情報処理学会論文誌　数理モデル化と応用　Vol.10　No.3　20–31 (Dec. 2017)

決定木を用いた基礎的有機化合物の発火点決定ルール抽出

林 亮子[1,a)]

受付日 2017年5月25日，再受付日 2017年7月17日，
採録日 2017年7月27日

概要：発火点とは，可燃性の化学物質を空気中で加熱すると自然発火する温度である．発火点は安全性の観点から重要であるため，古くから実験的に調べられている．しかし近年では新規の化合物が大量に生成されており，発火点が予測できると実用上役立つものと考えられる．そこで，著者らは近年データマイニングを用いて発火点の予測を試みている．本稿では，古くから知られていて発火点などの諸量がよく調べられている炭化水素および類例分子を選び，データマイニングの手法の 1 つである決定木を用いて発火点の分類ルールを検討した結果を報告する．入力データに分子量，融点，沸点，炭素原子個数，酸素原子個数および特徴的な部分構造の件数を使用し，統計プログラミング言語 R の決定木パッケージ rpart を用いて決定木を作成した．その結果，今回用いた物質の発火点に対して最も影響力が大きい分岐ルールは，ベンゼン環の有無であることが分かった．さらに，化学的な部分構造の観点から入力データを再検討し，データの追加も行った結果，決定木の過学習状態が改善し，発火点と分子の部分構造に関する経験的に知られたルールを決定木で確認できたので，本稿で報告する．

キーワード：データマイニング，決定木，発火点，分子，有機化合物

Abstraction of Ignition Point Decision Rules for Fundamental Organic Chemical Compounds Using Decision Tree

Ryoko Hayashi[1,a)]

Received: May 25, 2017, Revised: July 17, 2017,
Accepted: July 27, 2017

Abstract: Ignition point is the temperature at which an flammable chemical compound begins to burn naturally. Since ignition points are important from the viewpoint of safety, those are examined experimentally. However, enormous amounts of new chemical compounds are generated recently so that prediction of ignition point will be helpful practically. Therefore I am trying to predict ignition point using data mining. This paper reports a trial prediction of the ignition point for hydrocarbon and similar well-known and well-examined molecules via decision tree; one of the known methods of data mining. I used fundamental material values and the number of characteristic partial structures of molecules as descriptors then I applied "rpart", a package of the statistical programming language "R" in order to make a decision tree. As the result, the most important rule for the ignition point was the number of benzene ring. Furthermore, I reexamined descriptors from the viewpoint of partial structures of chemical compounds and I added more chemical compounds so that I could improve over-fitting situation and could get more known rules between ignition point and molecular structures.

Keywords: data mining, decision tree, ignition point, molecule, organic compound

1. はじめに

近年のデータマイニング技術 [1], [2] は非常な発展をとげ，誰でも簡単にデータマイニングを利用できるようになった．また，データマイニングに代表されるデータ科学は物質科学関連分野においても，実験，理論，計算に次ぐ「第 4 の科学手法」といわれはじめるほど注目されている．本稿ではデータマイニングの物質科学関連分野への応用事例を報告する．

[1] 金沢工業大学
Kanazawa Institute of Technology, Nonoichi, Ishikawa 921–8501, Japan
[a)] ryoko@neptune.kanazawa-it.ac.jp

発火点 [3], [4], [5] は，可燃物を空気中で加熱すると自然発火する温度であり，工業的に重要な量である．そのため，古くから知られた多くの物質において発火点は調べられており，データが蓄積されている．一方，化合物 [6], [7] はすでに数千万種類が知られており，さらに日々新しい物質が生成されている．発火点の調査は危険をともなうため，データマイニングによって新規の物質の発火点予測ができると，安全対策や化合物の有効利用の観点から実用的にも役立つものと考えられる．

化学分野においては，計算機の黎明期から化学物質の性質を調べるために計算機の積極的な利用が試みられており，ケモインフォマティクス [8] と呼ばれる分野が形成されている．ケモインフォマティクスでは，分子の性質を表す量を記述子と呼ぶ．記述子設定の方法論についてはまだ定説はなく，データ分析の目的によって必要な記述子を検討する必要がある．ケモインフォマティクス分野においても発火点の予測は試みられている [9], [10]．しかし，分子の性質や構造と発火点の関係については，まだ不明な点が多い．

本稿では，主要なデータマイニング手法の 1 つである決定木を用いて発火点予測を試みる．決定木を使用すると発火点の分岐ルールが得られ，記述子の発火点における寄与度を議論することができるため，先行研究とは異なる観点から発火点を議論できる可能性がある．本稿の著者グループは，これまでに文献 [11], [12] において発火点を調べたが，十分な予測精度が得られなかった．そこで，記述子を再検討するとともにデータ件数を追加した結果を本稿で報告する．

本稿では，多様な化学物質の中でも炭素原子，水素原子，酸素原子のみで構成された分子を扱う．これらの分子は化学物質の中でも構造が簡単な基本的なものであり，古くから存在が知られていて，性質もよく調べられている．既知の知見と分岐ルールや予測結果を比較することで，データマイニングが適切に行われたかどうかを確認できる．

化学では「基」という概念があり，複数の分子で特徴的に見られる原子の組合せと結合形態をまとめたグループを指す．基は分子の性質と関係が深いことが知られている．原子の種類に従って多様な基が存在するので，扱う原子の種類を限定することによって，考慮すべき基の種類を限定し，問題の複雑さを調整することができる．今回は扱う原子を 3 種類に限定したが，それでもなお分子中の部分構造として 10 種類以上の記述子が考えられる．

近年では，決定木よりも過学習が起こりにくい手法が開発されていて，代表的なものにランダムフォレストがある [2]．ランダムフォレストではより良い予測精度が得られることが期待でき，記述子の重要度も得られる．しかし，たとえば「ベンゼン環の有無で発火点は平均 150°C 異なる」というような，発火点を記述子の値で予測するような具体的なルールは得られない．分子の部分構造を示す記述子を用いた決定木で得られる発火点予測ルールは，発火点と分子部分構造の関係を示し，一定の価値があるものと考えられる．そこで本稿では，決定木作成により，記述子を用いた発火点予測ルールの抽出を試みる．

2 章以降の本稿の構成を述べる．2 章は類例研究を紹介する．3 章はデータの作成方法を説明し，記述子の設定内容を述べる．4 章では決定木を作成し，得られた発火点の分類ルールを議論する．5 章は本稿で得られた結果をまとめ，今後の課題を述べる．

2. 類例研究

まず，定量構造活性相関分野における先行研究を紹介する．Tsai らは 820 件のデータに 4 個の記述子とその線形結合式を用いて最大誤差 89 K，平均誤差 36 K で有機化合物の発火点を予測した [9]．Tsai らは，計算化学により得られる大量の数値データから，記述子として有効なものを 4 個選択している．それらは平均電子トポロジ状態指数（Mean electrotopological state），芳香率（Aromatic ratio），回転結合分数（Rotatable bond fraction），中心原子フラグメントであり，詳細はほかに譲るが，いずれも数値計算によって得られる量である．

Shi らは分子を部分的に分割した 8〜12 種類のフラグメントを記述子とし，265 件の有機化合物データを使用して，実際の発火点との誤差が 50 K から 90 K 程度で発火点を予測した [10]．Shi らの使用した記述子は「In Silico design and Data Analysis」（ISIDA）と呼ばれるツールを用いて作成しており，発火点を線形モデルで近似できるように，分子中の原子列および結合列を機械学習を用いて決定している．

これらの研究は発火点の予測精度の観点からは一定の成果をあげている．一方で，これらの研究はいずれも線形式で予測モデルを構成しており，発火点の予測モデルは簡素であるが，自然の複雑さを係数と記述子で表現しているものと考えられる．そのため記述子には，分子の基礎的な量をそのまま使用するというよりは，基礎的な量を組み合わせて作成する複雑な量を用いる傾向があり，係数を数値的に決定する．そのため，発火点決定ルールの可読性，すなわち分子の基礎的な物性量や構造情報と発火点との関係に関する知見としては，議論の余地があるものと考えられる．そこで本研究では，記述子には分子の基礎的な量を用いて，近年発達の著しいデータマイニング手法を利用して予測モデルを統計的に決定し，分子の基礎的な物性量や構造情報と発火点との関係を人間が理解しやすいルールとして得ることを目指す．

岡田と，本稿著者である林は炭化水素と類例分子 21 種類の分子量，融点および沸点をデータに用いて自己組織化マップを作成し，分子間の類似度を利用した発火点の予測

を試みた [11]．その結果，炭素原子が単結合だけで直鎖型に結合したアルカンでは，ある程度の予測が可能であったが，芳香族の発火点予測はできなかった．そこで，林と中田は炭素，酸素，水素のみから構成されている炭素原子 10 個以内の有機化合物に注目して 245 種類の学習データと 10 種類の試験データを用意し，発火点を決定木によって調べた [12]．しかし，得られた決定木は多くの部分が過学習であり，改善の余地があった．

3. 使用データの概要

3.1 分子データの作成法

本章では，使用したデータの概要を述べる．本研究では発火点予測を目的とするため，実用上十分な精度で発火点がすでに決定した物質に関する情報を収集する必要がある．発火点の情報をまとめて公開しているものに「国際化学物質安全性カード」[4] がある．これは，化学物質を安全に利用するための重要な情報を公共に提供することを目的とした ICSC プロジェクトによるもので，世界保健機関（WHO），国連環境計画（UNEP），国際労働機関（ILO）の共同事業である国際化学物質安全性計画（IPCS）が作成している．日本においては国立医薬品食品衛生研究所が担当機関であり，ICSC の日本語版を作成して公開している．

ICSC は，単独の分子だけを含む純物質についてデータを作成することを基本としているので，本研究の目的にも適合する．本稿でデータに含める物質は，国際化学物質安全性カードにおいて，分子の元素組成を示す組成式，沸点，融点，発火点が記載されているものを中心とする．なお国際化学物質安全性カードでは，発火点は「発火温度」と記載されている．

本稿では記述子として，特徴的な結合や基がその分子に含まれる個数を使用する．それらを元素の種類とその比率のみを示す組成式から決定するのは困難であるため，分子構造を表示する構造式を目視し，さらに SMILES 記法 [8] を参照して決定した．SMILES 記法は次節で紹介する．構造式と SMILES 記法は，科学技術振興機構（JST）が作成する「日本化学物質辞書 Web」[6] を参照した．沸点，融点，構造式，SMILES 記法のいずれかがこれまで紹介したウェブサイトに掲載されていない物質が 10 件程度あったため，wikipedia，「職場のあんぜんサイト」[5] および「Chemical Book」[7] を補助的に利用した．

本研究では発火点が既知の物質を扱うが，発火点よりも融点と沸点のほうがむしろ物質の性質を示す基本的な量であるため，発火点が既知の物質では，融点と沸点がすでに調べられていることが多い．そのため，本研究では融点，沸点，発火点に欠損値を含む分子は基本的に扱わないこととする．また，本研究の結果得られる決定木を利用して発火点を予測する際には融点と沸点が既知であることを前提とするが，それでもなお多くの物質を扱うことができる．

本稿のデータ作成で行った例外的な扱いを述べる．

昇華する物質　一部の物質は，固体を加熱すると固体から直接気体になる「昇華」を起こす．そのため，大気圧程度の状態では昇華点のみが知られている物質がある．その場合は，昇華を「液化と気化が同時に同じ温度で起こる」と解釈し，沸点と融点の両方に昇華点を用いた．この扱いにより，たとえば身近な物質で炭素間の三重結合を持つアセチレンをデータに加えることができる．

混合物質　一部の物質は，異性体が混合した状態で発火点を決定していて，その場合は構造が一意に決まらない．構造が大きく異なる異性体は一般に物理的性質も大きく異なって分離精製が可能であり，多くの場合は異性体を分離して発火点を決定している．一方混合状態で発火点を決定する場合は，たとえば鏡像異性体などのように，沸点や融点の差異が小さく，特徴的な部分構造が共通する物質が混合していて，実用的な分離が困難である．そのような場合は，異性体でも本研究で使用する基や結合の種類は共通するのでそのまま扱って問題はない．そのため，可能な範囲で構造式を調査し，基や結合の記述子を決定した．

沸点の欠損　物質の中には，発火点は決定していても，加熱すると分解してしまうために沸点が決定できない物質もある．そのような物質は今回使用しないものとした．

国際化学物質安全性カードでは，化学式や分子を構成する原子個数を用いた検索ができないため，登録されている物質の化学式を目視してデータに含めるかどうかを決定する必要がある．なお，国際化学物質安全性カードに登録されている物質は，2016 年 2 月 18 日現在で 1,702 物質である．本稿では，炭素原子，酸素原子および水素原子のみを含む分子を対象とするが，たとえば物質名に「クロロ」が含まれている分子は必ず塩素原子を含むために，調査対象から除外することができる．そのため，国際化学物質安全性カードに登録済みの物質名一覧表を用いて，対象となる可能性がある全物質の登録を目視で確認することは，現実に可能である．

発火点予測性能を改善するために，まず文献 [12] のデータ内容を確認した結果，同一分子の二重登録や欠損値があった．それらを削除すると文献 [12] のデータ件数は 240 となった．同時に入力ミスや分子の部分構造の錯誤があったため，データの修正を行った．また，データ内容を確認する過程で，全物質の登録内容を確認し，データ件数の増加も試みた．文献 [12] では炭素原子個数を 10 個以内としたが，今回は炭素原子個数が 10 個を超える分子も含めることとした．その結果，44 件の物質を新規に追加することができた．

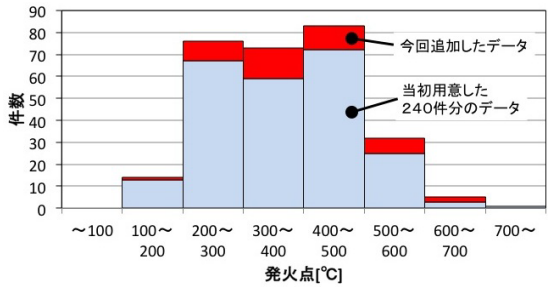

図 1　使用データにおける発火点の分布

Fig. 1　Histogram of ignition point for original 240 datas and additional 44 datas.

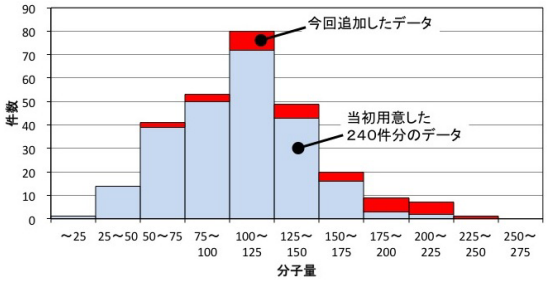

図 2　使用データにおける分子量の分布（この図の範囲外となる分子量 275 以上の分子が 9 件ある）

Fig. 2　Histogram of molecular weight for original 240 datas and additional 44 datas.

3.2　使用データの内容

本稿では，文献 [12] で使用した 240 件のデータと 44 件の新規データを扱うので，最大で 284 件のデータを使用する．本節では，今回用いたデータの性質を示す諸量の分布を述べる．まず発火点の分布を**図 1** に示す．図 1 は積み上げ棒グラフで，下方に文献 [12] で用意した 240 件の発火点の度数分布を示し，その上に今回追加したデータ 44 件分の分布を積み上げた．図 1 によると，最も温度の低い発火点でも 100°C 以上であり，最高の発火点は 700°C 台であった．ほとんどの物質は 200°C から 500°C の範囲にある．また，類例研究の発火点分布とおおむね同様である．追加したデータの発火点は特別な分布の偏りはなく，発火点に関しては本稿は文献 [12] とおおむね同様の分布となる．

次に，分子量の分布を**図 2** に示す．文献 [12] では炭素原子個数を 10 個以内と制限したが，今回追加したデータでは炭素原子個数に制限を設けないこととした．なお，図 2 に示した範囲よりも大きい分子量を持つ分子が 9 件存在する．それらは件数が少なく，図 2 に記入すると度数分布が見にくくなるため，本稿では記入を省略している．図 2 によると，分子量 100 から 125 までの分子が最も多く，その階級から離れるにつれて，件数は減少する．今回追加したデータでは，分子量が大きい階級で追加データ件数が多い．しかし，追加件数は相対的に少ないので，分子量の分布に大きな変化はない．

284 件のデータの類似度を可視化するため，発火点を除き，後述する記述子を用いて自己組織化マップを作成した．自己組織化マップは多次元データを 2 次元に投影するもので，2 層のニューラルネットで競合学習を行い，データ間の類似度を求めるものである．自己組織化マップでは，データ間の類似度距離に意味があって，類似度を適切に投影するために 2 次元空間の軸を設定するものであるから，縦軸と横軸には具体的な意味はない．自己組織化マップは，R では som パッケージ [1] を用いて作成することができる．

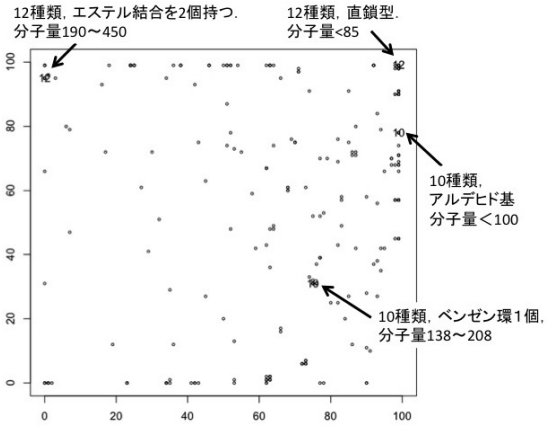

図 3　284 件の学習データで作成した自己組織化マップ

Fig. 3　Self-organizing map from 284 molecules' learning data.

284 件のデータの 18 個の記述子を入力とし，100 × 100 の正方格子のユニットを第 2 層に用いて競合学習した結果を**図 3** に示す．図 3 では，1 つの分子を 1 つの点で表現している．自己組織化マップでは同一のユニットに所属する分子が重ならないよう乱数でユニット領域内に分散して描画することが多いが，今回はユニットを多数使用していてほとんどの分子が異なるユニットに所属しているため，同一のユニットに所属する分子をそのまま重ねて描画する．今回作成した自己組織化マップでユニットが複数の分子を持つ場合，ほとんどは 5，6 個程度以内である．一方で 10 個以上の分子が同じユニットに所属する例が 4 カ所あったため，それらに共通する記述子上の特徴を図 3 に書き込んだ．図 3 によると，エステル結合，アルデヒド基，直鎖型およびベンゼン環の個数が，今回扱った分子の中で分子 10 個以上のグループを形成する記述子であることが分かる．

3.3　分子の構造と記述子

分子の構造から記述子を決めるために SMILES 記法を併用するので，その概要を紹介する．化合物の 1 つであるジケテンは組成式が $C_4H_4O_2$ であり，SMILES 記法で表すと C1(=O)OC(=C)C1 となる．詳細は文献 [8] などに譲るが，SMILES 記法は分子の化学構造を文字列で表現するもので，原子を元素記号で表し，二重結合を記号「=」，三重結合を記号「#」で表し，構造の主要部分からの分岐を () 内に入れて表す．環構造は結合部の原子に識別番号を付す．ジケテンでは，最初の炭素原子が環構造の起点で，その炭

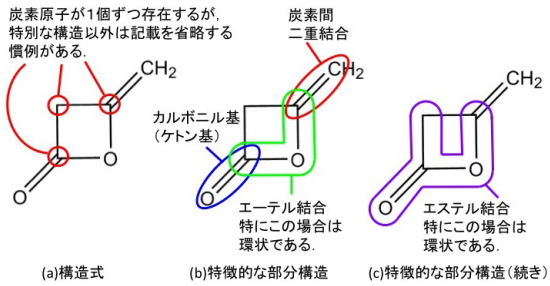

図 4　構造式と特徴的な部分構造の例（ジケテン $C_4H_4O_2$，SMILES 記法では C1(=O)OC(=C)C1）

Fig. 4　An example of structural formula and characteristic partial structure for diketene.

素原子に酸素が二重結合で枝分かれして結合しており，その炭素原子に酸素原子 1 個，次に炭素原子が 2 個単結合して最初の炭素原子に戻ることで環構造ができている．水素原子は 4 個含まれるが，結合可能な箇所に結合しているものとして省略して扱い，SMILES 記法では明記しない．

SMILES 記法では識別番号を用いて環構造を明示するため，SMILES 記法を併用することで環構造の個数を確定することができる．また，芳香族構造をなす原子は小文字で書くこととなっているため，通常の環構造と区別することができる．一方で，SMILES 記法は 1 つの化学構造に対して一意に決まらないことと三次元構造を記述できないことが欠点であり，近年では様々な改良版が提案されているが，歴史的な経緯もあり，多くのデータベースでは現在も SMILES 記法が記載されている．

次に，構造式と記述子の関係を説明する．構造式は，分子構造を簡略化して平面上に図示したものである．基本的には構造式は，情報科学で「グラフ」と呼ばれるデータ構造と考えることができ，節点が原子またはよく知られた部分構造に相当して，結合関係にある原子間を辺で結んだものである．辺が直線 1 本であるときは単結合であり，二重結合，三重結合の場合はそれぞれ二重線，三重線で辺を表現する．節点は基本的に原子記号を直接用いて記載するが，慣例として構造式では自明と考えられる水素原子の記載を省略し，また炭素原子も特別な場合を除いて辺の折れ曲がりで存在を表現する．構造式は原子位置の三次元情報を含まないので，立体構造を扱うことはできないが，本稿の範囲では大きな問題はない．

ジケテン $C_4H_4O_2$ を例として，構造式と特徴的な部分構造を**図 4** に示す．図 4 (a) のように，ジケテンは 3 個の炭素原子と 1 個の酸素原子が環構造をつくり，酸素が環構造の炭素原子 1 個と二重結合している．また，環構造のもう 1 つの炭素原子に CH_2 が二重結合している．

このジケテンが含む特徴的な部分構造を図 4 (b) および図 4 (c) に示す．図 4 (b) に示すように，炭素原子に酸素原子が二重結合している部分をカルボニル基（ケトン基と呼ばれることもある）という．また，炭素原子に CH_2 が二重結合している部分は炭素間二重結合として扱う．さらに，カルボニル基の炭素原子に酸素原子が結合し，その先に結合している炭素原子までも 1 つの特徴的な部分構造であり，エーテル結合と呼ばれている．エーテル結合からさらに原子が結合した先は，もとのエーテル結合のもう一方の原子に結合しているため，ジケテンは環状エーテルである．図 4 (b) に示した部分構造は互いに独立している．

またジケテンでは，図 4 (c) に示すように，エーテル結合の片方の炭素原子とカルボニル基が直接結合している構造を持つが，これらをまとめてエステル結合と呼ぶ．このように，分子の特徴的な部分構造は，入れ子の関係を持つ場合があり，記述子として部分構造を使用する際には，入れ子の関係を持つものの取扱いを決める必要がある．ジケテンでは，エステル結合から先がもとのエステル結合に戻るため，環状エステルでもある．

本稿における分子の特徴的な部分構造の取扱いを説明する．本稿ではベンゼン環，炭素間二重結合，炭素間三重結合，水酸基，アルデヒド基，カルボニル基，エーテル結合，環状エーテル結合，カルボキシル基，エステル結合，環状エステル結合，環構造，直鎖型構造の数を記述子とする．これらの記述子で，炭素，酸素，水素からなる化合物が持つ特殊な構造の主要なものを網羅している．これらは構造式を目視して数を判定するが，SMILES 記法をあわせて確認することで誤りを減らすことができる．また，すでに述べたようにいくつかの構造は他の構造を包含しているが，発火点に対して上位階層の構造が影響する場合と，下位階層の構造が単独で影響する場合のいずれも可能性があるので，包含関係にある構造はそれぞれの階層で重複して数えるものとする．

本稿および文献 [12] で採用した手法に基づいて作成した図 4 のジケテンのデータを**表 1** に示す．表 1 では，文献 [12] 中での記述子の扱いを「旧版」，本稿での扱いを「新版」として両方を示した．表 1 に示すように，本稿では分子の性質を表す連続値として分子量，沸点，融点，発火点を用いる．そして，分子の特徴的な原子個数として炭素原子個数，酸素原子個数を用いる．さらに，記述子として特徴的な部分構造の個数を用いる．構造が存在しない場合は 0 とする．環状構造，枝分かれがない構造である直鎖構造およびアルデヒド基も発火点に影響する可能性があるので，本稿では新規に加えた．本稿は「決定木を用いて発火点の決定ルールを調べる」という考え方であるため，入力データには予測対象である発火点を含む．本稿で扱うデータでは，以上の内容を欠損値なしで持つ．

ジケテンは図 4 に示したように環状エステル構造を 1 個持つ．旧版の記述子では包含関係にある記述子では上位階層の記述子だけを数える．そのため表 1 の旧版ではエステル結合だけを 1 としており，カルボニル基とエーテル結合は数えず 0 としている．一方，新版の記述子では包含関係

表 1　使用データ例（環状エステル，ジケテン $C_4H_4O_2$，旧版で定義していない記述子は「-」と記載する）
Table 1　An example of input data for diketene.

記述子		旧版	新版
分子量		84.1	84.1
沸点 [°C]		127	127
融点 [°C]		−7	−7
炭素原子個数		4	4
酸素原子個数		2	2
ベンゼン環		0	0
炭素間二重結合	C=C	1	1
炭素間三重結合	C#C	0	0
水酸基	−OH	0	0
アルデヒド基	−CHO	-	0
カルボニル基	−C(=O)−	0	1
エーテル結合	−O−	0	1
環状エーテル		-	1
カルボキシル基	−COOH	0	0
エステル結合	−COO−	1	1
環状エステル		-	1
環構造		-	1
直鎖構造		-	0
発火点 [°C]		275	275

にある記述子を重複して数えるため，エステル結合，カルボニル基，環状エーテル，エーテル結合，環構造のすべてが 1 である．

4. 発火点を分類する決定木

4.1 決定木と発火点予測精度の関係

まず，実用上必要と考えられる要求精度を評価する．発火点測定方法は国際規格で決められており，たとえば液体試料の発火点測定方法は文献 [3] で紹介されている．文献 [3] によると，液体試料の発火点測定手順においては，あらかじめ設定温度を定めて容器を加熱したのちに，液体試料を注入して鏡を経由して容器内を目視し，発火するかどうかを確認する．そのため実験においては，たとえば ±50°C 程度までであれば容認できる誤差であると考えられるが，100°C 異なると，実験環境の耐熱性能などの事前の想定に影響するものと考えられる．そこで本研究では，要求誤差を平均二乗誤差 50 K とする．また，誤差の最悪値は小さいほうが望ましいので，最大誤差の絶対値を参考にする．

決定木では，根ノードに全学習データが所属し，葉ノード以外の各ノードにおいて分岐ルールがあり，その分岐ルールに従って学習データを 2 つのグループに分ける．そして，1 つの葉ノードは，最終的な分類結果の 1 つのグループを表す．そのため，「決定木を用いて発火点を予測する」ということは，「予測したい分子を葉ノードのどれかに割り当て，その葉ノードに所属する学習データの平均発火点を予測発火点とする」ということである．発火点を良好な精度で予測するには，それぞれの葉ノードが持つ平均発火点の差分が要求誤差程度になる必要がある．3.2 節で述べたように，今回扱う学習データでは，発火点は 100°C から 700°C の範囲に分布しており，特にほとんどの物質は 200°C から 500°C の範囲にある．要求誤差が平均二乗誤差 50 K であるため，決定木は葉ノードを 10 個程度以上持つことを目標とする．

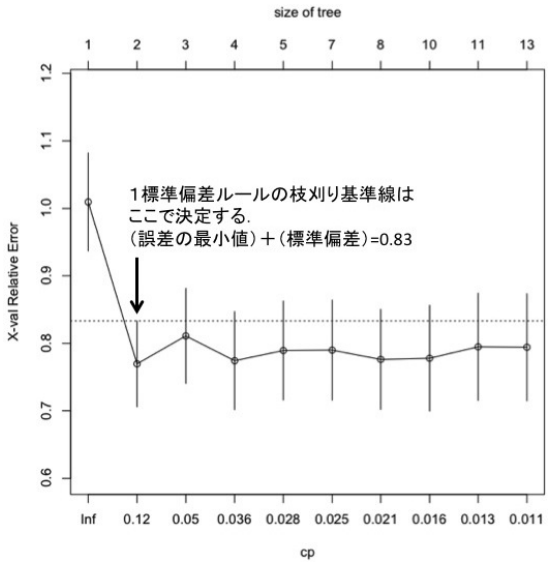

図 5　デフォルト設定で作成した決定木における複雑度と交差判定誤差の関係（旧版記述子，学習データ 240 件）
Fig. 5　Relationships between complexity parameter and cross validation relative error on the decision tree from old descriptors and 240 learning datas.

4.2 旧版記述子を用いた決定木

まず，240 件の分子で旧版の記述子を用いたデータによる決定木を作成する．決定木を作成すると過学習が起こっていることが多いので，文献 [1] の手順に従い，枝刈りを検討する．今回使用する rpart パッケージ中の関数 rpart は，決定木作成時にデータセットをランダムに分割して交差判定を行い，デフォルト設定では 10 分割交差判定を行う．そして，関数 plotcp は枝刈りに必要な情報を図示するのに用いる．

240 件の分子で旧版の記述子によるデータを入力し，デフォルト設定で作成した決定木における plotcp の出力を**図 5** に示す．図 5 は下側の第 1 横軸が複雑度 cp，上側の第 2 横軸が決定木の葉ノード個数であり，縦軸は相対交差判定誤差である．デフォルトでは $cp = 0.01$ まで分岐を行うため，図 5 の第 1 横軸右端は $cp = 0.01$ である．複雑度 cp は木が成長するとおおむね減少するが，単調減少することは保証されておらず，増加することもある．

決定木では，一標準偏差ルールと呼ばれる経験則があり，（相対交差判定誤差の最小値）＋（標準偏差）を枝刈りの基準値としてよく用いる [1], [2]．一標準偏差ルールについて

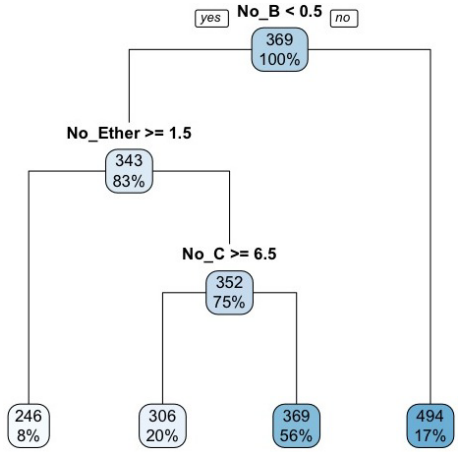

図 6　葉ノードを 4 個採用した枝刈り後の決定木（旧版記述子，学習データ 240 件）

Fig. 6　Pruned decision tree from old descriptors and 240 learning datas.

は特に文献 [2] pp.185–187 に詳しいのでそちらに譲り，本稿では結果のみを使用する．図 5 では相対交差判定誤差が最小となるのが葉ノード 2 個のときであり，その点の相対交差判定誤差に標準偏差を加えた 0.83 が枝刈りの基準値となり，図 5 中の水平線として表示されている．図 5 中の各点は標準偏差をエラーバーとして記入しているため，図 5 では葉ノード 2 個の点におけるエラーバーの上端が枝刈りの基準値と一致している．この事情は，本稿の以降の同様の図にも共通する．一標準偏差ルールに従うと，基準値の水平線より下になる最小の葉ノード数までを採用し，それよりも先の分岐は過学習によるものとして採用しないことで枝刈りを行う．

図 5 によると，葉ノード数 2 で相対交差判定誤差はいったん基準値よりも明確に小さくなり，その後基準値を超えることはないが，基準値付近にとどまった状態である．図 5 において基準値を厳格に適用すると，葉ノード 2，すなわち根のみの分岐で木の成長を止めることとなる．しかし，この基準値と複雑度は試行によりある程度の数値変動があり，図 5 では相対交差判定誤差は葉ノード数 3 でいったん増加して葉ノード 4 以降はおおむね最小値付近で安定するので，ここでは葉ノード数 4 までを採用した木を参考に示す．

240 件の分子で旧版の記述子を用いて作成した，葉ノード数 4 までを採用した決定木を**図 6** に示す．図 6 では 3 つの分岐があるが，根に近い分岐ほど重要なルールである．図 6 で各節点内の上側の数値はその節点に属する分子の平均発火点であり，下側はその節点に属する分子が全データに占める割合をパーセント表示したものである．そのため，根ノードでは上側は全データの平均発火点，下側は 100%となる．各節点の直上に分岐ルールが記載されており，そのルールを満たす分子は左子へ，満たさない分子は右子へ分類される．

図 6 は入力ファイルで用いた記述子名をそのまま用いた表示であるため，根ノードの分岐ルールから始め，子ノードに向かってノードをたどる順番に説明を加えると，以下のようになる．

ルール 1：ベンゼン環を持たない分子は全体の 83%あり，平均発火点は 343°C，ベンゼン環を持つ分子は 17%あって平均発火点は 494°C である．

ルール 2：（ベンゼン環を持たない分子のうち）エーテル結合を 2 個以上持つ分子は全体の 8%あって，平均発火点は 246°C である．エーテル結合を 1 個持つ，または持たない分子は全体の 75%あって平均発火点は 352°C である．

ルール 3：（ベンゼン環を持たない，エーテル結合が 1 個以下の分子のうち）炭素原子個数が 7 個以上の分子は 20%あり，平均発火点は 306°C，炭素原子個数が 6 個以下の分子は 56%あって平均発火点は 369°C である．

以上のルールのうち，ベンゼン環の有無は文献 [12] でも根ノードのルールとなっており，共通した結果となっている．2 つ目のルールはエーテル結合個数を用いているが，これは文献 [12] では上位になかったルールであり，データの錯誤を修正したため現れたルールと考えられる．3 つ目のルールは炭素原子個数を用いているが，その本質は分子サイズと考えられ，数値的に差が出やすい記述子としてたまたま炭素原子個数を採用したものと考えられる．炭素原子個数のほかに分子サイズに関連の深い記述子は，分子量，沸点，融点であることが知られている．文献 [12] でも分子サイズを用いたルールが根の右子と左子の両方に現れており，分子サイズは重要度の高い記述子であることが分かる．

4.3　旧版記述子と新版記述子の比較

次に，新版の記述子による 240 件の分子の学習データを用いて決定木を作成する．まずデフォルト設定で作成した決定木における plotcp 関数の出力を**図 7** に示す．図 7 では，葉ノード 7 個で相対交差判定誤差が最小となるため，枝刈りの基準値はこの点で決まり，0.71 となる．図 7 によると，一標準偏差ルールに従っても 4 個の葉ノードが利用できるので，葉ノード 4 個で枝刈りを行う．新版の記述子では，一標準偏差ルールに従っているので，図 6 の決定木よりも過学習の影響が少ないことが期待できる．

新版の記述子による，枝刈り後の決定木を**図 8** に示す．図 8 のルールは以下のとおりである．

ルール 1：ベンゼン環を持たない分子は全体の 83%あり，平均発火点は 343°C，ベンゼン環を持つ分子は 17%あって平均発火点は 494°C である．

ルール 2：（ベンゼン環を持たない分子のうち）カルボニル基を持たない分子は全体の 55%あって，平均発火点は 314°C である．カルボニル基を 1 個以上持つ分子は全体の 28%あって平均発火点は 400°C である．

ルール 3：（ベンゼン環を持たない，カルボニル基を持たな

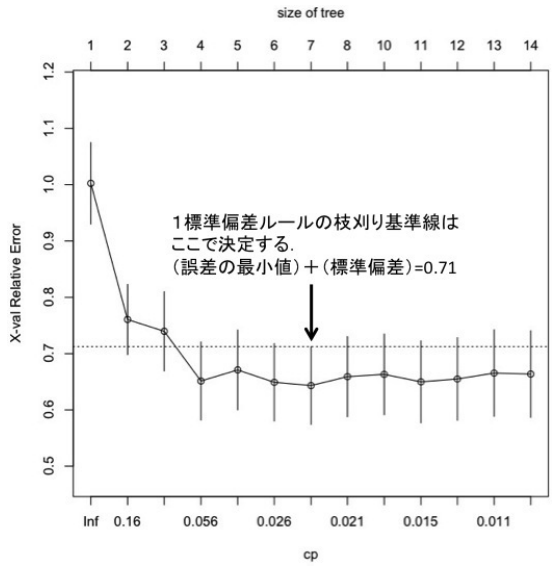

図 7　デフォルト設定で作成した決定木における複雑度と交差判定誤差の関係（新版記述子，学習データ 240 件）

Fig. 7　Relationships between complexity parameter and cross validation relative error on the decision tree from new descriptors and 240 learning datas.

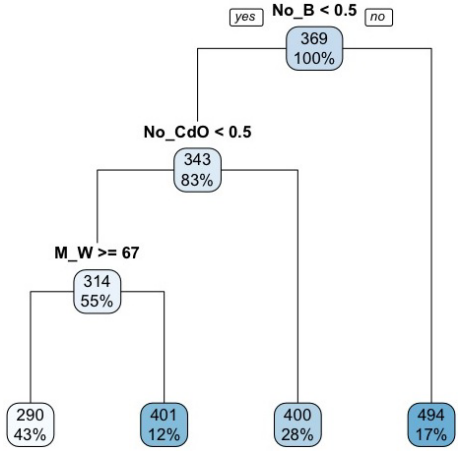

図 8　枝刈り後の決定木（新版記述子，学習データ 240 件）

Fig. 8　Pruned decision tree from new descriptors and 240 learning datas.

い分子のうち）分子量が 67 以上の分子は 43%あり，平均発火点は 290°C，分子量が 67 未満の分子は 12%あって平均発火点は 401°C である．

ルール 1 は旧版記述子で得られたものとまったく同じであり，ベンゼン環の扱いは旧版と新版でまったく同じであるために同じ結果になっている．2 個目のルールは，部分構造に関する記述子の扱いが旧版と異なるために，新規に現れたルールと考えられる．3 個目のルールは旧版記述子で得られたものと同様に，本質は分子サイズに関するルールであるものと考えられる．

4.4　学習データを増加して作成した決定木

次に，分子データを追加する効果を検討する．新版記述子を使用し，284 件の学習データを用いて決定木を作成する．まずデフォルト設定で作成した決定木における plotcp 関数の出力を**図 9** に示す．図 9 では，葉ノード 12 個で相対交差判定誤差が最小となるため，枝刈りの基準値はこの点で決まり，0.62 となる．図 9 によると，一標準偏差ルールに従っても 5 個の葉ノードが利用できるので，葉ノード 5 個で枝刈りを行う．44 件のデータを追加した結果，過学習の影響が少ないルールが 1 つ追加で得られ，葉ノードを 1 個追加することができる．

284 件の分子による，枝刈り後の決定木を**図 10** に示す．この決定木で追加できたルールはベンゼン環を持つ分子に関するルールであり，図 8 における根の右子に相当する．

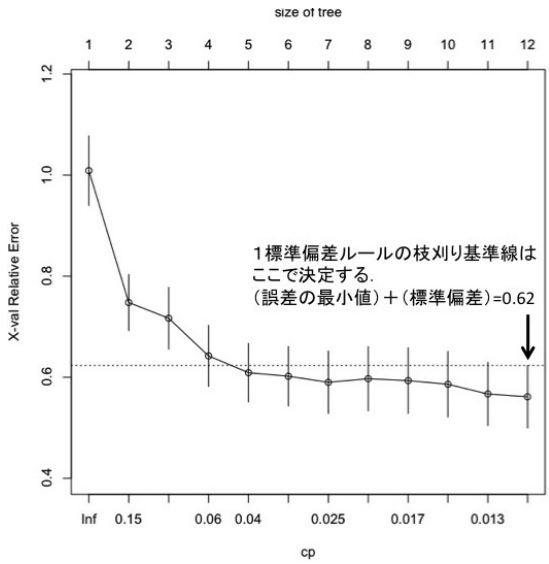

図 9　デフォルト設定で作成した決定木における複雑度と交差判定誤差の関係（新版記述子，学習データ 284 件）

Fig. 9　Relationships between complexity parameter and cross validation relative error on the decision tree from new descriptors and 284 learning datas.

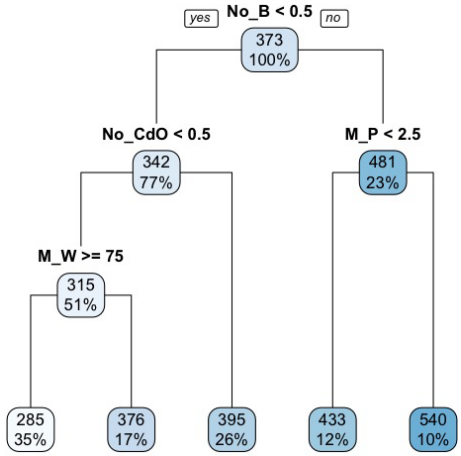

図 10　枝刈り後の決定木（新版記述子，学習データ 284 件）

Fig. 10　Pruned decision tree from new descriptors and 284 learning datas.

図 8 と対比するために，以下では図 8 中のルールに相当するものを同じルール番号とし，図 10 で新規に追加できた，根の右子ノードのルールを「ルール 4」とする．図 10 のルールは以下のとおりである．

ルール 1：ベンゼン環を持たない分子は全体の 77%あり，平均発火点は 342°C，ベンゼン環を持つ分子は 23%あって平均発火点は 481°C である．

ルール 2：（ベンゼン環を持たない分子のうち）カルボニル基を持たない分子は全体の 51%あって，平均発火点は 315°C である．カルボニル基を 1 個以上持つ分子は全体の 26%あって平均発火点は 395°C である．

ルール 3：（ベンゼン環を持たない，カルボニル基を持たない分子のうち）分子量が 75 以上の分子は 35%あり，平均発火点は 285°C，分子量が 75 未満の分子は 17%あって平均発火点は 376°C である．

ルール 4：（ベンゼン環を持つ分子のうち）融点が 2.5°C 未満の分子は 12%あり，平均発火点は 423°C，融点が 2.5°C 以上の分子は 10%あって平均発火点は 540°C である．

ルール 1 はこれまでの決定木で，特に過学習の影響が少なく得られたものとまったく同じであり，かなり確度の高いルールであると考えられる．ルール 2 およびルール 3 も，図 7 と数値の若干の差異はあるが，同様のルールである．この決定木作成に際して追加した 44 件のデータのうち，ベンゼン環を持つ分子が 23 件あったため，特にベンゼン環を持つ分子に関して過学習の影響が少ないルールが新規に得られたものと考えられる．

4.5　分子サイズに関する記述子を統合する効果

本稿でこれまで作成した決定木では，分子サイズに関するルールとして複数の記述子が使用される現象があった．そこで，分子サイズに関係が深いと考えられる記述子を 1 つだけ選択して残し，決定木の作成を試みる．本稿で使用した表 1 の新版記述子のうちで，分子サイズに関係が深い記述子は分子量，沸点，融点，炭素原子個数である．分子量は直接分子サイズを示す量であるため，基本的に分子量を使用するものとし，他の量と分子量の関係を確認する．

図 11 は分子量と沸点の関係を示す散布図である．図 11 では，分子量が 400 を超える分子では沸点が低くなる傾向があるが，特に分子量が 300 程度までは分子量と沸点の間に正の相関がありそうである．全データを用いた分子量と沸点の間の相関係数は 0.74 であり，正の相関があるものと考えられる．そのため，記述子として沸点の傾向を分子量で代替可能であるものと考えられる．

図 12 は分子量と融点の関係を示す散布図である．図 12 では，分子量が 200 程度以下かつ融点が 100°C 程度以下の分子とそれ以外の分子では傾向が異なることが分かる．全データを用いた分子量と融点の間の相関係数は 0.28 であり，相関は明確ではないが，分子量が 200 以下かつ融点が 100°C 以下の分子に限ると相関係数は 0.43 であり，弱い正の相関があるものと考えられる．融点に関しては，単純に線形回帰はできないものと予想されるが，分子量で代替することを試みる．

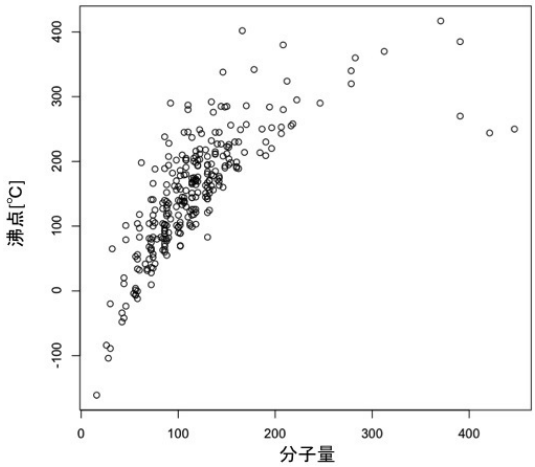

図 11　分子量と沸点の散布図（すべてのデータを用いた相関係数 = 0.74）

Fig. 11　Scatter plot of molecular weight and boiling point.

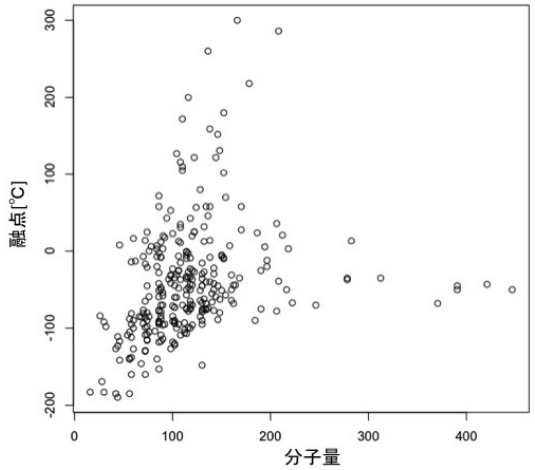

図 12　分子量と融点の散布図（全データを用いた相関係数 = 0.28，分子量 200 以下かつ融点 100°C 以下のデータを用いた相関係数 = 0.43）

Fig. 12　Scatter plot of molecular weight and melting point.

図 13 は分子量と炭素原子個数の関係を示す散布図である．炭素原子個数は 1 から 30 程度までの離散値であるため，図 13 は階段状に見える．有機化合物では分子量のほとんどは炭素原子個数で決まるため，図 12 では，分子量と炭素原子個数の間に高い正の相関が確認できる．全データを用いた分子量と炭素原子個数の相関係数は 0.94 であり，高い正の相関がある．そのため，炭素原子個数は分子量で代替可能であるものと考えられる．

以上の議論から，記述子のうち融点，沸点および炭素原子個数を分子量で代替するものとし，これらを削除した記

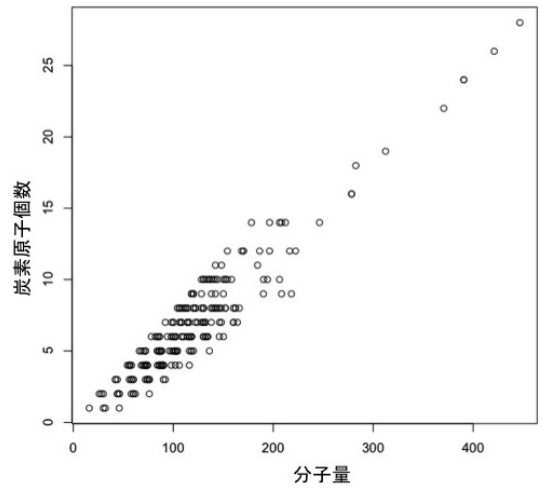

図 13　分子量と炭素原子個数の散布図（すべてのデータを用いた相関係数 = 0.94）

Fig. 13　Scatter plot of molecular weight and the number of carbon atoms.

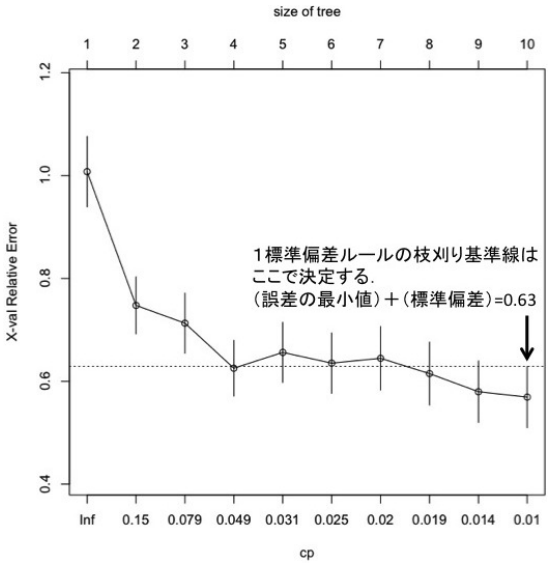

図 14　デフォルト設定で作成した決定木における複雑度と交差判定誤差の関係（新版記述子から融点，沸点，炭素原子個数を削除，学習データ 284 件）

Fig. 14　Relationships between complexity parameter and cross validation relative error on the decision tree from intensive descriptors and 284 learning datas.

述子セットを以後「集約記述子」と呼ぶものとする．集約記述子を用いて決定木を作成する．まずデフォルト設定で作成した決定木における plotcp 関数の出力を**図 14** に示す．図 14 では，葉ノード 10 個で相対交差判定誤差が最小となるため，枝刈りの基準値はこの点で決まり，0.63 となる．図 14 によると，葉ノードが 4 で一標準偏差ルールの基準値に非常に近いので，ここまでが安定したルールであって，以降のルールは過学習の影響が比較的大きい可能性がある．明確に基準値を下回ったのは葉ノード 8 個であるため，今回は葉ノード 8 個で枝刈りを行う．

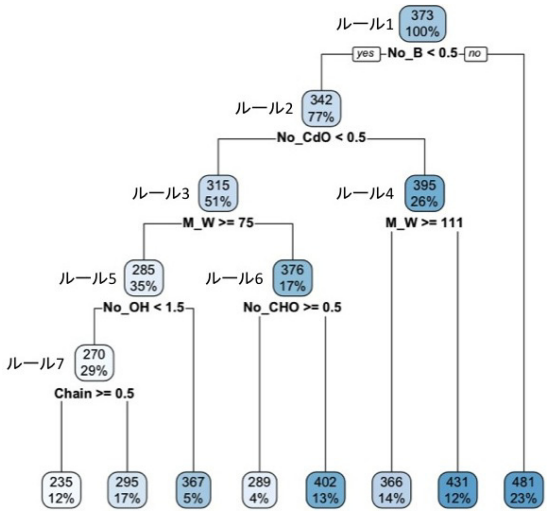

図 15　枝刈り後の決定木（新版記述子から融点，沸点，炭素原子個数を削除，学習データ 284 件）

Fig. 15　Pruned decision tree from intensive descriptors and 284 learning datas.

284 件の分子による，枝刈り後の決定木を**図 15** に示す．図 15 中のルール 1 および 2 は，新版記述子を用いたこれまでの決定木で得られたルールと比べて，数値の変動はあるが同じ記述子を使用している．また，ルール 3 はこれまでの決定木では分子量か融点を用いたルールであり，分子サイズに関するルールと考えると同等であると考えられる．そのため，ルール 1 から 3 については，新版記述子を用いたこれまでの決定木とルールの内容が同様であるため説明を省略する．図 15 で新規に得られたルールは以下のとおりである．いずれも，ベンゼン環を持たないことを前提とする．

ルール 4：カルボニル基を持つ分子のうち，分子量が 111 以上のものは全体の 14%あって，平均発火点は 366°C である．分子量が 111 未満の分子は全体の 12%あって平均発火点は 431°C である．

以下のルールは，ベンゼン環とカルボニル基を持たないことが前提である．

ルール 5：（分子量が 75 以上の分子のうち）水酸基を 1 個持つかまたはまったく持たない分子は 29%あり，平均発火点は 270°C，水酸基を 2 個以上持つ分子は 5%あって平均発火点は 367°C である．

ルール 6：（分子量が 75 未満の分子のうち）アルデヒド基を持つ分子は 4%あり，平均発火点は 289°C，アルデヒド基を持たない分子は 13%あって平均発火点は 402°C である．

ルール 7：（分子量が 75 以上で水酸基を 1 個持つかまたはまったく持たない分子のうち）直鎖構造を持つ分子は 12%あり，平均発火点は 235°C，直鎖構造を持たない分子は 17%あって平均発火点は 295°C である．

図 15 で得られた発火点予測ルールは，これまでの決定

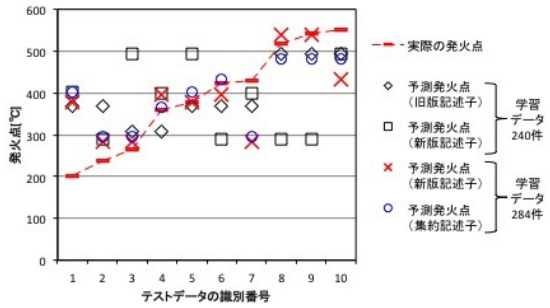

図 16　テストデータを用いた発火点予測結果

Fig. 16　Prediction results of decision trees using four types of descriptors for ten test datas.

木に比べて分子の部分構造に関するものが多くなったが，それは分子サイズに関する記述子を統合した効果であるものと考えられる．一方で，枝刈りの議論では，4 個を超える葉ノード数において，顕著な誤差改善は見られなかったので，ルール 4 以降は過学習の影響が大きい可能性がある．

次に，図 15 の決定木の発火点予測性能を検討する．今回用いたデータの発火点分布を示す図 1 によると，発火点の多くが分布するのは 200°C から 600°C であるため，発火点予測に用いるという観点からは，この範囲にある葉ノード個数が多いことが望ましい．図 15 で得られた葉ノードは，200°C から 400°C の間に 5 個ある．しかし，左から数えて 2 番目と 4 番目の葉ノードの差異は 6°C であり，3 番目と 6 番目の葉ノードの差異は 1°C だけであるため，発火点予測に利用する観点からは効果が小さい．400°C から 500°C の間には 3 個の葉ノードがあるため，この温度範囲では，ある程度の発火点予測性能が期待できる可能性がある．

4.6　発火点予測性能

本稿でこれまで示した 4 つの決定木を用いて，学習データからあらかじめ試験用に除外したテストデータ 10 件の発火点を予測した結果を**図 16** に示す．図 16 は横軸を 10 件のテストデータとし，縦軸を発火点とする．テストデータはあらかじめ発火点の低いものから高いものに並べたのち使用している．実際の発火点は赤色の横線マークで示し，見やすくするため破線で結合した．予測発火点は互いに異なるマークで示した．

図 16 中の菱形のマークは，学習データ 240 件で旧版記述子を用いた決定木による予測発火点であり，発火点が 300°C 以下で誤差が大きいが，その他は実際の発火点の ±50°C 程度の範囲に入る，比較的良好な結果が得られている．図 16 中の正方形のマークは，学習データ 240 件で新版記述子を用いた決定木による予測発火点であるが，10 件のテストデータのうち 4 件で 200°C 程度の誤差があり，さらに 2 件で 100°C 程度の誤差がある．また，明らかに他の予測発火点よりも誤差が大きいように見受けられる．図 16 中の赤色の交差マークは，学習データ 284 件で新版記述子を用いた決定木による予測発火点であり，10 件のテストデータのうち 7 件で ±50°C 程度の良好な予測を示しており，学習データ 240 件の結果よりもおおむね予測性能は向上したように見受けられる．図 16 中の円形マークは，学習データ 284 件で新版記述子から沸点，融点および炭素原子個数を除いた集約記述子を用いた決定木による予測発火点であり，10 件のテストデータのうち 8 件で ±50°C 程度の良好な予測を示している．一方で，残り 2 件は 100°C または 200°C の大きな誤差がある．

表 2　テストデータを用いた発火点予測における二乗平均誤差の平方根と最大誤差の絶対値

Table 2　RMSE and maximum absolute error of the prediction results.

使用した決定木（学習データ件数と記述子の種類）	二乗平均誤差の平方根	最大誤差の絶対値
240 件，旧版記述子	79.0	167
240 件，新版記述子	157.0	253
284 件，新版記述子	83.9	174
284 件，集約記述子	85.4	200

図 16 の 10 件のテストデータを用いた発火点予測結果における，平均二乗誤差と最大誤差を**表 2** に示す．表 2 によると，学習データが 240 件の 2 つの決定木では，旧版記述子のほうが平均二乗誤差と最大誤差の絶対値の両方とも新版記述子よりも小さい．新版記述子でも学習データが 284 件の決定木は平均二乗誤差と最大誤差の絶対値の両方とも学習データが 240 件の結果よりも改善されている．新版記述子は旧版よりも 5 個記述子が増加しており，学習データ件数が 240 件の場合は，記述子個数に対して学習データ件数が小さいことが影響している可能性がある．集約記述子を用いた決定木では，平均二乗誤差と最大誤差の絶対値はどちらも，同じ学習データ件数で新版記述子を用いた決定木よりもむしろ増加している．集約記述子を用いた決定木では，分子量と相関が小さい融点を削減していることが影響している可能性がある．

表 2 の結果は，いずれの決定木でも平均二乗誤差が 50 K を超えており，まだ要求精度に達していない．文献 [10] が本稿と同程度の 265 件のデータで 10 種類程度の記述子を用いて，誤差が 50 K から 90 K 程度である．文献 [10] が発火点予測性能を優先していることを考慮すれば，本稿の結果は発火点決定ルールの可読性を優先して決定木を用いた中では最良の結果の 1 つとはいえそうである．文献 [9] では 820 件のデータと 4 個の記述子を用いて最大誤差 89 K，平均誤差 36 K で発火点を予測している．本稿では記述子の個数が多いため，文献 [9] と同程度の性能を得るためには，炭素，酸素および水素のみからなる分子の発火点データが 2,000 件程度以上必要であると考えられる．または，発火点への影響が少ない記述子を削減することも有効であ

ると考えられる．データの追加を検討するとともに，記述子を選択する手法を検討したい．

表 2 に示した 4 つの決定木では，旧版記述子の決定木が最も過学習の影響が大きいと考えられ，集約記述子の決定木が最も過学習の影響が少ないと考えられる．しかし，今回発火点の予測性能が最良であったのは旧版記述子による決定木であった．今回使用したテストデータは文献 [12] と同じものを使用しており，今回得られた予測性能はテストデータの選び方に依存する可能性もあるため，予測性能については今後も引き続き検討したい．

5. おわりに

本稿では，R の rpart パッケージを用いて炭化水素および類例分子の発火点を分類する決定木を作成した．各分子を特徴付ける記述子には，分子量，融点，沸点などの物理量に加えて特徴的な部分構造の個数を使用した．得られた決定木は過学習の影響が大きかったため，特徴的な部分構造の種類を増やし，部分構造どうしの包含関係の扱いを見直して記述子を再設定した．その結果，過学習の影響を減らした決定木を得ることができた．さらに，扱う分子を 44 件追加し，分子サイズに関する記述子を 4 個から 1 個に集約した結果，決定木を改善することができた．

今回得られたルールのうち上位の 3 つは，データの増減や記述子の増減を行っても繰り返し現れたため，炭素，酸素，水素のみからなる分子の発火点に関してはかなり確からしいルールといえそうである．それらはベンゼン環の有無，カルボニル基の有無，分子サイズに関するルールで，ベンゼン環や分子サイズと発火点の関係はある程度知られており，さらにカルボニル基も有機化学では反応性の高い重要な基として知られている．旧版記述子ではカルボニル基に関するルールは得られなかったが，新版記述子では繰り返し得られているので，知られたルールを発見できるかどうか，という観点からも新版記述子のほうが旧版よりも適切であると考えられる．

一方で，発火点予測精度の観点からは，実用上の要求精度と考えられる二乗平均誤差 50 K はまだ達成できていない．「決定木を用いて発火点を予測する」ということは，「いずれかの葉ノードに予測したい分子を割り当てる」ということであった．そのため，発火点を良好な精度で予測するには，葉ノードが 10 個程度以上あって，葉ノードが持つ平均発火点の差分が要求精度程度になる必要があった．現段階では枝刈り後の葉ノード個数が少なく，葉ノードが代表する平均発火点の分布が偏っているため，発火点を予測することはまだ困難である．今後は，葉ノードの平均発火点が発火点予測に適した分布になる状態を目指したい．また，過学習に強いランダムフォレストなどの他の手法の利用も検討する．

本稿で扱った有機化合物は，炭素，酸素，水素のみを含むものであったが，これは有機化合物の中でもかなり限定された範囲であり，実際には窒素や硫黄を含む有機化合物が多数存在して，それらの多くは可燃物である．扱う元素の種類を増やすと記述子の種類も増えるため，データの増え方と記述子の増え方が発火点予測に与える影響を調査することも興味深い．今後も引き続き調査を行いたい．

謝辞　本研究の一部は科研費 16K13739 の助成を受けて行われた．関係各位に感謝する．

参考文献

[1] 豊田秀樹：データマイニング入門，東京図書株式会社 (2008).
[2] 平井有三：はじめてのパターン認識，森北出版株式会社 (2012).
[3] 国立研究開発法人産業技術総合研究所安全科学研究部門：消防法危険性確認試験データベース 6 発火点測定，入手先 ⟨http://explosion-safety.db.aist.go.jp/INFOMATION/shoubou_3_6.htm⟩（参照 2017-07-16）.
[4] 国立医薬品食品衛生研究所（NIHS）：国際化学物質安全性カード（ICSC）日本語版，入手先 ⟨http://www.nihs.go.jp/ICSC/⟩（参照 2016-03-17）.
[5] 厚生労働省：職場のあんぜんサイト，入手先 ⟨http://anzeninfo.mhlw.go.jp/⟩（参照 2016-03-17）.
[6] 科学技術振興機構（JST）：科学技術総合リンクセンター J-GLOBAL，入手先 ⟨http://jglobal.jst.go.jp/⟩（参照 2016-03-17）.
[7] Chemical Book, available from ⟨http://www.chemicalbook.com/⟩(accessed 2016-03-17).
[8] Gasteriger, J. and Engel, T.（編），船津公人，佐藤寛子，増井秀行（訳）：ケモインフォマティクス—予測と設計のための化学情報学，丸善株式会社 (2005).
[9] Tsai, F.-Y., Chen, C.-C. and Liaw, H.-J.: A model for predicting the auto-ignition temperature using quantitative structure property relationship approach, *Procedia Engineering*, 45, pp.512–517 (2012).
[10] Shi, J., Chen, L., Chen, W.: Prediction on the auto-ignition temperature using substructural molecular fragments, *Procedia Engineering*, 84, pp.879–886 (2014).
[11] 岡田　彩，林　亮子：競合学習を用いた炭化水素分子の類似度マップ，平成 26 年度電気関係学会北陸支部連合大会，講演論文集 F27 (2014).
[12] 中田侑江，林　亮子：炭化水素および類例分子の発火点決定木，情報処理学会研究報告，Vol.2016-MPS-107, No.14 (2016).

林 亮子　（正会員）

東北大学理学部物理学科卒業．北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（情報科学）．1998 年北陸先端科学技術大学院大学情報科学研究科助手．2005 年金沢工業大学講師，2017 年より准教授．この間，高性能計算，可視化，データマイニングに関する研究に従事する．IEEE-CS，電子情報通信学会，可視化情報学会，人工知能学会，日本化学会情報化学部会各会員．