# 半順序構造をもつ属性からなる事例への決定木適用と化学物質分析への応用

速水　亜希子†　　田中　栄太朗†　　稲積　宏誠‡
青山学院大学大学院理工学研究科経営工学専攻†
青山学院大学理工学部情報テクノロジー学科‡

## 1. はじめに

各種知識発見技術の中で、構造情報からの知識発見アルゴリズムの開発は重要な検討課題とされている。我々は、特に化学物質の構造情報に注目し、各物質をそこに含まれる部分構造により表現することにより決定木分析を可能とし[2]、有効な決定木分析を実現する試みを行っている[3]。ここで得られる事例表現においては、属性が部分構造であるために、それらの間に包含関係が成り立っている。このような半順序構造をもつ属性に対して、通常の決定木分析を行うことは、分類能力と説明能力を著しく低下させることになる。

本稿では、[3]の取組みをさらに発展させて、半順序構造を持つ属性群から、属性の重複利用を許して全順序構造を満たす属性系列を生成し、これを新たな属性とすることによって決定木分析を行う方法を提案する。特に、生理活性物質データベースから抽出された部分構造を用いて本手法の有用性を示す。

## 2. 系列情報の属性化

半順序構造から重複を許して属性を抜き出し、全順序構造を満たす系列を生成する。その系列を新たな属性、その順序表現を属性値と定義することによって、事例表現を変換することが可能となる。これを決定木分析の対象とすることによって、順序構造のどの部分が特徴分類に寄与するかという情報の抽出が可能となる。

### 2.1 系列情報の抽出法

まず、属性の半順序関係を有向グラフ表現する。そのグラフから、末端ノードをルートとする木構造を抽出する。これによって順序関係を損なわない部分木が生成される。この木構造を線形の系列の組み合わせとみなすことによって各系列を生成し、各系列を属性とする。

この手順は次のとおりである。

step.1　半順序グラフからすべての事例に包含されるノードを削除し、さらに推移的なリンクを削除する

step.2　末端ノードをルートとし、そこからリンクをたどり、木構造を抽出する

step.3　各抽出木について末端ノードを起点としてリンクを辿る。リンクが分岐するときは、相手先ノードをなんらかの統一した尺度にしたがって選択し、ノード系列を抽出する

step.4　step.3 で抽出した系列に含まれるリンクを step.2 で抽出した木構造から削除し、残った部分木に対して、たどるべきリンクがなくなるまで step.2～step.4 をくりかえす。

この結果を用いて、各系列を属性に、系列に対する順序数を属性値とし、事例表現の変換を行う。

このような変換方法により、系列と木構造、木構造と半順序グラフの変換が可逆変換となり、また順序数を割り当てることにより、各ノードと一対一対応し、もとの属性を一意に特定できることになる。決定木生成アルゴリズム C5.0 では連続値属性に対する閾値による分類が自動化されている。そのため各系列に対して最もクラス分類に寄与するであろう構造が特定されることが期待される。また、step.2 で抽出した木構造から抜き出した系列に対して、同じ木構造から抽出されたことを表すように属性表現を工夫することによって、共通の木構造から生じた属性であるという情報を保持させることができる。

### 2.2 ノードの縮約

各ノードの順序関係の中で、隣接する二つのノードに対して、それを含む事例の分布に変化

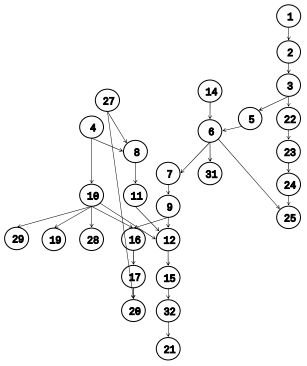

図 2.1　順序グラフ

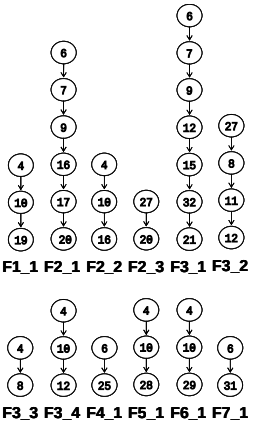


図 2.2　抽出された系列

Decision tree for the example consisting of attributes with partial order structure and application to molecule databases
†Akiko HAYAMI, Eitaro TANAKA　Graduate School of Science and Engineering, Aoyama Gakuin University
‡Hiroshige INAZUMI　School of Science and Engineering, Aoyama Gakuin University

がないときには、この二つのノードの示す情報の変化にはクラス分類の能力はないとみなし、ノードを縮約する。

これによって、クラス分類に寄与しない情報が除去され、順序系列を簡潔に表現することができることになる。これらによって得られた系列は図 2.3 のようになった。

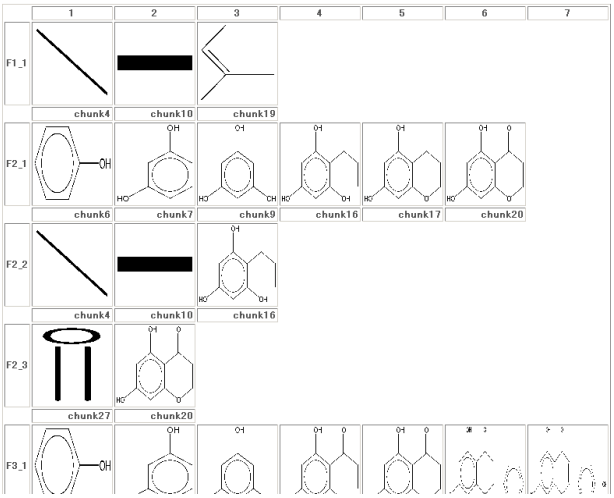


図 2.3　系列情報

## 3. 実験

抗菌活性値により分類されるフラボノイド類 325 種類に対して実験を行って抽出された 32 個の部分構造(Chunk1-Chunk32)によって表現された各物質に対して、本稿で提案した系列情報の属性化を用いて分析を行う。決定木生成アルゴリズムは C5.0 を使用した[1]。

系列情報の属性化の例を図 2.1 と図 2.2 に示す。また、連続値属性に対して重み付けをしながら離散クラスに分割する手法である、事例数変換法も同時に用いた。[3]

系列情報の属性化により、部分構造のどの部分の変化がクラス分類に影響を与えるのかが明示される。また系列間の関係をあらわす値から、決定木の解釈と順序関係との対応が容易に示される。

図 3.1 はこれらの手法を用いて生成した決定木、図 3.2、図 3.3 は決定木から抽出した高活性物質をあらわすルール表現である。系列情報の属性化を行ったことにより差分情報が明示されている。また図 3.4、図 3.5 は各ルールにあてはまる訓練事例の活性度による分布を示している。

## 4. 結論

本稿では、属性間に半順序関係のある事例データに対しての知識発見をテーマとした。これに対して、系列情報の属性化を行うことにより、半順序を全順序に変換し、系列を属性として事例データを変換することによって順序関係を活かした決定木を生成することができた。今後、本手法の理論的な解釈を厳密に行うことと、系列情報を活かした知識表現方法やルール解釈を確立させることについて検討していく予定である。

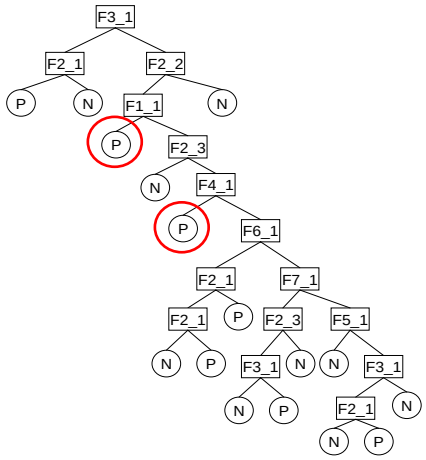


図 3.1　C5.0 により生成された決定木

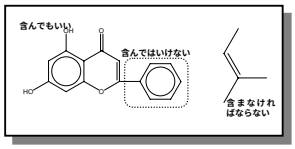


図 3.2 決定木によるルール 1

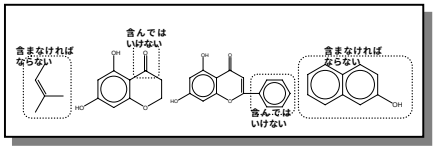


図 3.3 決定木によるルール 2

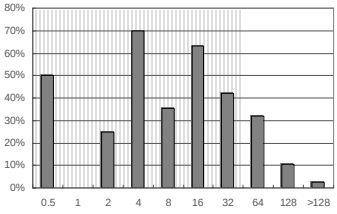


図 3.4 ルール 1 の事例分布

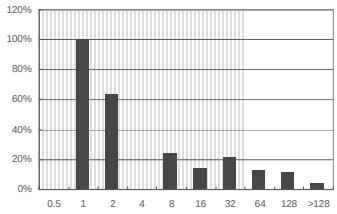


図 3.5 ルール 2 の事例分布

参考文献

[1]J.R.Quinlan.:”C4.5 Programs for Machine Learning” Morgan Kaufmann Publishers 2929 Campus Drive, Suite 260 San Mateo, CA94403

[2] 田中栄太朗、津田哲夫、吉澤有美、稲積宏誠：“化学構造データベースからの有効な部分構造抽出法に関する考察”，情報処理学会第 65 回全国大会，3 巻 pp.137-138（2003）

[3]速水亜希子、田中栄太朗、吉澤有美、稲積宏誠：“化学構造情報を用いた知識発見と知識表現に関する考察”，情報処理学会第 65 回全国大会，3 巻 pp.141-142（2003）