G-20

# 構成的帰納を利用した属性生成規則の獲得

## Acquisition of Feature Generation Rules Using Constructive Induction

大西 悠季生† 大原 剛三† 馬場口 登†
Onishi Yukio Kouzou Ohara Noboru Babaguchi

## 1. はじめに

機械学習などの知識発見技術は，近年急速に発展しつつあるものの，専門家でないユーザはその利用に際して「アルゴリズムの選択」と「データ前処理」という2つの問題に直面することが知られている．このうち「データ前処理」に焦点を当てると，学習システムの多くは，属性として与えられた背景知識を用いて正例と負例を満たす概念を学習するため，所与の属性（以下，初期属性）が学習目標となる概念に対して不適当であると学習は失敗する．そこで統計的手法やドメインに対する知識を用いて初期属性から学習に適切な属性（以下，新属性）を構成することが必要不可欠となる．本稿ではユーザがこのような専門知識を持たない場合でも，適切な新属性を生成可能にする属性生成規則を獲得する手法を提案する．提案手法では，種々のドメインにおける有用な属性生成事例を初期属性の持つ性質(以下，メタ属性)により蓄積することで，その生成規則の獲得を可能にする．

## 2. 提案手法を用いた新属性生成システム

提案手法を用いた新属性生成システムの概念図を図1に示す．提案手法により実現される学習部では構成的帰納と呼ばれる手法を用いて，複数のドメインに対して学習に有用な新属性を生成し，その生成方法を一般化することで属性生成規則を獲得する．生成部では，学習部で獲得した属性生成規則を用いて対象ドメインに対する新属性を生成する．本稿では，学習部と生成部で用いられるドメインをそれぞれ学習ドメイン，適用ドメインと呼ぶ．両ドメインに対してユーザがメタ属性を与えることにより，学習ドメインと適用ドメインが異なる場合においても，メタ属性に共通点があれば学習ドメインにおいて獲得された属性生成規則が適用ドメインにおいて利用可能となる．

新属性生成方法の一般化に関しては，構成的帰納により生成された新属性を属性生成事例として，またユーザに与えられたメタ属性を背景知識として既存学習システムに与える．これにより，メタ属性で記述されたドメインに依存しない属性生成規則を獲得する．本研究では，メタ属性と属性生成事例の記述に柔軟な表現力と数学的裏付けのある論理プログラムを，学習システムとして論理プログラムからの学習が可能な帰納論理プログラミングシステムを用いる．以降では，メタ属性と属性生成事例の生成について説明する．

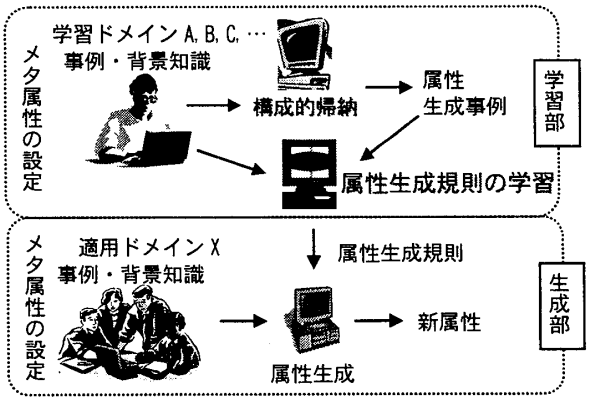
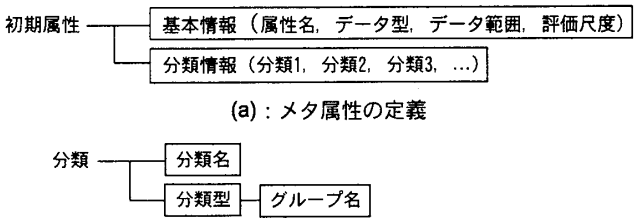



図1：新属性生成システムの概念図




(a)：メタ属性の定義




(b)：分類情報の定義

図2：メタ属性と分類情報の定義

## 3. メタ属性を用いた構成的帰納

### 3.1. メタ属性の定義

メタ属性の定義を図2(a)に示す．メタ属性は，基本情報と分類情報からなる．基本情報が初期属性の属性名やデータ型などを規定するのに対し，分類情報は初期属性の分類を規定する．各分類の定義は図2(b)に示す通りであり，分類名，分類基準を表す分類型，およびその分類型の下で各初期属性が所属し得るグループの種類を表すグループ名からなる．例えば，鳥に関するドメインにおいて，初期属性として｛翼の長さ，翼の重さ，胴体の長さ，胴体の重さ｝がある場合，分類名「g_scale」に対して分類型として「スケール」，グループ名として「長さ」と「重さ」を定義することで，初期属性を「スケール」という基準により2つのグループに分類できる．

このようなメタ属性の設定には統計学や対象ドメインに対する深い知識を必要としないため，データ構築者レベルのユーザでも設定することが可能である．また，メタ属性を用いることで，学習部で行われた新属性の生成過程をドメインに依存しない抽象的な表現で記述することができる．

### 3.2. メタ属性を利用した構成的帰納システム

構成的帰納とは，初期属性から構成演算子を用いて対象ドメインに適切な新属性を構成する手法である．既存の研究では初期属性を AND などの構成演算子で結合するもの[1]から，あらかじめ各初期属性に割り当てられた四則演算や集合演算を用いるもの[2]などが提案されている．本研究では，様々な学習ドメインに対して有用な新属性を生成するために，多様な新属性を生成可能な属性生成構文を構成的帰納に導入する．ここで用いた属性生成構文を図3に示す．属性生成構文は，新属性を生成するための可能な演算と，演算ごとの対象属性の組み合わせを規定するものである．新属性は，属性生成構文中の定義から二値化属性，結合属性，集計属性，変換属性のいずれかとなり，さらに各属性の定義を用いることで当該新属性を生成するための

† 大阪大学 産業科学研究所, *I.S.I.R*, Osaka University

```
〈新属性〉::=〈二値化属性〉|〈結合属性〉|〈集計属性〉|〈変換属性〉
〈二値化属性〉::='binarlizeAttr('〈属性集合〉','〈二値判定子〉','〈定数〉').'
〈結合属性〉::='groupingAttr('〈集合演算〉','〈属性集合〉').' |
              'relativeAttr('〈相対演算〉','〈初期属性〉','〈初期属性〉').'
〈集計属性〉::='compileAttr('〈統計演算〉','〈初期属性〉').'
〈変換属性〉::='transAttr('〈変換演算〉','〈初期属性〉').'
〈属性集合〉::='attrSet('〈グループ名〉').' {'attrSet('〈グループ名〉').'}
〈統計演算〉::='ratio'|'asds'|'rank'
〈集合演算〉::='sum'|'ave'|'max'|'min'|'stdev'|'integ'
〈二値判定子〉::='const'|'over'
〈相対演算〉::='difference'|'increase_ratio'
〈変換演算〉::='log2'|'log10'|'sqr'|'sqrt'
```

図３：属性生成構文

定義式を得ることができる．例えば，メタ属性にグループ g が存在した場合に新属性を結合属性とするならば，定義式 *groupingAttr*(*sum*,*attrSet*(*g*).). を生成することができる．この定義式により規定される新属性は，「グループ g の属性を合計する」という意味となり，ユーザがメタ属性として与えた分類情報が集合演算 sum の集計の対象を決定するのに利用される．

上記のように生成した新属性を帰納論理プログラミングシステムにより一般化するために，各々を 1 つの属性生成事例とみなし，各学習ドメインの正，負例を分類するのに有用である新属性を属性生成規則学習のための正例，そうでないものを負例とする．具体的には，以下の評価関数を用いて各新属性 $n$ を評価する．

$$\begin{cases} f_{posi}(n) = \underset{1\le i\le c(n)}{MAX}\left[(e(n_i)-e'(n_i))\left(1-H\left(\dfrac{e(n_i)}{e(n_i)+e'(n_i)}\right)\right)\right] \\ f_{neg}(n) = \underset{1\le i\le c(n)}{MAX}\left[(e'(n_i)-e(n_i))\left(1-H\left(\dfrac{e'(n_i)}{e(n_i)+e'(n_i)}\right)\right)\right] \end{cases}$$

ただし，

$$H(x) = -x\log(x)-(1-x)\log(1-x)$$

であり，また，$c(n)$は属性 $n$ の属性値の数，$e(n_j)$，$e'(n_j)$はそれぞれ属性 $n$ に対して属性値 $n_j$をもつ正，負例数である．

直観的には，$f_{posi}(n)$ は，学習ドメインのより多くの正例に満たされ，かつほとんどの負例に満たされないような属性値 $n_{MAX}$ をもつ新属性 $n$ に対して高い評価を与える．ここで，$n_{MAX}$ を満たす学習ドメインの正例の集合を $Pn$，$f_{posi}(n)$の高い新属性を順に $n_1$, $n_2$, $n_3$ ,・・・とする．このとき $Pn_i$ の和集合により学習ドメインにおける全正例が被覆されるまで評価値の高い新属性を上位から抽出し，属性生成規則学習のための正例として用いる．同様に，学習ドメインにおける全負例を被覆するまで評価値 $f_{neg}(n)$ が高い新属性を上位から抽出し，それらも属性生成規則学習のための正例とする．また，属性生成規則学習のための負例に関しては，閾値 $\delta$ より評価値の低い新属性を用いる．

## 4. 実験と評価

本章では，提案システムにより獲得される属性生成規則の適用性（実験 1）と属性生成規則の有用性（実験 2）を実験により検証する．実験ではカリフォルニア大学アーバイン校において公開されているデータベースから，学習ドメインとして abalone, balance, iris, mfeat-pix, 適用ドメインとして heart を用いた．属性生成規則の獲得および，分類規則学習の実験には帰納論理プログラミングシステム G-REX[3]を用いた．

```
1: newAttribute(A):-trans_attr(A,log2,B),groupIn(B,g).
2: Attribute(A):=relative_attr(A,difference,B,C),groupIn(B,whole).
```

図４：属性生成規則（実験１）

表１：適合率（実験 2）

| 初期属性 | 初期属性＋新属性 |
|---|---|
| 68.8% | 71.0% |

### 4.1. 実験 1 結果

実験 1 では上記 4 つのデータベースから 624 個（正例 244 個，負例 380 個）の属性生成事例を生成し，それらを用いて G-REX により属性生成規則を獲得した．その結果獲得された属性生成規則は全部で 24 個，そのうちドメイン heart に適用可能な属性生成規則は 7 個であった．この結果から，提案システムにより獲得される属性生成規則の他のドメインへの適用性が確認された．一方，獲得された規則には図 4 の規則 1「グループ g の属性の対数を計算する」のように新属性を数個だけ生成するものだけでなく，図 4 の規則 2「グループ whole の属性と全ての属性の差分を計算する」のように大量の新属性を生成してしまう抽象度の高い規則も獲得された．従って，このような規則を獲得しないような何らかの制約，あるいは適宜規則を取捨選択する手法の検討が必要であるといえる．

### 4.2. 実験 2 結果

実験 2 では，実験 1 で学習された属性生成規則を heart に適用し新属性を生成するとともに，当該新属性の適用ドメインにおける分類規則学習に対する有効性を検証するために，学習に新属性を用いた場合と，用いなかった場合の適合率を 4 重交差検定により比較した．その結果，8 個の新属性が生成され，そのうち 6 個が学習した分類規則に使用されていた．このことから，提案システムにより獲得される属性生成規則は適用ドメインの学習において有用であるといえる．一方，4 重交差検定による適合率は，表 1 に示すような結果となり，新属性を用いることにより適合率が上昇したものの，その差は約 2％程度であった．この結果から，今後，適合率向上のための改善を検討する必要があるといえる．なお，ここでの適合率とは正例として分類された事例数のうち，正しく分類された正例数の割合である．

## 5. まとめ

本稿では，様々なドメインにおいて蓄積された属性生成事例からドメインに依存しない属性生成規則を獲得する手法を提案した．提案手法では，メタ属性を用いることでドメインに依存しない属性生成事例を生成し，これを一般化することで汎用性の高い属性生成規則の獲得を図った．実験により異なるドメインに適用可能な属性生成規則の獲得を確認し，また，属性生成規則の有用性を確認した．

今後の課題として，属性生成構文を拡張し，新属性候補をより多様化させることが挙げられる．

## 参考文献


[1] Y.Hu and D.Kibler, Generation of Attributes for Learning Algorithms, Proc.of AAAI-1996, pp.806-811, 1996

[2] S.Markovitch and D.Rosenstein, Feature Generation Using General Constructor Functions, Machine Learning, vol.49, pp.59-98, 2002

[3] 大原，高，馬場口，北橋，クラス階層における目標概念の一般性を動的に決定するデフォルト規則学習システム，人工知能学会誌，vol.17, No.2, pp.153-161, 2002