Vol.2022-MPS-138 No.13
Vol.2022-BIO-70 No.13
2022/6/27

基礎的有機化合物の融点と沸点の決定ルール

林 亮子[1,a)]

概要：融点と沸点は物質の性質を示す基礎的な量であり，古くから多くの物質において調べられているため，多くのデータが蓄積されている．そして，融点と沸点には分子の大きさや構造が影響することが知られているが，定量的にどの程度の影響があるのかはまだ調査の余地がある．本稿では，炭化水素および類例の分子の融点と沸点を決定木とランダムフォレストを用いて分類し，分類ルールを調査した結果を報告する．

キーワード：沸点，融点，決定木，ランダムフォレスト，有機化合物

1. はじめに

近年，機械学習の利用技術 [1][2] は非常な発展を遂げ，誰でも簡単に機械学習を利用できるようになった．また，蓄積したデータに機械学習を適用して隠れたルールを調査しようとするデータ科学は，理論，実験，シミュレーションに次ぐ「新たな研究手法」と言われるほど注目されている．

物質が固体から液体に変わり始める融点と，液体から気体に変わり始める沸点は，物質の性質を示す基礎的な量である．そのため，古くから多くの物質において融点と沸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 [3][4][5][6] 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．融点と沸点は分子の性質を反映していて，分子の大きさや構造が融点と沸点に影響することが知られている．

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 [7] 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，融点と沸点の予測は古くから行われている [8]．融点と沸点のメカニズムはある程度知られているが，定量的にはまだ調査の余地があるものと考えられる．

本稿の著者グループは，これまでに文献 [9] において発火点を調べてきた．その過程で，ケモインフォマティクス分野の研究者から融点と沸点についても調べるように助言を戴いた．そこで，まず発火点調査で使用したデータを使用して主要なデータマイニング手法である決定木とランダムフォレストで融点と沸点の分類と予測を行い，これまで知られた融点と沸点の依存性がどの程度実際に現れるのかを調べた結果を本稿で報告する．

第 2 節以降の本論文の構成を述べる．第 2 節は今回使用したデータの概況を説明し，説明変数の設定内容を示す．第 3 節では決定木およびランダムフォレストを用いて分類器を作成した結果を示し，テストデータから分類器で融点と沸点を予測した結果を紹介する．第 4 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. 使用データの概要

2.1 分子データの収集

本小節では，使用したデータの収集方法を述べる．本稿では筆者が文献 [9] で用いたデータから発火点を除外し，融点と沸点を目的変数に設定して使用する．使用したデータは，国際化学物質安全性計画 (IPCS) が作成している「国際化学物質安全性カード」[3] を基本とし，欠損値を補うために wikipedia,「職場のあんぜんサイト」[4] および「Chemical Book」[6] を補助的に利用した．また，本稿では説明変数として，特徴的な部分構造がその分子に含まれる個数を使用するが，それらのデータは独自に作成した．作成の詳細は文献 [9] に譲るが，データの内容を 2.3 で説明する．

2.2 使用データにおける基礎的な諸量の分布

文献 [9] では 294 件のデータを使用した．まず，それらのデータの分子量の度数分布を図 1 に示す．分子量とは，

[1] 金沢工業大学
Kanazawa Institute of Technology
[a)] ryoko@neptune.kanazawa-it.ac.jp

Vol.2022-MPS-138 No.13
Vol.2022-BIO-70 No.13
2022/6/27

分子の大きさを示す量で，分子を構成する元素の原子量の総和であり，分子の相対的質量を表す．図 1 によると，分子量 50 から 150 までの分子が最も多く，その階級から離れるにつれて，データ件数は減少する．

次に，融点の度数分布を図 2 に示す．図 2 では，融点が −100 ∼ 0°C が最も多く，50°C を超える物質は少ない．このことから，今回のデータに含まれる物質は，常温で固体のものが少ないことがわかる．図 3 に 294 件のデータにおける沸点の度数分布を示すが，沸点が 0°C 以下の物質は少なく，ほとんどが 50°C 以上である．このことから，今回扱う物質のほとんどは常温では液体であることがわかる．

さらに，分子量と融点の関係を確認する．図 4 は 294 件のデータにおける分子量と融点の関係を示す散布図である．図 4 では，おおむね分子量が増加すると融点も高くなる傾向がわかるが，同じ分子量でも融点には幅があり，例えば分子量 100 程度の物質の融点には，最大で 300°C ∼ 400°C 程度の幅があることがわかる．なお，図 4 の全てのデータで計算した相関係数は 0.29 であったため，弱い正の相関がある．しかし，分子量 250 程度以上の分子では，それ以下のサイズの分子と比較すると，分子量に対して融点が低いことから外れ値として扱うことも考えられ，その場合は相関係数はより大きくなると予想できる．

図 5 は今回使用したデータの分子量と沸点の関係を示す散布図である．図 5 では，分子量が 300 を超える分子では沸点が大きくならない傾向があるが，分子量が 300 程度までは分子量と沸点の間に大きな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 であるため，正の相関があると考えることができる．図 4 と比べると沸点では，同じ分子量の分子間での沸点の差異は小さいので，沸点の分子量への依存性が融点よりも大きいと予想できる．

今回用意したデータに分子量 350 程度以上の分子は 5 件ある．これらのデータは融点と沸点の両方で他のデータと傾向が異なるように見えるので，以降のデータ処理では外れ値として除外する．すると 289 件のデータを扱うこととなるので，学習に使用しないテストデータを 29 件ランダムに選択し，残り 260 件を学習データに使用した．

2.3 分子の性質を示す説明変数

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 6 に示す．図 6(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，環構造のもう一つの炭素原子に 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 6(b) および図 6(c) に示す．図 6(b) および図 6(c) には以下の部分構造がある．

カルボニル基　図 6(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

炭素間二重結合　炭素原子に CH_2 が二重結合している部分

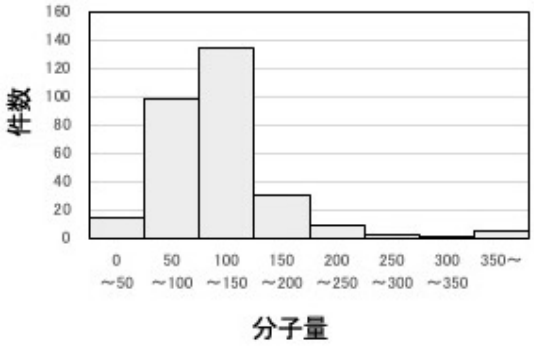

図 1　分子量の度数分布（294 件）

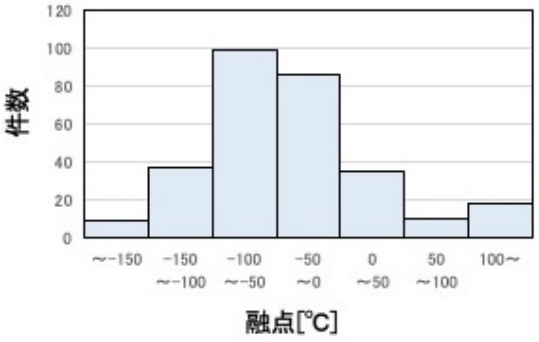

図 2　融点の度数分布（294 件）

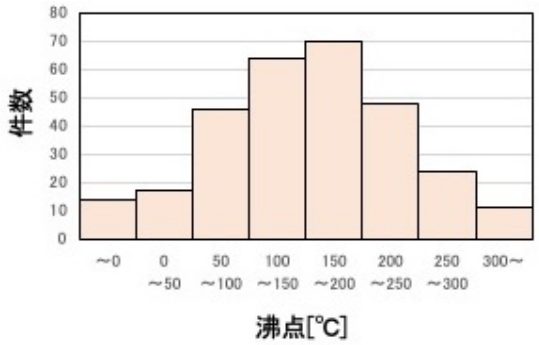

図 3　沸点の度数分布（294 件）

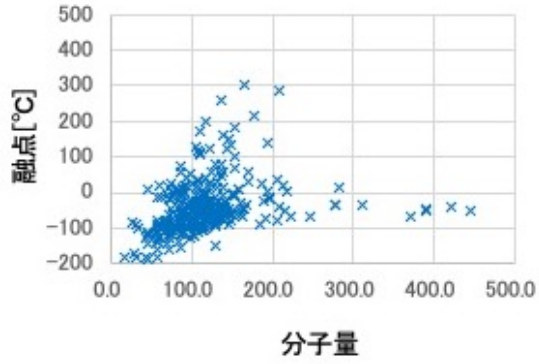

図 4　使用データにおける分子量と融点の関係（294 件，相関係数 =0.29）

Vol.2022-MPS-138 No.13
Vol.2022-BIO-70 No.13
2022/6/27

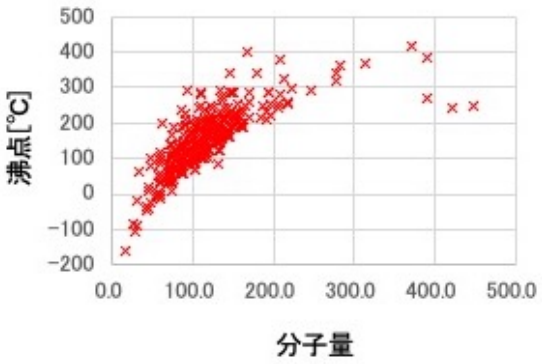

図 5　使用データにおける分子量と沸点の関係（294 件，相関係数 =0.75）

エーテル結合　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の部分

環状エーテル　エーテル結合からさらに原子が結合した先で，もとのエーテル結合のもう一方の原子に結合しており，これらをまとめた部分

エステル結合　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っており，これらをまとめた部分

環状エステル　エステル結合から原子が結合した先がもとのエステル結合に戻るため，これらをまとめた部分

これらの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，説明変数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．入れ子の関係を持つ部分構造では，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，本稿では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．

本稿で使用するデータの例として，図 6 のジケテンのデータを表 1 に示す．表 1 に示すように，本稿では目的変数を沸点と融点とし，分子の性質を表す連続値の説明変数として分子量を用いる．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，説明変数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．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，カルボキシル基，エステル結合，環状エステル，環状構造，枝分かれがない構造である直鎖構造を説明変数に使用する．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 6 に示したように環状エステル構造を 1 個持つ．今回のデータでは，包含関係にある説明変数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，環構造の全てが 1 である．

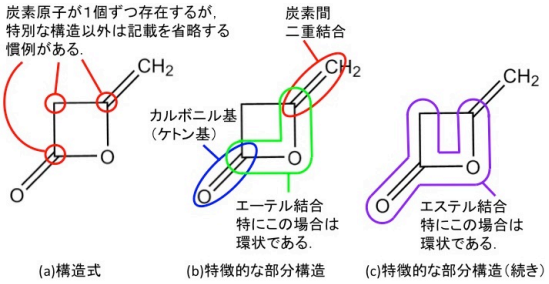

図 6　構造式と特徴的な部分構造の例（ジケテン $C_4H_4O_2$，SMILES 記法では C1(=O)OC(=C)C1）

3. 融点と沸点のデータマイニング

3.1 データの分類手法

本稿では統計プログラミング言語 R を用いてデータ処理を行なった．R のバージョンは 4.2.0 である．本稿ではデータの分類ルールの可読性の観点から決定木，過学習への頑健さからランダムフォレストを用いてデータ処理を行う．決定木は R の rpart パッケージ（バージョン 4.1.16），ランダムフォレストは randomForest パッケージ（バージョン 4.7-1）を使用した．決定木は機械学習を用いて人間が理解しやすい分岐ルールを設定してデータ集合を再帰的に 2 分割するが，過学習が起こりやすいことが知られている．一方ランダムフォレストは，決定木を複数個作成してアンサンブル学習を行う手法で，決定木よりも過学習に強いことが知られている．

3.2 決定木による学習結果

図 7 は 260 件の学習データで作成し，枝刈りを行った融点の決定木である．図 7 は，R の rpart.plot 関数が出力す

表 1　使用データ例（環状エステル，ジケテン $C_4H_4O_2$）

目的変数		データ例	データ範囲
沸点 [°C]		127	127
融点 [°C]		-7	-7
説明変数		データ例	データ範囲
分子量		84.1	84.1
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	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	1
環構造		1	1
直鎖構造		0	0

Vol.2022-MPS-138 No.13
Vol.2022-BIO-70 No.13
2022/6/27

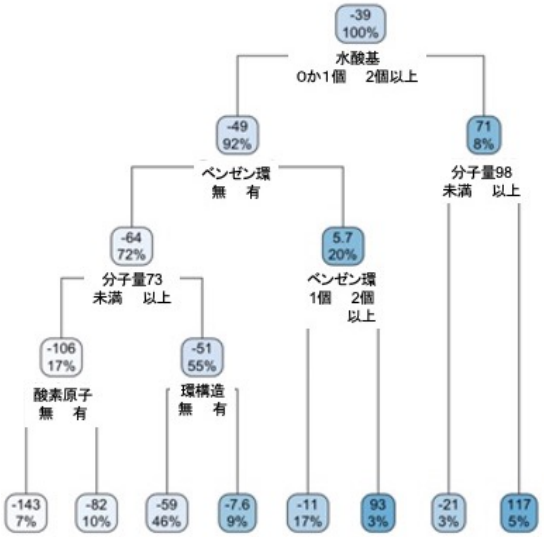

図 7　枝刈り後の融点の決定木（学習データ 260 件）

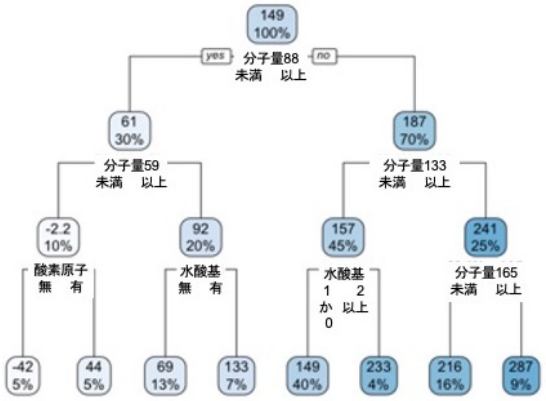

図 8　枝刈り後の沸点の決定木（学習データ 260 件）

る図上に分岐ルールの意味を重ね書きして作成している．一番上が全ての学習データ集合を表す木のノードで，全体の平均融点が-39°C であることを表す．ノードのすぐ下には最初の分岐ルールが記載されており，水酸基が 2 個未満の場合は左のノード（平均融点 −49°C で，92 件のデータが含まれる．），2 個以上の場合は右のノード（平均融点 71°C で，8 件のデータが含まれる．）に分岐することを示す．決定木では，上のほうが重要なルールである．図 7 では，水酸基やベンゼン環の個数が重要なルールとなっている．また，これまで知られていたように，分子量の大きさも重要なルールとなっている．

図 8 は 260 件の学習データで作成し，枝刈りを行った沸点の決定木である．図 8 は図 7 と同様に，R の rpart.plot 関数が出力する図に分岐ルールの意味を重ね書きして作成している．見方は図 7 と同様である．図 8 では，上の方の分岐ルール 3 件が分子量に関するものであり，沸点では分子の大きさが支配的であることになっている．分岐が進むと水酸基の個数や酸素原子の有無に関するルールが現れるが，融点とは異なり，ベンゼン環に関するルールは現れない．

図 7 と 8 を比べると，融点では分子構造に関するルールが上位にあり，沸点では分子量に関するルールが上位にある．融点は，分子が固体からどれだけ液体になりやすいかを示すものと考えられ，固体では液体や気体に比べて一般に分子間距離が小さいので，分子の構造の影響が大きい可能性がある．一方沸点は分子が液体からどれだけ気体になりやすいかを示すものと考えられ，一般に液体では分子間距離がある程度大きいので，分子内の詳細な構造というよりは分子の大きさすなわち分子量の影響が大きいことを示す可能性がある．

3.3　ランダムフォレストによる学習結果

次に，260 件の学習データにランダムフォレストを適用して分類器を作成し，その重要度を調べた結果を示す．図 9 は融点の分類器における説明変数の重要度である．図 9 によると，最も重要な説明変数は分子量で，その次が水酸基の個数，さらにベンゼン環個数である．次いで炭素原子個数と酸素原子個数があるが，炭素原子個数は分子量に寄与しており，酸素原子個数は水酸基個数に寄与しているので，連動して重要度が高くなっているものと考えられる．さらに，カルボキシル基，カルボニル基，エーテル結合，環構造もある程度重要であるものと考えられる．

図 10 はランダムフォレストを用いて作成した沸点の分類器における説明変数の重要度である．図 10 によると，最も重要な説明変数は分子量で，その次は炭素原子個数であるが，炭素原子個数は分子量に寄与しているため，融点での状況と同様に，分子量に連動して重要度が高くなっているものと考えられる．さらに，水酸基個数と酸素原子個数の重要度が続くが，融点と同様に，本質は水酸基と考えられる．ベンゼン環個数もある程度の重要度があるものと考えられる．以降の説明変数は重要度が非常に小さいので，今回使用したデータにおいては実質の重要度はないものと考えられる．

図 9 と図 10 を比べると，どちらも分子量が最も重要度が大きい．これは決定木の結果とは様子が異なるが，これまで化学分野で知られた知見では，「分子量が大きいほど沸点と融点は高くなる」ことが分子間力の議論から知られており，その知見には合う結果である．また，図 9 と図 10 ではどちらも実質的に分子量の次に重要度が高い説明変数は水酸基個数であり，これが妥当なルールかどうかは更なる調査が必要である．また，図 9 と図 10 ではどちらも分子量の重要度が最も大きいが，図 9 では他の量の重要度も大きいのに対して図 10 では分子量以外の量の重要度が分子量に対して小さいので，沸点では融点よりも分子量への依存性が大きいものと考えられる．

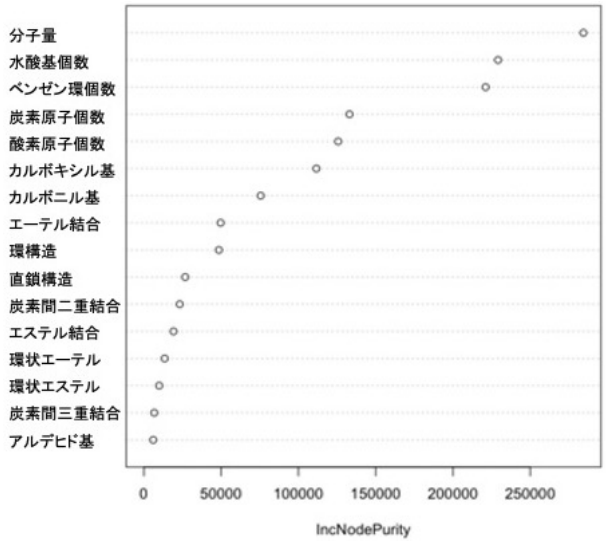

図 9 ランダムフォレストを用いた融点の分類器における説明変数の重要度（学習データ 260 件, mtry=5, 生成した木の数は 500）

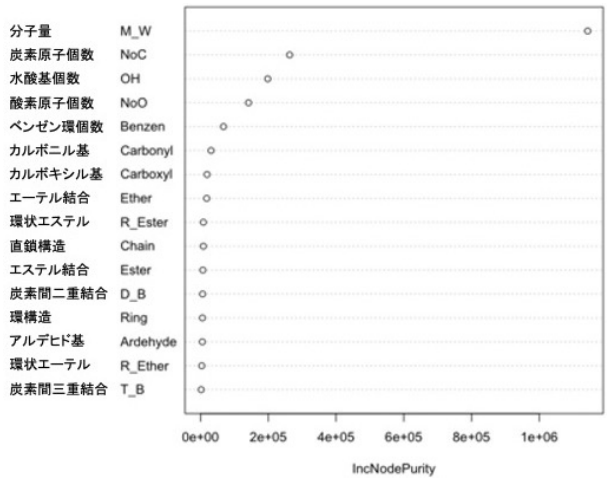

図 10 沸点を分類するランダムフォレストにおける説明変数の重要度（学習データ 260 件, mtry=10, 生成した木の数は 500）

3.4 沸点と融点の予測性能

決定木とランダムフォレストによる分類器の性能評価のため，学習データに含まないテストデータ 29 件の，実際の融点と予測した融点の関係を図 11 に示す．図 11 では，横軸が実際の融点，縦軸が予測融点を表し，1 件のデータが図 11 中の 1 つの点となる．図 11 では，正確に融点を予測できると縦軸の量と横軸の量が一致するので，縦軸の量＝横軸の量となる理想曲線をあわせて記載した．図 11 では，ランダムフォレストの予測結果が決定木の予測結果よりもおおむね理想曲線に近い領域にあり，ランダムフォレストのほうが良好な予測ができていることがわかる．なお，二乗平均平方根誤差 (RMSE) を計算したところ，決定木は 62.4 であり，ランダムフォレストはより小さい 42.9 であったため，平均的にランダムフォレストのほうが予測性能が良いことがわかる．

次に，同じテストデータ 29 件の，実際の沸点と予測した

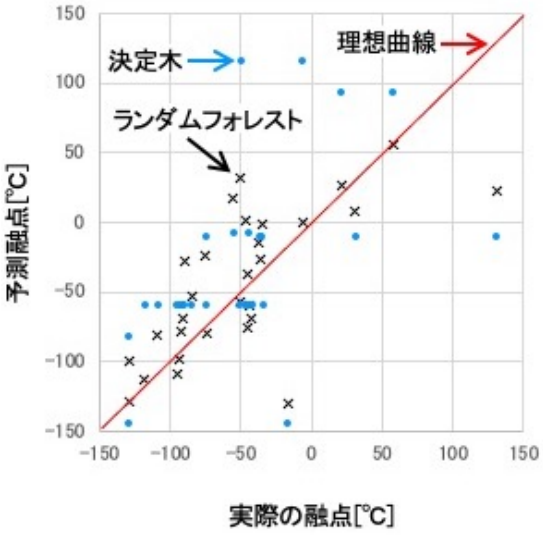

図 11 テストデータを用いた融点予測結果（決定木の RMSE=62.4, ランダムフォレストの RMSE=42.9）

沸点の関係を図 12 に示す．図 12 も図 11 と同様に，横軸が実際の融点，縦軸が予測融点を表し，1 件のデータが図 12 中の 1 つの点となる．こちらも縦軸の量＝横軸の量となる理想曲線をあわせて記載した．図 12 では，ランダムフォレストの予測結果が決定木の予測結果よりもおおむね理想曲線に近い領域にあり，ランダムフォレストのほうが良好な予測ができている．なお，二乗平均平方根誤差 (以後 RMSE と呼ぶものとする) を計算したところ，決定木は 41.8 であり，ランダムフォレストはより小さい 23.9 であったため，平均的にランダムフォレストのほうが予測性能が良いことがわかる．

図 11 と図 12 を比較する．図 11 の温度範囲は 300°C で，図 12 では 350°C なので，2 つの図の温度範囲はおおむね同程度である．図 11 は決定木もランダムフォレストも理想曲線から大きく外れた点があり，図 12 では理想曲線から大きく外れた点が少ない．また，沸点の RMSE のほうが融点よりも小さい．そのため，沸点のほうが正確に予測しやすいことが示唆される．

4. おわりに

本稿では，決定木とランダムフォレストを用いて炭化水素および類例分子の融点と沸点の分類器を作成した．その結果，これまで化学分野で知られていたように，分子量が融点および沸点に大きく影響することが今回使用したデータの範囲でも確認できた．また，融点の予測は沸点よりも難しく，ランダムフォレストでも予測融点が実際の融点から 100°C 以上異なる場合があった．

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物質であり，実際には可燃物の範囲でも，窒素や硫黄を含む有機化合物が多数存在する．扱う元素の種類を増

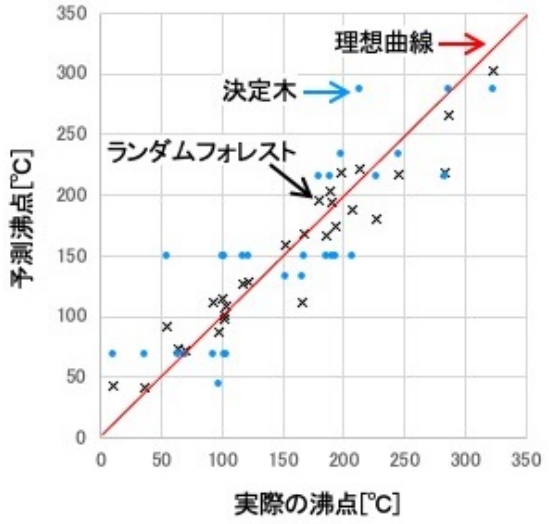

図 12 テストデータを用いた沸点予測結果（決定木の RMSE=41.8, ランダムフォレストの RMSE=23.9）

やすと分子の特徴的な部分構造の種類も増えるため説明変数を増やす必要がある．データを追加して引き続き調査を試みたい．

謝辞 本研究の一部は科学研究費補助金 19K12007 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, http://www.nihs.go.jp/ICSC/ (2016.3.17).
[4] 厚生労働省：職場のあんぜんサイト, http://anzeninfo.mhlw.go.jp/ (2016.3.17).
[5] 科学技術振興機構（JST）：科学技術総合リンクセンター J-GLOBAL, http://jglobal.jst.go.jp/ (2016.3.17).
[6] Chemical Book, http://www.chemicalbook.com/ (2016.3.17).
[7] 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).
[8] N. Shimizu, H. Kaneko, ”Construction Regression Models with High Prediction Accuracy and Interpretability Based on Decision Tree and Random Forests”, J. Comput.Chem. Jpn., Vol. 20, No. 2, pp.71-87 (2021).
[9] 林 亮子：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，情報処理学会論文誌数理モデル化と応用 (TOM)，Vol. 10, No. 3, pp.20-31, (2017).