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　５ 年　 ６ 月 ２６ 日現在

機関番号：１４３０１
研究種目：基盤研究(A)（一般）
研究期間：2018～2022
課題番号：１８Ｈ０４１１３

研究課題名（和文）離散原像問題の解析と応用

研究課題名（英文）Analysis and Applications of Discrete Preimage Problems

研究代表者
阿久津　達也（Akutsu, Tatauya）

京都大学・化学研究所・教授

研究者番号：９０２６１８５９

交付決定額（研究期間全体）：（直接経費）　　29,200,000円

研究成果の概要（和文）：離散データを入力としてその特性などを出力とする予測関数を機械学習法などで得た後で、望ましい特性を与えてその逆像を計算することにより新規の構造データを生成するという離散原像問題について、化学構造の設計を主対象に研究を行い、混合整数線形計画法に基づく実用的な手法を開発した。予測関数としては主にニューラルネットワークを用い、逆像問題を効率的に解くために化学構造を階層的に表現する二層モデルなどの新たな概念を提案し、中規模の化学構造について現実的な時間で離散逆像問題が解けることを示した。一方、理論的観点からは、線形閾値関数に基づく自己符号化器の圧縮能力と頂点数などの関係を解析するなどの成果を得た。

研究成果の学術的意義や社会的意義
本研究により、新規構造データ生成のための新たな方法論である離散原像問題という枠組みを確立した。原像（逆像）の計算自体は一般に計算困難なクラスに属するが、混合整数線形計画法を効果的に適用するための計算手法や数理モデルを開発し、中規模の化学構造データに対し、実際に原像が計算可能なことを示した。近年、生成AIが注目を集めているが、既存手法とは大きく異なる方法論を示したこともあり、独創的でかつ発展性の高い成果が得られたと考えられる。
この方法論を発展・拡張することにより新規で有用な化合物、さらには、タンパク質などの設計につながる可能性があり、社会的観点からも応用可能性の高い成果が得られたと考えられる。

研究成果の概要（英文）：We studied the discrete preimage problem, in which prediction functions for discrete data are obtained using machine learning methods and then novel discrete data are obtained by computing the preimages for given properties. In this project, we developed methods for the problem based on mixed integer linear programming with focusing on design of chemical structures. As for the prediction functions, we mainly used artificial neural networks, and developed novel representation models such as the two-layered model for efficiently handle chemical structures. As a result, our developed methods could compute preimages for moderate-size chemical structures. From a theoretical viewpoint, we obtained several results on discrete models, which include analysis of relations between the compression ratio and the numbers of layers and nodes in autoencoders using linear threshold activation functions.

研究分野：数理生物情報学

キーワード：逆問題　ニューラルネットワーク　整数計画法　ケモインフォマティクス　バイオインフォマティクス　グラフアルゴリズム　特徴ベクトル　生成AI

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 人工知能やバイオインフォマティクスなどの分野において「データからの性質もしくはクラスの予測」は中心的課題の一つであり、これまでに数多くの研究が行われてきた。具体的に DNA 配列、タンパク質配列、自然言語などの文字列データを直接予測関数に入力するのは困難であるため、これまでの多くの研究では「入力データから特徴量を取り出し、それらをベクトル化した特徴ベクトルを得て、特徴ベクトルに対して学習や予測を行う」という手順により行われてきた。しかしながら、近年は深層学習技術の発展により特徴ベクトルを経由せずに入力データから直接、性質やクラスを予測することが可能になりつつある。そこで自然な「問い」として、「予測技術を新規物質などの設計に活用できないか？」という考えが生まれる。

(2) 制御理論や機械工学などにおいては「一部の変数のみに対する時系列データを観測することにより過去のある時点の内部状態を推定する」という観測問題が研究されてきた。設計問題ではないが「結果から原因（元データ）を推定する」という点において共通点がある。もちろん制御理論においては多くの研究蓄積があるが、複雑ネットワークという観点から近年は新たな展開が図られつつある。

(3) これらの議論に基づき、順方向の関数が与えられた時、もしくは、順方向の関数を機械学習技術などにより得た後、逆関数を計算することにより新規構造を設計したり内部状態を推定したりすることが考えられる。もちろん、逆関数や逆問題についてはこれまで様々な研究が行われてきたが、化学構造や配列データなどの離散データを対象として研究は十分に行われていなかった。

２．研究の目的

多くの予測・分類問題ではデータが与えられた場合に性質やクラスなどを予測することが目的となる。近年では望ましい性質が与えられた場合に、その性質を満たすデータを設計する一種の逆問題が研究されつつある。本研究では、この問題を写像と写像後のデータが与えられた場合に、元のデータ、すなわち、原像を推定する問題として定式化し、その理論的性質を解析するとともに、大規模システムに対して実際に原像を計算する手法を開発する。応用問題は離散データを扱うものを主対象とし、特に化学構造の設計問題や、時系列データから離散システムの内部状態を推定するためのマーカー検出問題を中心に新規で有用な計算手法を開発する。

なお、離散原像問題に基づく新規データ設計のためには、良い精度で予測を行う関数を得ることや、そのメカニズムを解析することも必要であるので、タンパク質配列データなどに対する新規の予測手法の開発や、ニューラルネットワークの離散モデルの理論解析なども並行して行う。

３．研究の方法

構造データの設計問題については、現実問題における主要なグラフ構造の一つである化学構造データを中心に研究を行う。予測関数としては、ReLU(Rectified Linear Unit)関数を活性化関数とする階層型のニューラルネットワークを主対象とする。原像の計算は一般に計算困難なクラスに属するので、それらのクラスに対して実際に最適解や厳密解が計算可能なことが知られている混合整数線形計画法を主に用いる。ただし、すべてのグラフ構造を対象とすると効率的な処理が困難になるため、多くの化学構造を表現可能な数理モデルを開発する。なお、開発にあたっては実際の化学構造データベースなどを活用し、中規模の化学構造データに対して、実際に原像を計算可能とすることを目標とする。

予測問題については、階層型のニューラルネットワークを主に使用し、タンパク質配列データや遺伝子発現データなどを主対象に高精度の予測を行う手法を開発する。また、ニューラルネットワークの理論解析にあたっては、線形閾値関数に基づく階層型のモデルを主に用いて、離散的手法による解析を行う。

４．研究成果

(1)区分線形関数（ReLU 関数を含む)を活性化関数とする階層型のニューラルネットワークに対し、出力ベクトルから入力ベクトルを求めるという原像問題を混合整数線形計画問題（MILP）として定式化し、MILP ソルバーを用いて入力ベクトルを計算する手法を開発した。この手法を実装した結果、入力層、中間層が数百頂点からなり、出力層が１個の頂点からなるネットワークに対し、通常の CPU を用いて数秒以内で計算結果（厳密解）を得ることができた。

(2)上記の(1)で示した手法と、以前より開発していた特徴ベクトルからの木状化合物の列挙方法を組み合わせ、与えられた性質を持つ木状化合物を列挙する手法を開発した。さらに、この MILP の定式化の改良を行い、制約を満たす化学構造が少なくとも一つ存在する場合に限り、MILP が解を持つようにした。それに加え、１個の環がある場合、および、２個の環がある場合に対応できるよう、構造の定義法、MILP の定式化などを拡張した。これらの手法について、実際の化

学構造データを用いた計算機実験によりその有効性を示した。

(3)より一般的な化学構造に対応するために、化学グラフを内部と外部に分けて捉える二層モデルという概念を導入し、そのモデルに対する原像問題を解くための MILP による定式化を開発、実装し、計算機実験により有効性を確認した。さらに、制約を満たす多数の化学グラフを系統的に列挙するために、MILP で１個の解を求めた後に動的計画法を適用するという新規な列挙法を開発し、計算機実験によりその有効性を確認した。また、これらの研究に関連して開発してきた一連のソフトウェアを mol-infer と命名し、そのソースコードを GitHub 上で公開した（https://github.com/ku-dml/mol-infer)。

(4)化学グラフに関する原像問題について、ニューラルネットワーク以外の予測モデルと MILP との組み合わせについても研究を行った。具体的には、Lasso と呼ばれる制約つきの線形回帰モデルを用いる方法、および、決定木モデルを用いる方法などを開発・実装した。そして計算機実験により、それらの有効性を確認した。

(5)特徴ベクトルから環を持つ化学構造を効率的に列挙するための新たなアルゴリズムを開発した。具体的には１個の２連結成分を持ち、その辺数が頂点数より１個多い化学グラフについて、重複も過不足もなく異なる構造をすべて列挙するアルゴリズムを開発、実装し、計算機実験により有効性を確認した。

(6)ニューラルネットワークの応用面についても研究を行い、遺伝子発現データと他の情報を統合して解析するための 2 種類の手法を開発した。一つは遺伝子発現データとタンパク質相互作用ネットワークデータを統合して解析する手法であり、タンパク質相互作用ネットワークデータをグラフラプラシアンを用いて２次元点集合に変換し、遺伝子発現データを対応する点の強度とすることにより、各サンプルのデータを画像データとして扱えるようにし、それに対し深層学習による画像解析技法を適用することにより腫瘍細胞の分類を行う。もう一つは遺伝子発現データと遺伝子間の進化的距離を統合して解析する手法であり、進化的距離データに多次元尺度構成法を適用することにより各遺伝子を２次元点集合に変換し、前者と同様の手法を適用することにより、腫瘍細胞のサブタイプの分類を行う。いずれも公開データから取得した遺伝子発現データを用いた計算機実験により、その有用性を示した。

(7)学習したニューラルネットワークからブール関数を抽出する問題に取り組み、既存手法と比較してより広いクラスのブール関数を抽出する手法を開発した。具体的には Nested Canalyzing Function というクラス、多数決関数というクラスとそれらを組み合わせたクラスのブール関数を抽出できるような手法を開発した。さらに、各入力変数や入力変数の組の値と出力との確率的な関係を抽出するために、動的計画法に基づく手法も開発した。これらの手法について計算機実験を行い、その有効性を評価した。

(8)遺伝子ネットワークの離散数理モデルであるブーリアンネットワークにおいて、一部の頂点の状態を知るだけでネットワーク全体がどの定常状態にいるのかを同定する問題に以前の研究で取り組み、そのための最小頂点数の計算手法を提案していた。これは遺伝子ネットワークからマーカー遺伝子を検出する問題をモデル化したものである。本研究では、このモデルをノイズのある場合に拡張し、それに対応した新たな計算手法を開発し、理論解析および計算機実験による解析を行った。

(9)ブーリアンネットワークにおいて周期的定常状態を効率良く検出することは、その内部状態を推定する上で重要である。定常状態に関する事前知識を活用することにより、従来手法より理論的に効率よく検出する手法を開発した。具体的には、定常状態における各頂点の状態の 0,1 の確率が与えられている場合に、全体の確率が高い順から低い順へと探索することにより、平均的に高速に動作するアルゴリズムを開発した。そしてシミュレーションデータおよび細胞周期などに関する大規模ブーリアンネットワークモデルを用いた計算機実験によりその有効性を確認した。

(10)線形閾値関数を活性化関数とする階層型ニューラルネットワークに基づく自己符号化器において、中間層における圧縮データのサイズと、層数、頂点数の関係を理論的に解析した。なお、モデル化にあたっては、入力ベクトルと出力ベクトルが完全に一致するという条件のもとで解析を行った。その結果として、中間層一層の場合に、最適な符号化は行えるが、復号化は不可能である場合が存在することなどを示した。さらに、得られた理論的結果の一部を計算機シミュレーションにより検証した。

５．主な発表論文等

〔雑誌論文〕　計28件（うち査読付論文　28件／うち国際共著　12件／うちオープンアクセス　12件）

１．著者名 Azam Naveed Ahmed, Zhu Jianshen, Sun Yanming, Shi Yu, Shurbevski Aleksandar, Zhao Liang, Nagamochi Hiroshi, Akutsu Tatsuya	４．巻 16
２．論文標題 A novel method for inference of acyclic chemical compounds with bounded branch-height based on artificial neural networks and integer programming	５．発行年 2021年
３．雑誌名 Algorithms for Molecular Biology	６．最初と最後の頁 18
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s13015-021-00197-2	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Nagamochi Hiroshi, Zhu Jianshen, Azam Naveed Ahmed, Haraguchi Kazuya, Zhao Liang, Akutsu Tatsuya	４．巻 20
２．論文標題 機械学習QSARの整数計画法に基づく逆解析法	５．発行年 2021年
３．雑誌名 Journal of Computer Chemistry, Japan	６．最初と最後の頁 106～111
掲載論文のDOI（デジタルオブジェクト識別子） 10.2477/jccj.2021-0030	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Cheng Xiaoqing, Ching Wai-Ki, Guo Sini, Akutsu Tatsuya	４．巻 130
２．論文標題 Discrimination of attractors with noisy nodes in Boolean networks	５．発行年 2021年
３．雑誌名 Automatica	６．最初と最後の頁 109630
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.automatica.2021.109630	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Munzner Ulrike, Mori Tomoya, Krantz Marcus, Klipp Edda, Akutsu Tatsuya	４．巻 18
２．論文標題 Identification of periodic attractors in Boolean networks using a priori information	５．発行年 2022年
３．雑誌名 PLOS Computational Biology	６．最初と最後の頁 e1009702
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pcbi.1009702	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Zhu Jianshen, Wang Chenxi, Shurbevski Aleksandar, Nagamochi Hiroshi, Akutsu Tatsuya	4．巻 13
2．論文標題 A novel method for inference of chemical compounds of cycle index two with desired properties based on artificial neural networks and integer programming	5．発行年 2020年
3．雑誌名 Algorithms	6．最初と最後の頁 124
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a13050124	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Shi Yu, Zhu Jianshen, Azam Naveed Ahmed, Haraguchi Kazuya, Zhao Liang, Nagamochi Hiroshi, Akutsu Tatsuya	4．巻 22
2．論文標題 An inverse QSAR method based on a two-layered model and integer programming	5．発行年 2021年
3．雑誌名 International Journal of Molecular Sciences	6．最初と最後の頁 2847～2847
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/ijms22062847	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Liu Pengyu, Melkman Avraham A., Akutsu Tatsuya	4．巻 126
2．論文標題 Extracting boolean and probabilistic rules from trained neural networks	5．発行年 2020年
3．雑誌名 Neural Networks	6．最初と最後の頁 300～311
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.neunet.2020.03.024	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Matsubara Teppei, Ochiai Tomoshiro, Hayashida Morihiro, Akutsu Tatsuya, Nacher Jose C.	4．巻 17
2．論文標題 Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles	5．発行年 2019年
3．雑誌名 Journal of Bioinformatics and Computational Biology	6．最初と最後の頁 1940007
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S0219720019400079	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Chun-Yu, Ruan Peying, Li Ruiming, Yang Jinn-Moon, See Simon, Song Jiangning, Akutsu Tatsuya	4．巻 17
2．論文標題 Deep learning with evolutionary and genomic profiles for identifying cancer subtypes	5．発行年 2019年
3．雑誌名 Journal of Bioinformatics and Computational Biology	6．最初と最後の頁 1940005
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S0219720019400055	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Nishiyama Yuhei, Shurbevski Aleksandar, Nagamochi Hiroshi, Akutsu Tatsuya	4．巻 16
2．論文標題 Resource cut, a new bounding procedure to algorithms for enumerating tree-like chemical graphs	5．発行年 2019年
3．雑誌名 IEEE/ACM Transactions on Computational Biology and Bioinformatics	6．最初と最後の頁 77-90
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TCBB.2018.2832061	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計23件（うち招待講演　8件／うち国際学会　19件）

1．発表者名 Akutsu Tatsuya
2．発表標題 On the Compressive Power of Boolean Threshold Autoencoders
3．学会等名 The 1st Online Conference on Algorithms（招待講演）（国際学会）
4．発表年 2021年

1．発表者名 Nagamochi Hiroshi, Zhu Jianshen, Azam Naveed Ahmed, Haraguchi Kazuya, Zhao Liang, Akutsu Tatsuya
2．発表標題 機械学習QSARの整数計画法に基づく逆解析法
3．学会等名 日本コンピュータ化学会2021年春季年会
4．発表年 2021年

１．発表者名
Zhu Jianshen、Azam Naveed Ahmed、Haraguchi Kazuya、Zhao Liang、Nagamochi Hiroshi、Akutsu Tatsuya

２．発表標題
An Improved Integer Programming Formulation for Inferring Chemical Compounds with Prescribed Topological Structures

３．学会等名
34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems（国際学会）

４．発表年
2021年

１．発表者名
Azam Naveed Ahmed、Zhu Jianshen、Haraguchi Kazuya、Zhao Liang、Nagamochi Hiroshi、Akutsu Tatsuya

２．発表標題
Molecular Design Based on Artificial Neural Networks, Integer Programming and Grid Neighbor Search

３．学会等名
2021 IEEE International Conference on Bioinformatics and Biomedicine（国際学会）

４．発表年
2021年

１．発表者名
On Control and Observation of Attractors in Boolean Networks

２．発表標題
Akutsu Tatsuya

３．学会等名
International Symposium on the Genetics of Industrial Microorganisms 2019（招待講演）（国際学会）

４．発表年
2019年

〔図書〕　計1件

〔産業財産権〕

〔その他〕
化学構造推定のために開発した一連のソフトウェア群をmol-inferと命名し、GitHub（ソフトウェア開発の公開レポジトリ）上で以下のURLより公開した。
https://github.com/ku-dml/mol-infer

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
研究分担者	永持　仁 (NAGAMOCHI Hiroshi) (70202231)	京都大学・情報学研究科・教授 (14301)	
研究分担者	細川　浩 (HOSOKAWA Hiroshi) (90359779)	京都大学・情報学研究科・講師 (14301)	

	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
研究協力者	ソン　ジャンニン (SONG Jiangning)		
研究協力者	前川　真吾 (MAEGAWA Shingo) (30467401)	京都大学・情報学研究科・助教 (14301)	

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			
オーストラリア	Monash University			
その他の国・地域	国立交通大学			
中国	The University of Hong Kong	Xi'an Jiaotong University		