Vol.2016-MPS-107 No.14
2016/3/9

炭化水素および類例分子の発火点決定木

中田 侑江[1]　林 亮子[1,a)]

概要：発火点とは，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．発火点は主に実験的に調べられているが，実験条件への依存性が大きく，測定が困難な量である．また，発火は非定常な状態で起こる現象であるため，シミュレーションを用いて調べることは困難である．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分析と予測を試みている．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一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．入力データは分子量，融点，沸点，炭素原子個数，酸素原子個数およびいくつかの官能基の有無，特徴的な結合の有無を含むものとし，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わかった．さらに，得られたルールに基づいて分子の発火点を予測したところ，多くの場合は誤差 50 度程度で発火点を予測できた．

キーワード：決定木，発火点，分子

1. はじめに

近年ではデータマイニング技術 [1], [2], [3] が成熟し，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．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を試みた結果を報告する．現在確認されている化学物質 [4], [5] の数は約 2000 万種類以上あり，その数は日々増え続けている．新規の物質は詳細な性質が不明なことも多いため，データマイニング技術を用いて，未調査の性質を予測できると有用であると考えられる． そこで本稿では，統計プログラミング言語 R の決定木パッケージを用いて，実験およびシミュレーションが困難な量である発火点 [6], [7] の予測を試験的に行った結果を報告する．

発火点は測定条件によって変化し，例えば試料の加熱時間や加熱速度，測定条件中の空気の成分，圧力，器壁の状態などの影響を受けるため，発火点は物質固有の一意に定まる定数というよりは，ある実験環境の観測値として得られる．また本研究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，炭素が 10 個以下の分子を扱う．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．性質がわかっている分子を扱って，これまでに知られている性質および基礎的な調査結果と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．

一方化学分野においては，計算機の黎明期から計算機を利用して化学物質の性質を調べる分野が存在し，ケモインフィマティクス [8] と呼ばれている．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．記述子として何を用いると最適なのかは，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の観点から発火点を予測する試みも行われている [9] [10]．本稿では，主要なデータマイニング手法の一つである決定木を用いる．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，本稿では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．

本稿では「国際化学物質安全性カード」[5] から得られるデータを使用する．今回使用した分子は 250 種類程度あり，データマイニングの対象として最低限のデータ個数は得られる．そしていくつかの特徴的な官能基と結合を扱うことができる．

第 2 節以降の本論文の構成を述べる．第 2 節は類例研究を紹介する．第 3 節は本稿における問題設定内容を説明し，使用したデータの詳細を述べたのちに本稿が用いる統計プログラミング言語 R と決定木を紹介する．そして，決定木

[1] 金沢工業大学
KIT, 3-1 Yatsukaho, Hakusan, Ishikawa 924–0838, Japan
[a)] ryoko@neptune.kanazawa-it.ac.jp

を作成した結果を示し，適切な木について議論する．第 4 節は決定木を用いて発火点を予測した結果を述べる．第 5 節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

Tsai, Chen, Liaw は定量構造活性相関 (QSPR) の手法を用いて発火点予測を行った [9]．文献 [9] では 4 個の記述子とその線形結合式を用いて最大誤差 $89K$，平均誤差 $36K$ で有機化合物を混合した物質の発火点を予測した．岡田と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，発火点の予測を試みた [10]．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，ある程度の発火点予測が可能であった．一方で，ベンゼンは構造の似たシクロヘキサンよりも発火点は $100K$ 以上高いが，その予測ができなかった．

3. 発火点を予測する決定木

3.1 本稿の問題設定

本稿では発火点予測を目的とするため，最初に，現在知られている化学物質の燃焼開始過程モデル [6], [7] を説明する．化学物質が燃焼を開始するときは，化学物質を加熱することによって温度が上がり，沸点を超えて化学物質が気体になった後に燃焼を始める．そのため，沸点は燃焼に関連する重要な量である．燃焼開始時に口火と呼ばれる小さな火花が存在すれば引火と呼ばれ，口火がなくても燃焼を始める場合を自然発火という．発火点は自然発火が起こる温度であり，引火が起こる温度は引火点である．一般に発火点のほうが引火点よりも高温である．そして，炭化水素および類例物質の燃焼では，炭素と酸素が反応して二酸化炭素および一酸化炭素が生成される．そのため，炭素原子と酸素原子が燃焼現象における重要な原子である．

本稿では，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，炭素が 10 個以下の分子を扱うが，分子を炭素原子 10 個以下に限定した理由を述べる．発火点が調べられている分子は炭化水素および類例分子が多い．今後研究の進展にともなって，発火点とシミュレーション結果との関連を調査する可能性があり，シミュレーションにおける問題サイズは炭素原子個数でおおむね定義される．炭素原子 10 個程度であれば量子化学シミュレーションの計算モデルで詳細なものを用いても，数十分程度以内で計算できる．そのため，まず炭素原子が 10 個以下の分子を主に調べることとした．

次に，酸素原子を含む分子を対象に加えた理由を述べる．林と岡田が文献 [10] において発火点の予測を試みた際には，炭素原子と水素原子のみを含む 20 種類程度の炭化水素類例分子を扱った．しかしデータマイニングの対象としてはデータ件数が少なく，また，より正確な発火点の予測のためには，分子の種類にもある程度の幅が必要であると考えられた．また，近年のケモインフォマティクスでは，化学反応を官能基や結合の種類などの物質の化学構造から定量的に予測しようとする定量構造活性相関が盛んに研究されている．燃焼も化学反応であるため，分子の官能基と発火点についても何らかの関係がある可能性がある．対象物質に酸素原子を含む分子を加えることで扱う官能基の種類も増え，対象となる分子の種類も大幅に増加するため，本稿では酸素原子を含む分子を加えた結果を検討する．

本稿で使用するデータの中核部分は「国際化学物質安全性カード」[5] として公開されているものを使用する．このデータは沸点，融点，発火点（[5] では「発火温度」と記載されている）を含むが，発火点の詳細な実験データは記載されていない．これまで述べた問題設定をすると，分子は 250 種類程度あり，データマイニングの対象として最低限のデータ量は得られる．そして官能基は 3 種類，結合は 4 種類扱うことができる．データの詳細は第 3.2 節で述べる．

本稿では，統計プログラミング言語 R[1], [2], [3](以後，「R」と記す) を使用する．R はオープンソースであるため，これまで数多くのパッケージが開発され，公開されている．特に近年はデータマイニング手法を実装したパッケージが数多く公開されている．本稿ではパッケージ rpart[1], [2], [3] を用いて決定木を作成する．パッケージ rpart は決定木を作成する関数 rpart および関連するいくつかの関数を含む [2], [3]．決定木の概要は第 3.3 節で述べる．

3.2 使用データの概要

本稿で使用する分子量，融点，沸点，発火点データは，「国際化学物質安全性カード」ウェブページ [5] から得た．このウェブページでは官能基がわかる構造式も掲載されており，本稿では構造が簡単な分子を用いるため，文献 [4] も併用して構造式から炭素原子個数，酸素原子個数，官能基

表 1 使用データ例

物質名	メタン CH_4	ベンゼン C_6H_6
オブザベーション番号	1	33
分子量	16.0	78.11
沸点	-161.49	80.1
融点	-182.48	5.5
C 個数	1	6
O 個数	0	0
ベンゼン環	0	1
炭素間二重結合	0	0
ヒドロキシル基	0	0
カルボキシル基	0	0
ケトン基	0	0
エステル結合	0	0
エーテル結合	0	0
発火点	537.0	498.0

の個数，各種結合の個数が得られる．本稿ではウェブページ [5] に掲載されている物質のうち，発火点データが明示されたもので，第 1 節で示した条件を満たした物質を用いる．

決定木関連分野では，記述子は予測変数とも呼ばれ，予測対象は基準変数とも呼ばれる．本稿の予測対象は発火点である．今回は「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．表 1 に決定木の作成で使用した化学物質データの一例を示す．表 1 に示すように，本研究で入力データに用いる記述子は，以下のものである．

分子の性質を表す連続値： 分子量，沸点，融点，発火点．

分子の特徴的な原子個数（0 または自然数）： 炭素原子個数，酸素原子個数．炭素原子は必ず存在するが，酸素は存在しない場合があり，そのときは 0 とする．

特徴的な構造の個数（0 または自然数）： ベンゼン環，二重結合，ヒドロキシル基，カルボキシル基，ケトン基，エステル結合，エーテル結合．なお，存在しない場合は 0 とする．

以上のデータを各分子は欠損値なしで持つ．以後，1 個の分子を 1 個の「オブザベーション」と呼ぶこととする．本稿では，1 オブザベーションにつき連続値と離散値を含む 13 個のデータがある．

それぞれの記述子を使用する理由を述べる．分子量は分子の大きさを表す量としてよく用いられるものの一つである．炭素原子の最も基本的な同位体である 6 個の陽子と 6 個の中性子でできた ^{12}C 原子の質量を正確に 12 としたときの，各原子の平均相対質量を原子量というが，分子量は，分子を構成する原子の原子量の総和である．分子の大きさは分子の基本的な性質の一つであり，気体へのなりやすさに大きく影響することが知られているため，入力データに使用する．

第 3.1 節で述べたように，沸点および炭素原子個数，酸素原子個数は燃焼現象に関係が深い量であるため，記述子に含める．沸点に関連する量として，融点も入力データに含めた．水素原子は分子量に占める割合が非常に小さく，有機化学ではあまり考慮しないことが多い．水素原子は本稿が対象とするすべての分子に入っていて，分子の特徴には寄与しないと考えられるため，本稿においても特別な扱いは行わない．

次に，記述子に含めた結合と官能基を紹介する．炭素原子相互の結合には単結合，二重結合，三重結合およびベンゼン環がある．燃焼では炭素原子の結合が切れて CO_2 が発生するため，炭素原子の結合エネルギーが燃焼に深く関係する．そのため，炭素原子同士および炭素原子と酸素原子との間の主要な結合を記述子に含めた．ベンゼン環は通常の単結合や二重結合とは異なり，6 個の炭素原子同士の結合の全てが等価な結合形態をとり，結合が非常に安定であって，一般に発火点が高くなることが知られている．そのため，ベンゼン環の有無を本稿では記述子に含めた．ヒドロキシル基などの官能基は，分子の特徴ごとに発火点に違いがある可能性があるため使用した．本稿では一般的な官能基を用いる．

本稿で扱う化学物質は，おおまかに以下のような分類が可能である．最後に示した件数は，決定木作成に使用したオブザベーション数である．

鎖状飽和化合物 直鎖型で単結合のみで結合した化合物．メタン，エタンなど．129 件．

鎖状不飽和化合物 直鎖型で二重結合や三重結合を含む化合物．エチレン，プロペンなど．38 件．

環状飽和化合物 環状で単結合のみで結合した化合物．シクロプロパンなど．10 件．

環状不飽和化合物 環状で二重結合を含む化合物．シクロヘキセン，無水マレイン酸など．10 件．

芳香族化合物 ベンゼン環を持つ化合物．ベンゼン，フェノールなど．41 件．

複素環式化合物 炭素と酸素（エーテル結合）から成る環状を持つ化合物．フランなど．17 件．

これらの各分類における分子量と発火点の関係のうち，代表例を図 1 および図 2 に示す．これらの図は相互に比較できるように縦軸と横軸の範囲を揃えており，酸素原子個数で色分けしている．図 1 は鎖状飽和化合物，図 2 は芳香族化合物の分子量と発火点の関係である．図 1 において酸素 0 個の物質は分子量が増えると発火点が小さくなる傾向はあるが，他の物質では分子量と発火点の間に明確な関係は見られない．図 2 およびここで示していない分類においても，分子量と発火点の間に明確な関係は見られない．一方，図同士を比較すると，図 2 の芳香族化合物は他の化合物に比べて発火点が大きい傾向があることがわかる．人間が視覚的に理解できるルールを，決定木でも得られるかどうかをあわせて検討する．

3.3 決定木の概要

本稿では，データマイニングの代表的な手法の一つである決定木を用いる．決定木とは CART という機械学習のアルゴリズムが分析した統計的な相関関係の結果を木構造で表した図である [1]．予測や分類を行いたい量である基準変数が質的変数の場合は分類木の手法を用い，基準変数が連続変数の場合は回帰木の手法を用いる．本稿では連続変数である発火点を予測するため，回帰木を扱う．

次に，回帰木における一般的な分岐基準の選択方法の概要を説明する．回帰木では，平方和の分解を用いて分岐する．親ノード（回帰木作成の最初では根ノードであり，全オブザベーションを含む）において，基準変数の偏差平方和 S は次式で計算する．

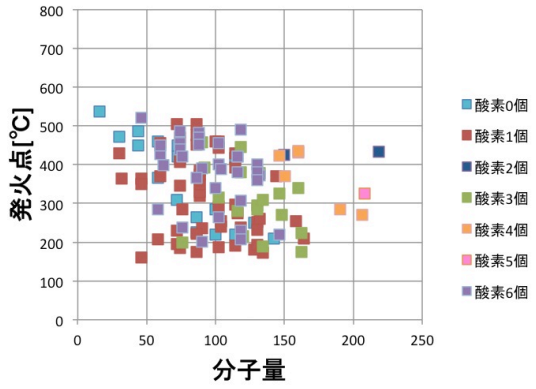

図 1 鎖状飽和化合物の分子量と発火点の関係

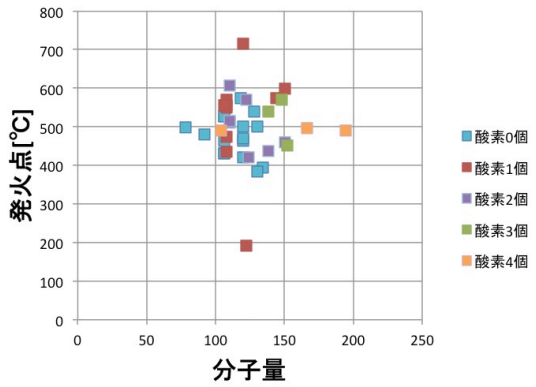

図 2 芳香族化合物の分子量と発火点の関係

$$S = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{1}$$

ここで，N は親ノードが持つオブザベーション個数，y は i 番目のオブザベーションが持つ基準変数値，$\bar{y}$ は親ノード内の基準変数の平均値である．次に，ある予測変数 T の特定の値を用いて 2 分岐する場合を考え，右子が N_R 個，左子が N_L 個のオブザベーションを含むものとする．このとき，右子の偏差平方和 S_R^T および左子の偏差平方和 S_L^T は次式で表せる．

$$S_R^T = \sum_{i=1}^{N_R} (y_i - \bar{y_R})^2 \tag{2}$$

$$S_L^T = \sum_{i=1}^{N_L} (y_i - \bar{y_L})^2 \tag{3}$$

ここで，$\bar{y_R}$ と $\bar{y_L}$ は右子と左子それぞれのノード内での基準変数の平均値である．予測変数の値候補について，式 (1)-(式 (2)+式 (3)) に相当する

$$S - (S_R^T + S_L^T) \tag{4}$$

を計算し，式 (4) が最大となる予測変数を採用して分岐する．式 (1) すなわち S は，この分岐を検討する段階では定数であるため，$S_R^T + S_L^T$ が最小となる予測変数 T の値が式 (4) を最大にする．

3.4 決定木の作成

最初に，今回集めた 245 件のオブザベーションを関数 rpart の入力に用い，決定木を作成した．作成した決定木を図 3 に示す．図 3 中の楕円は木のノードを表し，ノードに記入した数値は，そのノードに所属するオブザベーションの平均発火点を示す．丸の下の「n=」とついた数値は，そのノードに所属するオブザベーション個数である．各ノードの上に記入している式は，そこから分岐する際の分岐条件であり，分岐条件を満たすオブザベーションは次に左子へ，満たさないときは右子へ所属する．図 3 は分岐が非常に多く，分岐を適切に打ち切る「プルーニング」が必要な可能性がある．分岐が多い状態は過学習が起こっている状態でもある．そこで，文献 [3] の手順に従い，プルーニングを行うかどうか検討する．

文献 [3] によると，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確認も行い，デフォルトでは 10 分割交差確認を行う．関数 plotcp はプルーニングに必要な情報を図示するので，図 3 の決定木における plotcp の出力を図 4 に示す．図 4 は下側の第一横軸が木の大きさであり，縦軸は相対誤差である．なお，文献 [3] によると，plotcp の出力において，木の大きさは葉の数で示している．cp は木が成長すると単調に減少するため，上側の第 2 横軸で示しているが，cp は木の大きさに対して等間隔ではない．図 4 では相対誤差＝ 0.83 程度が妥当な基準として自動計算されて，図 4 中の水平線として表示されている．その根拠は（誤差の最小値）＋（誤差の標準偏差）がその数値となることによる．

図 4 によると，木のサイズ=2 で相対誤差は一旦基準値よりも小さくなり，その後基準値を超えることはないが，基準値付近に留まった状態である．デフォルトでは cp=0.01 で終了するため，図 4 の右端は cp=0.01 である．図 4 において基準値を厳格に適用すると，根のみの分岐で木の成長を止めることとなるため，図 3 は過学習の傾向はあるがプルーニングを行わないこととする．

図 3 で得られたルールを以下に述べる．この木は過学習の傾向があるので，木の深さ 2 までの，上方の 4 つのルールのみを列挙する．

ルール 1: ベンゼン環を持たない分子の平均発火点は 344 宝，ベンゼン環を持つ分子の平均発火点は 492 宝 である．

ルール 2: （ベンゼン環を持たない分子のうち）分子量が 67 以上の分子の平均発火点は 331 宝，分子量が 67 未満の分子の平均発火点は 411 宝である．

ルール 3: （ベンゼン環を持たず，分子量が 67 以上の分子のうち）カルボキシル基を持たない分子の平均発火

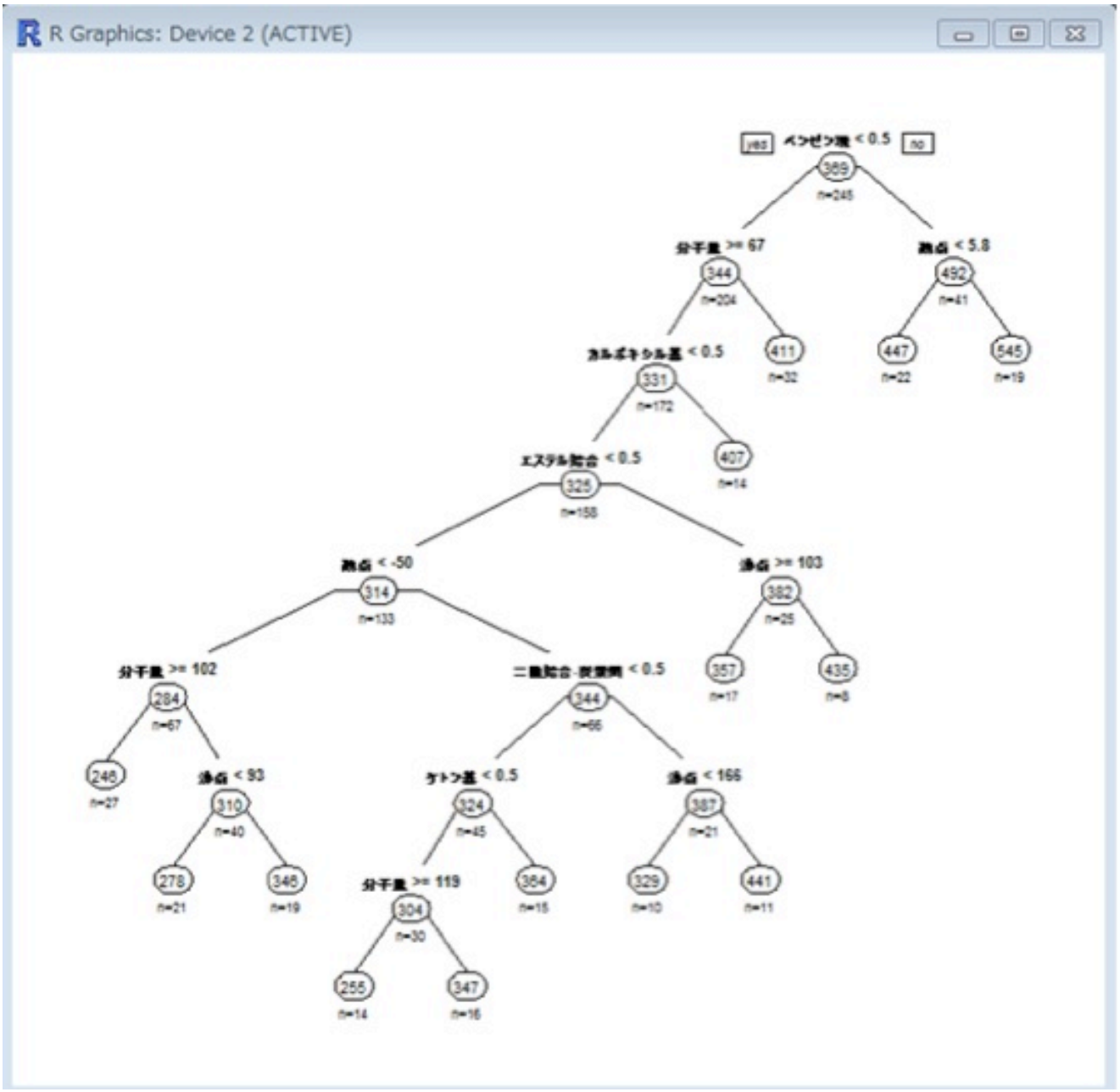

図 3 245 種類の分子における発火点の決定木

点は 325 宝，カルボキシル基を持つ分子の平均発火点は 407 宝である．

ルール 4: （ベンゼン環を持つ分子のうち）沸点が 5.8 宝未満の分子の平均発火点は 447 宝，沸点が 5.8 宝以上の分子の平均発火点は 545 宝である．

ルール 1 は，第 3.2 節で示した，ベンゼン環を持つ分子は他の種類の分子よりも発火点が高くなる傾向と適合する．ルール 2，3，4 は第 3.2 節の図 1 および同様の図からは読み取れないルールである．

4. 決定木を用いた発火点予測

次に，図 3 の決定木が，実際の化学物質の発火点予測に

表 2 図 3 の決定木を用いた発火点予測結果

オブザベーション番号（分類）		実際の発火点	決定木による予測発火点	差分	評価
246	鎖状不飽和	202	278	76	△
247	鎖状飽和	238	407	169	×
248	鎖状飽和	266	255	-11	○
249	鎖状飽和	360	407	47	○
250	鎖状不飽和	378	411	33	○
251	鎖状飽和	423	346	-77	△
252	鎖状飽和	430	246	-184	×
253	芳香族	518	545	27	○
254	芳香族	543	545	2	○
255	芳香族	550	447	-103	×

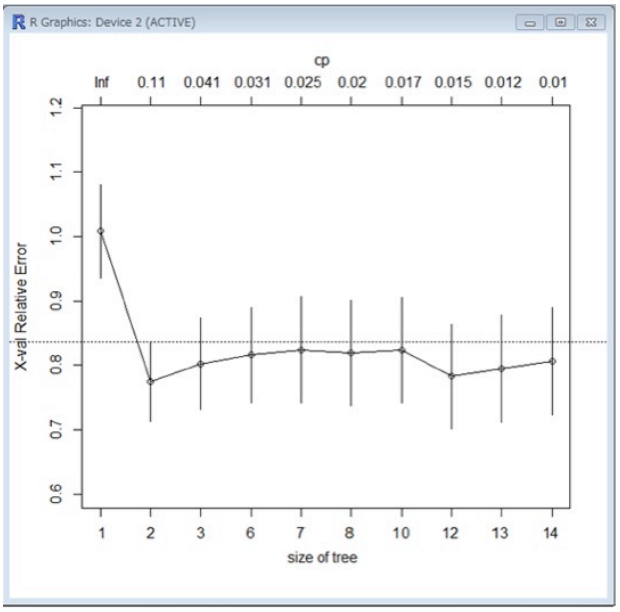

図 4 木の大きさおよび複雑度と相対誤差の関係

有効かどうかを検討する．決定木の生成時に使用していないオブザベーション 10 個を別に用意し，predict.rpart 関数を用いて図 3 の決定木による発火点予測を行った結果を表 2 に示す．表 2 では左から，使用オブザベーション番号とそのオブザベーションが所属する第 3.2 節で示した分子の分類，実際の発火点，決定木を用いて予測した発火点および（差分）＝（予測発火点）-（実際の発火点），評価を示した．評価は，差分の絶対値が 50 以下ならば○，50 〜 100 ならば△，100 以上ならば×とした．表 2 では，○が 5 個，△が 2 個，×が 3 個であった．すなわち 10 件のオブザベーション中の 5 件は誤差 50 宝 以内で発火点を予測できた．しかし，誤差 100 宝以上のオブザベーションも 3 件ある．

実験においては，50 宝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 宝異なると，実験環境の耐熱性能等の事前の想定に影響するものと考えられる．本稿で得た決定木は，いずれも 50〜100 宝の誤差が 10 件中 2 件，100 宝以上の誤差も 10 件中 3 件現れており，まだ実用に耐えるものではない．

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，その決定木を用いて発火点予測を行った．今回作成した木は最初の分岐規則以外は過学習の傾向があり，改善の余地がある．学習データと異なるデータを決定木に適用して発火点を予測した結果，10 件中 5 件は誤差が 50 宝未満であったが，誤差が 100 宝度を超える結果も 3 件あり，実用に供するためにはまだ課題が多い．

今後の課題は以下の通りである．今回作成した決定木は過学習の傾向があったため，オブザベーションを増やして決定木を作成し，決定木の妥当性を検討する必要がある．また，発火点はもともと実験環境への依存性が大きい量であるため，より適切に発火点を予測するには実験条件を考慮する必要がある．実験条件は温度などの数値データだけでなく使用した容器の形状などの数値では表しきれない情報や非言語情報が含まれる．今回使用したデータは実験条件が記載されていないため，実験条件の取得方法とその利用方法を検討する必要がある．

決定木はニューラルネット等よりも，分類で用いた規則が人間に理解しやすいが，得られたルールは最適である保証がないことが指摘されている [1], [3]．このような手法をどのように化学などの分野で利用していくのか，利用方法なども今後検討していきたい．

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] 岡田昌史，他：R パッケージガイドブック，東京図書株式会社 (2011).
[3] 金　明哲：[連載] フリーソフトによるデータ解析・マイニング　第１９回 R と樹木モデル（２），入手先 ⟨https://www1.doshisha.ac.jp/ mjin/R/19.html⟩ (2016.2.8).
[4] 吉村壽次代表編集：化学辞典 (第 2 版) 小型版，森北出版株式会社 (2009).
[5] 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/⟩ (2015.12.20).
[6] J. G. Quintiere，大宮喜文，若月薫訳：基礎　火災現象原論，共立出版 (2009).
[7] 久保田浪之介：トコトンやさしい燃焼学の本，日刊工業新聞社 (2012).
[8] J.Gasteriger，T.Engel 編集，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., Liaw, H.-J.: *A model for predicting the auto-ignition temperature using quantitative structure property relationship approach*, Procedia Engineering 45, 512-517, (2012).
[10] 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).