様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 ２５ 日現在

機関番号：３３３０２
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２００７

研究課題名（和文）データ科学と計算科学の協働に基づく物質探索システム

研究課題名（英文）Material search system based on collaboration between data science and computational science

研究代表者
林　亮子（Hayashi, Ryoko）

金沢工業大学・工学部・准教授

研究者番号：３０３０３３３２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：近年物質・材料科学において，これまで蓄積したデータとデータ科学および計算科学を活用して新規の材料を研究開発する機運が国内外で高まっている．本研究はデータ科学と計算科学を適切に組み合わせて物質探索を目指すものである．
本研究では，試験的に沸点と融点の分類ルールを調査した．融点と沸点は物質の性質を示す基礎的な量であり，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．そこで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た．

研究成果の学術的意義や社会的意義
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．融点と沸点は分子の性質を反映しており，分子の大きさや構造が融点と沸点に影響することが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．そこで本研究では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性質を定量的に評価できるかどうかを調べた．

研究成果の概要（英文）：In recent years, in material science, the momentum for researching and developing new materials by utilizing the data accumulated so far and data science and computational science is increasing at home and abroad. This research aims at material search by appropriately combining data science and computational science.
In this study, we investigated the classification rules of boiling point and melting point on a trial basis. Melting point and boiling point are basic quantities that indicate the properties of a substance, and many data have been accumulated. It is known that the size and structure of molecules affect the melting point and boiling point, but there is still room for investigation as to how much they affect quantitatively. Therefore, the melting points and boiling points of hydrocarbons and similar molecules were classified using decision trees and random forests.

研究分野：データ科学

キーワード：データマイニング　ケモインフォマティクス　分子　分子間力

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年物質・材料科学において，これまで蓄積したデータとデータ科学および計算科学を活用して新規の材料を研究開発する機運が国内外で高まっている．本研究はデータ科学と計算科学を適切に組み合わせて物質探索を目指すものである．研究代表者はシミュレーションの高速化を専門としており，計算科学分野の研究者とも共同研究の経験があるが，その過程で結果データ処理の重要性に気づき，近年では計算科学とデータ科学を組み合わせた物質探索支援システムに関する研究開発を行っている．

本研究では分子ワイヤの探索を目標とし，必要となる研究開発を行うものである．近年の実験技術や計算技術の発達により，周囲の状態に応じて変形する分子や分子機械などが扱えるようになってきた．それらの物質は共有結合などよりも１桁～３桁小さい力である分子間力で結合しており，今後の研究開発が期待される物質である．本課題では，データ科学と計算科学を活用して，分子が動的に構造を変えて高次の機能を実現する高機能分子材料の設計支援を目的とする．そして，分子間力で結合した導電性分子ワイヤの探索を具体的な実問題として扱う．分子ワイヤは分子レベルで作る結線であり，近年分子機械や分子回路が現実に作成されるようになってきたために注目されている物質である．そのため，本研究が目的とする分子ワイヤ物質の探索は，ナノテクノロジーや材料科学の観点からも意義がある．

２．研究の目的

本課題の概要を述べる．シミュレーションを用いたこれまでの材料設計では，研究プロセスに研究者が介在することが多い．しかし人間の処理能力には限界があるために研究者が研究プロセスのボトルネックになってしまい，またデータを人間が理解するために情報を削減するので計算結果が活用できない．そこで本課題では，研究プロセスに含まれる定型業務を自動化して研究プロセスの支援を行うことを目指す．

結果データの分析から次に探索すべき条件を決定することはまだ人間の判断に委ねられており，マテリアルインフォマティクスによる自動化も試みられているがまだ方法論は確立されていない．また物質・材料科学では，物質を作成してのちその物性を調べる「順問題」的手法による研究開発が主流であり，必要な物性を得るための物質探索は「物質探索における逆問題」に相当し，まだ研究開発の途上である．これらの問題を解決するために，本研究ではデータマイニングを化学や物理などの材料関連分野に応用し，最終的には物質探索を支援することを本研究の目的とする．

３．研究の方法

本研究課題が目的とする導電性分子ワイヤは分子間力によって結合したり結合を解除したりするものを想定しているが，分子間力の影響が大きい現象を示す身近な量に融点と沸点がある．融点と沸点は，物質の基本的な性質を示す量でもあるため，古くから多くの物質で調べられており，データが豊富にある．そこで，融点と沸点でデータマイニングを行い，定性的に知られたルールを定量的に評価できるかどうかを調べてデータマイニングの物性問題への応用可能性を検討する．

今回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．

４．研究成果

本研究の成果の一部を「基礎的有機化合物の融点と沸点の決定ルール」と題して研究会発表するので，その内容を紹介する．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．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．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．

研究代表者は，これまでにデータマイニングを用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知見がどの程度データから再現できるのかを調べた．

今回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．今回扱うデータでは，以上の内容を欠損値なしで持つ．

図１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１は，R の rpart.plot 関数が出力す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根にあたるノードで，全体の平均融点が-39℃であることを表す．各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℃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℃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１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２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２は図１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１と同様である．図２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１と図２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，一般に液体では固体よりも分子間距離がある程度大きいため，液体から気体になるときは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．

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調べた結果，ランダムフォレストのほうが良好な予測ができており，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良い予測性能を得られることがわかる．同様に沸点においても実際の値と予測した沸点の関係を調べた結果，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良好な予測ができ，決定木の RMSE 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．

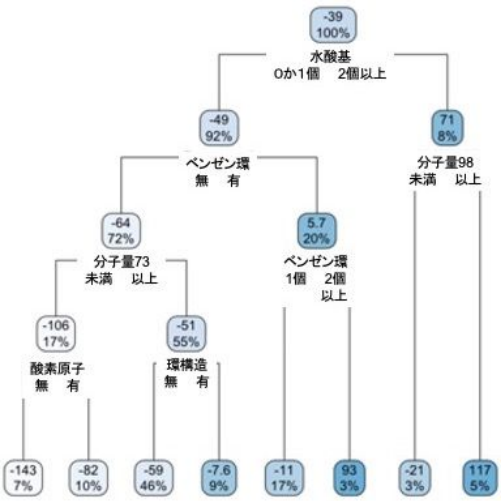

図１．枝刈り後の融点の決定木
（学習データ 260 件）

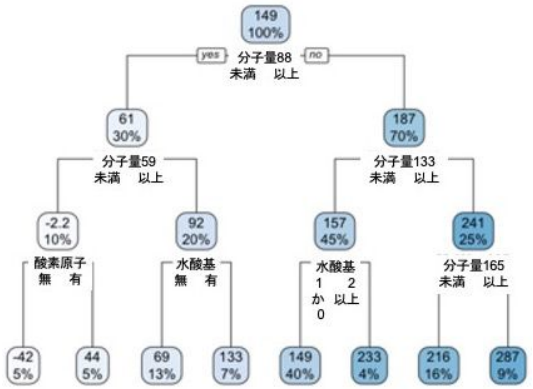

図２．枝刈り後の沸点の決定木
（学習データ 260 件）

＜引用文献＞
「基礎的有機化合物の融点と沸点の決定ルール」，林　亮子，研究報告数理モデル化と問題解決（MPS），2022-MPS-138(13),1-6（2022-06-20），2188-8833．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 林　亮子
２．発表標題 機械学習を用いた分子に関する諸量調査の試み
３．学会等名 第４２回ケモインフォマティクス討論会
４．発表年 2019年

１．発表者名 林　亮子
２．発表標題 基礎的有機化合物の融点と沸点の決定ルール
３．学会等名 情報処理学会第138回MPS研究会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関