3AH－3

束構造を用いた重複概念学習の検討

岩佐 英彦† 馬場口 登‡ 北橋 忠宏‡ 横矢 直和†

†奈良先端科学技術大学院大学情報科学研究科
‡大阪大学産業科学研究所

1 はじめに

クラス情報の付与されていない事例集合から，漸次的に概念を学習する手法を一般に概念形成，あるいは概念クラスタリングと呼ぶ．COBWEB[1] はその代表的なアルゴリズムであり，階層木を用いて概念の階層構造を学習する．しかし，階層木による分類は重複概念を適切に扱えないという欠点があることが指摘されている [2, 3]．重複概念とは 男性」と 成人」という二つの概念のように部分的に事例を共有する概念のことをいうが，事例の集合を異なる観点から分類する場合に必ず現われる概念でありその扱いは重要である．

従来の重複概念の学習手法 [2, 3] は，排他的な概念の対を複数求めるものであり，概念の階層的な構造を適切に表現できるものではない．重複概念の階層構造の特徴として，ある概念が複数の概念に包含されることが挙げられる．GALOIS[4] はこの様な概念階層構造の記述に束を用いているが，考え得る全ての概念を網羅的に学習するアプローチが採用されており，事例数の増加が概念数や概念間のリンク数の爆発的増加をもたらすという欠点がある．本稿では，漸次的に重複概念の階層構造を束として学習する手法を提案しその動作例を示す．

2 重複概念学習手法の概要

2.1 概念対とその更新操作

提案手法においては，排他的な概念の集合 (以下，概念対と呼ぶ) を複数求めることにより，異なる観点に基づき事例を分類する．束構造は概念対の獲得後に概念間の包含関係に基づいて生成される．以下では，事例は定められた属性に対する属性値 (記号データのみ) のベクトル表現として与えられる．ここで概念とは事例の集合のことであるが、その意味は所属する事例によって規定される．本研究では，概念は所属する事例の属性値の出現確率として表現される．

提案手法では，学習過程において新しい事例が与えられると，以下の四つの操作のいずれかが実行される．

(1) 事例を既存の概念対の既存の概念に追加
(2) 事例を既存の概念対の新しい概念として登録
(3) 同一階層に新しい概念対を生成し，事例を新しい概念として登録
(4) 既存の概念に事例を追加した後，概念を部分集合に分割し，下位階層に概念対を生成

2.2 学習過程制御のための評価基準

学習過程において上述の四つの操作を選択するための評価基準として，概念対の集合$\theta = \{S_1 \ldots S_M\}$に対する指標$U(\theta)$を導入する．

$$U(\theta) = \sqrt[M]{\prod_{l=1}^{M} SU(S_l)} \tag{1}$$

$$SU(S_l) = \sum_{\{C_k \in S_l\}} CU(C_k) \tag{2}$$

$$CU(C_k) = \begin{cases} P(C_k|\theta)\times \\ \sum_i \sum_j P(A_i|\theta)P(A_i = V_{ij}|C_k,\theta)^2 & (3a) \\ \sqrt[N]{\prod_{n=1}^{N} SU(S_n)} & (3b) \end{cases}$$

ここで，Mは最初の階層における概念対の総数，C_kは概念対S_l中のk番目の概念，A_iは事例を記述するi番目の属性，V_{ij}は属性A_iの属性値集合の中のj番目の属性値である．また，式 (3a) は概念が下位階層に分割されていない場合，式 (3b) は概念が概念対として分割されている場合に対応しており，Nは概念対の総数である．

式 (3a) で定義される$CU(C_k)$は，ある概念において各属性の属性値の出現確率の分散が大きい場合に大きな値をとる．同一概念対に含まれる概念は互いに排他的であることから，各概念に対応する$CU(C_k)$の値が大きいということは，概念内の事例の類似性が高く，異なる概念間の事例の類似性が低いことを意味する．すなわち，$U(\theta)$の値が大きい概念対の集合は，類似する事例を一つの概念に集約しているといえる．

Learning Overlaping Concepts with Lattice
†Hidehiko Iwasa, ‡Noboru Babaguchi, ‡Tadahiro Kitahashi and †Naokazu Yokoya
†NAIST, 8916-5 Takayama, Ikoma Nara, Japan
‡ISIR, Osaka Univ., Mihogaoka Ibaraki Osaka, Japan

3 学習アルゴリズムと動作例

学習は既存の概念対の集合θに対して新事例が与えられた時に四つの操作のいずれかを適用し，$U(\theta')$を最大にするθ'を得ることに相当する．事例は複数の概念に所属し得るので操作は複数回実行される．

提案手法では，この操作系列を山登り法による探索により求める．つまり，θに対して$U(\theta')$を最大にする操作を選択してθをθ'へと更新し，引続きθ'を対象とした操作の選択を繰り返す．新事例の追加による概念の学習アルゴリズムを以下に示す．

(1) 全ての一階層目の概念対を操作対象として選択
(2) $U(\theta)$を最大にする操作を選択し，θを更新
(3) 事例が追加された概念が属する一階層目の概念対を操作対象から除去
(4) 2,3の操作を$U(\theta)$の値が増加しないか，追加された事例の全属性値が，事例が追加された概念のいずれかにおいて予測可能になるまで反復
(5) 概念の包含関係を探索し，包含関係にある概念間にリンクを張り概念束を生成

ステップ4の条件は，反復処理の停止条件である．ここで，ある事例の属性値V_{ij}が概念C_kにおいて予測可能であるとは，$P(A_i = V_{ij}|C_k)$の値が他のA_iの属性値のC_kでの出現確率よりも大きいことをいう．

提案手法の動作を確認するために，動物に関するデータベース[1]を用いて提案アルゴリズムの動作確認実験を行なった．各事例は動物の生物的特徴を示す16属性により記述されている．図1に，ワシとライオンの事例を含む二つの概念(丸で表現)からなる一つの概念対(四角で表現)が存在し，ヒョウという事例が追加された場合に選択され得る六通りの操作を示す．図中の1,2は既存概念への事例の追加，3は既存の概念対への新概念の登録，4は概念対の生成，5,6は既存概念への事例の追加後の下位階層への概念対の生成を各々表し，それぞれ点線で示される概念や概念対を生成する．

図2に，ライオン，ワシ，ヒョウ，ムクドリ，クマ，ペンギン，カモ，コマドリの8事例から得られた概念階層構造を示す．概念間の線分は束構造における包含関係を表す．概念に記された名前は人間が与えたものである．図より，動物の胎性，生息範囲，飛翔性という異なる観点から重複概念が学習されていることがわかる．COBWEBが学習する階層木ではこの三つの分類を適切に表現することはできない．さらに，陸生動物は二足動物と四足動物という二つの概念へと分割されており，階層木と同様に概念の細分化が実現されている．

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html を参照のこと

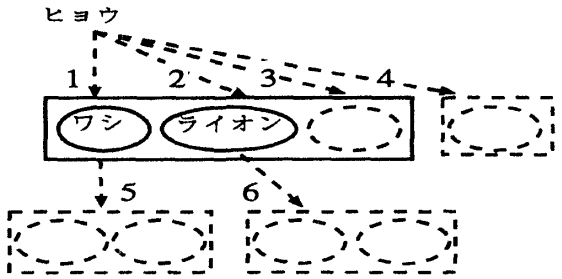

図 1: 事例の追加による概念対集合の更新の様子

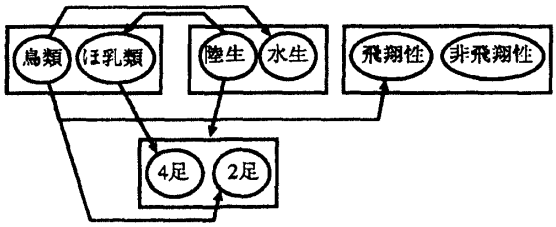

図 2: 学習された概念階層構造の例

提案手法と同様に束を学習するGALOISは，与えられた事例から生成し得る全ての概念から束を構成する．これに対し，提案手法は新概念の追加を$U(\theta)$の増加率と属性値の予測可能性に基づいて制限するため，同じ事例から生成される概念の数はGALOISよりも少ないことが予想される．大規模な事例集合による比較は今後の課題とする．

4 むすび

重複概念を学習するためのアルゴリズムを提案し，その動作例を示した．今後は，未知属性値の予測実験，事例の分類実験の実施や，事例数と概念数の関係の実験的検証が課題である．なお，本研究の一部は文部省科学研究費(奨励A:09780340)の補助による．

参考文献

[1] Fisher, D. H.: Knowledge Acquisition via Incremental Conceptual Clustering, *Machine Learning*, 2, pp.139-172(1987).

[2] Joel, D. Martin: Acquiring and Combining Overlapping Concepts, *Machine Learning*, 16, pp.121-155(1994).

[3] 内田 泰宏, 岩佐 英彦, 馬場口 登, 北橋 忠宏: "重複概念の獲得手法について", 1996年電子情報通信学会総合大会講演論文集, D-152(1996).

[4] Carprineto, C., and Romano, G.: A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval, *Machine Learning*, 24, pp.95-122(1996).