化学構造情報を用いた知識発見と知識表現に対する考察

速水 亜希子† 田中 栄太朗† 吉澤 有美† 稲積 宏誠†
青山学院大学理工学部†

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。本稿では、部分構造の組み合わせとして表現された各化学物質から、いかにして有効な知識を抽出するか、さらにどのような表現方法が有効かについて検討する。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とその包含関係を用いて、決定木により活性度の高い物質の特徴抽出を行う[1][2]。その際に、連続値で表現される被説明属性の取り扱いについて事例数変換法を提案する。さらに、各部分構造がそれぞれ包含関係を有する点に注目し、包含関係を系列情報の組み合わせとして解釈することによって統合化された属性に変換することを提案し、その有効性について検討する。

2. 事例数変換

連続値を被説明変数として決定木学習を行う場合、連続値を離散クラス化する必要がある。閾値をどのように決定するかについては、確率分布からの帰納学習法の中でも検討されているが[3]、閾値付近の事例は本来どちらのクラスに分類されるか曖昧であり、閾値からの距離に応じて事例がクラス分類に貢献する重要度は明らかに異なると考えられる。そこで、[3]を参考にして閾値を決定し、事例ごとに閾値とその被説明属性値との間に距離を定義し、それに応じて仮想的に同一事例を付加することとする。その結果、被説明属性の数値レベルを反映した決定木を生成することができるものと考えられる。閾値の決定方法と距離の定義についてはなお検討の余地があるが、本稿では対象事例の領域知識に応じて定義するものとする。

3. 系列情報の属性化

化学物質から抽出される部分構造にはそれらの間に包含関係が存在する。しかし決定木分析はそれぞれの属性の独立性が前提とされている場合が多い。そこで包含関係を系列情報の組み合わせとして解釈することによって統合化された属性に変換することを考える。

まず、各部分構造の包含関係を包含グラフとしてグラフ表現する。その包含グラフから末端ノードをルートとする木構造を取り出す。これにより包含関係の情報のまったく欠落しない系列ができあがる。しかし、このような木構造を直接属性化するのは非常に困難である。そこで、求められた木構造を、線形の系列の組み合わせに変換し、これを包含系列とする。これは、部分構造に含まれる原子数に対して全順序をもつ系列となっており、各部分構造に対応する原子数を属性値とすることができる。このようにして、与えられた部分構造を、部分構造情報を属性値とする属性に統合することができる。

後述する実験に用いる属性間の包含関係を図3.1 に示す。包含グラフから包含系列を求める手順は次のとおりである。

step.1 推移的なリンクを削除する。
step.2 原子数の最も多い末端ノードからリンクをたどる。
step.3 リンク中で相手先ノードの構造中の原子数の最も多いノードへのリンクを残し、それ以外を除去する。
step.4 相手先ノードに対して step.3 を繰り返す。
step.5 たどるべきリンクがなくなれば、その系列を抽出し、末端ノードがなくなるまで step.2 を実行する。

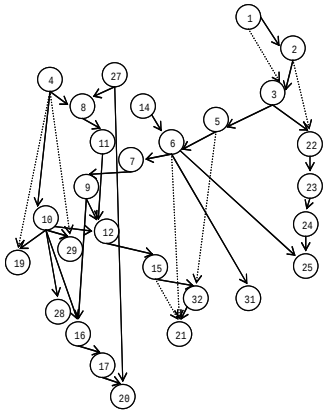

図 3.1 包含グラフ

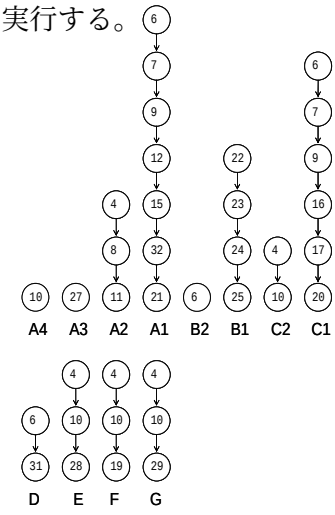

図 3.2 包含系列

The knowledge discovery and the knowledge representation from substructures of molecules
†Akiko HAYAMI, Eitaro TANAKA, Yumi YOSHIZAWA, Hiroshige INAZUMI
School of Science and Engineering, Aoyama Gakuin University.

この結果、各包含系列を属性に、各部分構造に対応する原子数を属性値とみなすことができ、部分構造の組み合わせで表現された化学物質データを包含系列の組み合わせで表現しなおすことができる。原子数を属性値とすることで属性値と各部分構造が一対一対応していることから、属性値から部分構造の特定が可能となる。また、決定木では連続値属性に対して閾値を求めることによって分類基準を作成することができるので、最もクラス分類に影響する構造が特定できることになる。さらに部分構造包含関係の差分情報が明示される形になり、各部分構造の関係に対して評価がしやすくなる。この手法を用いて図 3.1 の包含グラフから求められた包含系列を図 3.2 に示す。

4. 実験

325 種類のフラボノイドに対して GBI 法によって抽出された部分構造を用いて決定木を生成する。決定木生成アルゴリズムは C5.0 および１ノードに複数の属性を置くことのできる領域分割決定木 DTMACC とする。本稿で提案した 2 つの手法を用いて、事例数変換をする場合としない場合、系列情報の属性化をした場合としない場合、それぞれの組み合わせに対してこの二つの決定木生成アルゴリズムを用いて計 8 種の決定木生成実験を行った。事例数変換においては正事例(高活性物質)に対する重みを重視し、高活性物質を特徴的に取り出すように重みを定義した。また、系列情報の属性化については前章で示した図 3.2 の包含系列を系列属性として変換した。

図 4.1 は包含系列を属性とし、さらに事例数変換を行って、C5.0 で生成された決定木である。図 4.2 は包含系列を属性として用い、事例数変換後のデータで C5.0 により生成された高活性物質を表現するルール表現である。また、図 4.3 は各部分構造を属性として用い、事例数変換後のデータで DTMACC による決定木を生成し抽出したルールである。

特に C5.0 では、部分構造そのものを属性として扱った場合には、明らかに各属性が独立してクラスを決定していないため、適切な表現が期待できない。そこで系列情報を属性として表現する効果が示されている。さらに、構造間の差分情報が明示的に表現されているため、活性度の変化に影響を与える部分構造が明示されるようになることがわかる。

5. 結論

本稿では、化学物質が部分構造の組み合わせとして表現され、被説明属性が連続値で表現されている事例データを対象とした知識発見をテーマとした。それに対して、事例数変換により離散値を活かし、系列情報の属性化を用いることによって構造情報の特徴を活かした決定木を生成することができた。抽出されたルールの解釈方法の確立と、発見された知識の活用方法の検討が今後の課題である。

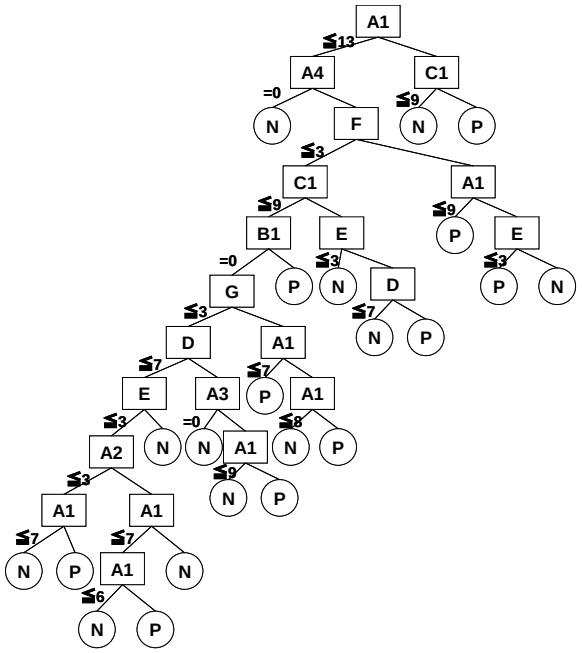

図 4.1 系列情報属性化・事例数変換によって C5.0 による木

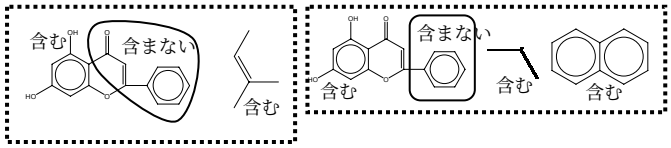

図 4.2 C5.0 系列情報によるルール

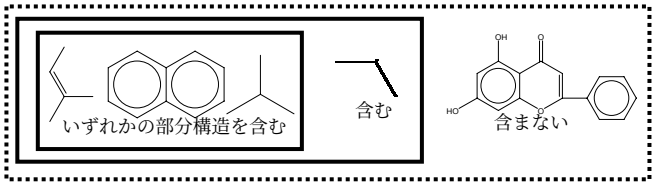

図 4.3 DTMACC 構造情報によるルール

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 (1993)

[2]櫛雄介，稲積宏誠，複合属性による領域分割を用いた決定木 DTMACC，人工知能学会論文誌，Vol.17，No.1,pp.44-52（2002）

[3]森田千絵，月本洋：最尤法と無差別原理を用いた確率分布からの帰納学習，人工知能学会誌 Vol.12，No.2 pp.297-304（1997）