ優先の再帰的手続きを図 1に示す．引数は以下の通りである．

- P は conjunction である．
- *Node* は既存知識の探索されているノードである．
- *ExpertTreeP* は，既存知識のルートノードから *Node* に至るパスに対応する conjunction である．
- n は仮事例の全数とする．この値は，$RR \cdot |\text{実事例}|$ に取られる．

Procedure ComputeArtExs(P, *Node*,*ExpertTreeP*, n)
  If *Node* is a leaf, then
    $D = \{p : p \in P$ and $p \notin ExpertTreeP\}$
    $Factor = 1/\Pi_{(x_i,v_j)\in D}$ [# of values of x_i]
    Let p be the probability of this leaf.
    Let c be the class at this leaf.
    Increment *NumberOfArtExs*[c] by $p \times n \times Factor$
  Else
    Let x_i be the feature tested at *Node*
    If there exists v_j such that $(x_i, v_j) \in P$ then
      Let *Child* be the child of *Node* corresponding to the value v_j of x_i
      ComputeArtExs(P, *Child*, *ExpertTreeP* $\cup\{(x_i, v_j)\}$, n)
    Else
      For each child *Child* of *Node*
        Let v be the value of x corresponding to *Child*
        ComputeArtExs(P, *Child*, *ExpertTreeP* $\cup\{(x_i, v_j)\}$, n)
  Return.

図 1: 与えられた conjunction を満たす仮事例の個数計算手続き.

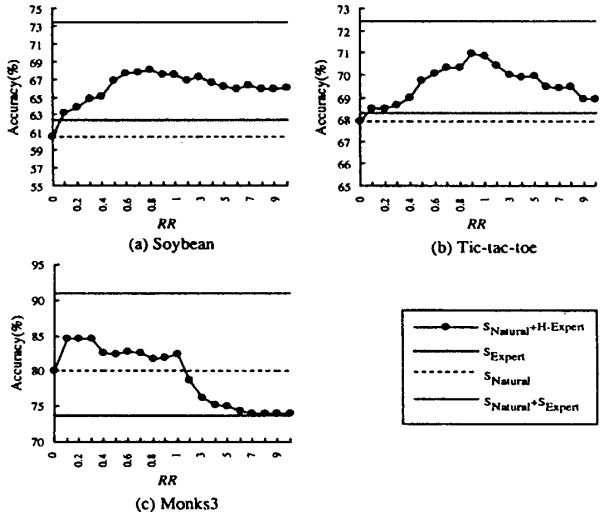

図 2: Soybean, Tic-tac-toe, Monks3 の実験結果.

従来手法では，仮事例を実際に生成しているため，属性に値を入れる必要あり，期待確率 (各属性値が均等に出現) と異なる属性値の分布となる．これに対して，本方法では，期待される属性値の分布をそのままスコア計算に利用できる．アルゴリズムの詳細は文献 [2] に譲る.

4 実験的評価

以上述べたアルゴリズムを，Quinlan の C4.5 を修正して実現した．UCI データベースから Soybean-Large, Tic-tac-toe, Monk3 を選んだ．また，訓練事例の一部から既存知識を作成し，残りの訓練事例を用いて，5-fold cross-validation を行なった.

結果を，図 2 (a), 2 (b), 2(c) に示してある．Soybean では，精度のカーブは，RR= 0.8 でピークを示している．ここで，既存知識と実事例から学習された知識の精度は，実事例のみの場合よりも 7% 高く，既存知識よりも 5% 高い．Tic-tac-toe では，RR= 0.9 でピークを示している．この状況は，Monks3 では，多少異なっている．このドメインでは，既存知識の精度が極めて低い．しかし，最終的な知識の精度は，実事例のみから生成された知識にくらべて 4% の向上を示している.

5 おわりに

仮事例生成による既存知識・事例融合型学習手法について，学習アルゴリズム自身に既存知識を持ち込む手法を提案した．仮事例を実際に生成すると，属性値に具体値を入れなければならないので，事例の統計的性質が偏る．これに対して，本手法では，等確率を保証するので，より正確に既存知識を反映し得る.

参考文献

[1] 石井他, "既存知識と事例に基づく融合型学習手法", 平成 7 年度 *AI* 全大 *7-05*, 1995.

[2] 金田他, "既存知識と事例に基づく融合型学習手法 (II)", 通信学会・*AI* 研究会 平成 *8* 年 1 月, 1996.