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　３ 年　 ６ 月　 ７ 日現在

機関番号：３４５０４
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００３２０

研究課題名（和文）質の異なる化合物データベースからの薬理活性予測モデルの学習

研究課題名（英文）Learning on Structure-Activity Relationship from Heterogenous Chemical Compound Databases

研究代表者
猪口　明博（Inokuchi, Akihiro）

関西学院大学・理工学部・教授

研究者番号：７０４５２４５６

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：成果の1つは，大量のグラフからなるデータベースから，クエリグラフに含まれるグラフを検索するための方法論を確立したことである．我々の手法は，データベース中のグラフを表現する表現形式であるグラフコードの接頭辞木に基づいている．この接頭辞木を索引として用いることで，データベース中の多数のグラフとクエリグラフの間の部分グラフ同型判定問題を同時に解くことができる．もう1つの成果は，深層学習法のGCNにおけるover smoothing現象の軽減に関するものである．我々の方法では，GCNとdense connectionを組み合わせることで，既存研究に比べ，数%の精度向上が得られることが分かった．

研究成果の学術的意義や社会的意義
深層学習法は，現在，盛んに研究が実施されている研究分野である．その中で，様々なタイプのデータが解析されているが，グラフは非常に高い表現力を有していることが知られ，それに対するデータ解析手法や学習手法は非常に重要である．我々の成果は，データベースと機械学習の基礎研究分野への貢献であり，それらを発展させていくことで，実用化を目指せるものである．実用化の具体例の1つは，創薬化学分野である．

研究成果の概要（英文）：In this work, we investigated various approaches for learning on structure-activity relationship from heterogenous chemical compound databases. The first outcome is establishing a methodology for efficiently searching graphs contained in a query graph from a database consisting of a huge amount of graphs. Our approach is based on the prefix tree of graph codes that represent the graphs in the database. By using the prefix tree as an index, we simultaneously compute the subgraph isomorphism problem (which is known to be NP-complete) between the query graph and multiple graphs in the database. The second outcome is reducing the over smoothing phenomenon in Graph Convolution Networks in the deep learning research domain. In our approach, we combined Graph Convolution networks with the dense connection, and increased a certain percentage of prediction accuracies for various benchmark datasets compared with some conventional methods.

研究分野：人工知能

キーワード：機械学習　データベース　グラフ　構造活性相関

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

新薬には莫大な研究開発費と期間が必要である．創薬における最近の問題は，市場に出る新規医薬品数の鈍化である．効率化された実験装置の導入により，研究開発段階で合成される化合物の数は飛躍的に増加する一方で，市場に出る新規医薬品数は大きくは増えていない．これに対しデータ解析分野では，過去の実験結果データを計算機に学習させ，まだ実験していない化合物の薬理活性や副作用を予測（構造活性相関）する研究が進んでいる．これにより，活性の高いと予測された化合物から優先的に合成・実験することで，研究効率が改善されると期待されている．

一方，本研究課題を申請した当時，画像認識において，深層学習法を用いた飛躍的な認識正答率の向上が見られた．ILSVRC というコンテストでは学習に 1200 万枚の画像が用いられ，152 層からなるニューラルネットワーク(NN)が認識正答率 96.43%を達成した．正答率はそれ以前の 5 年で約 25%も増加しているが，その一翼を担っているのが事前学習法と大規模データの活用であり，過学習を防ぐ効果がある．深層学習法は，化合物の活性予測にも適用されはじめていたが，学習に用いられる化合物数が少ないため，深層学習法を使わない従来手法と比較して，その改善は小幅に留まっていた．

深層のニューラルネットワーク(NN)による化合物活性予測では，少量の化合物データしか使われておらず，事前学習法を十分に活用できていない．一方で，我々は，人工的に化合物を列挙するソフトウェアを検討しており，それが完成すれば大規模な化合物 DB を活用できると考えた．この化合物 DB を NN の事前学習に活用できれば，より深層な NN を構築でき，予測精度の改善が達成できる可能性があると考えた．

２．研究の目的

本研究では深層学習法を用いて化合物の活性予測を行う．この際，過学習を避け予測精度を高めるには，大規模データが必要となる．一方で，学術分野で用いられている化合物データベースは僅か数百化合物程度しかない．そこで，大量の化合物を含む人工化合物データベースを活用し，事前学習を行うことで，学習対象を表現するのに適したパラメータを事前に会得し，目的化合物データベースに対する高精度な予測を目指す．

３．研究の方法

上記の目的を達成するには，以下の 3 つの技術が必要になると考えた．

(1) 人工化合物の列挙アルゴリズム
(2) 包摂グラフ検索データベース
(3) グラフデータに対する深層学習法の事前学習法

(1)に関しては，当時，文献[1]が発表されていた．文献[1]は，C（炭素），N（窒素），O（酸素），S（硫黄），ハロゲン元素で構成される 17 原子以下の化合物を人工的に列挙する手法に関するものである．この論文では 10 万 CPU 時間を掛けて，約 1660 個もの化合物を人工的に列挙した成果が掲載されている．その列挙には，GENG と呼ばれるアルゴリズムが使われ，一旦，1140 億個ものグラフが一旦生成されたのちに，その 99.995%ものグラフが廃棄されるため無駄が多い．廃棄するグラフを列挙しないように制御できれば，膨大な計算時間を削減できる可能性があるため，その改良を行う．

(1)の列挙アルゴリズムの動作内で，多数のフィルタ[1]が使われる．フィルタとは，化合物がある部分構造を含むと自然界では安定して存在しないために，列挙する化合物を制限するたえに使われる．例えば，3 個の炭素の間に 2 個の二重結合が連続した部分構造（C=C=C）をもつ不飽和化合物は，文献[1]の実験では列挙されない．フィルタをグラフで表すと以下のような問題に帰着できる．

グラフの集合 D={g1,g2,…,gn}とクエリグラフ q が入力として与えられたときに，q に部分グラフとして含まれるグラフを D から検索せよ．

ここで，D に含まれる各グラフがフィルタに相当し，グラフ列挙アルゴリズムが生成する 1 つのグラフが q に相当する．q をクエリグラフとして，D のグラフを検索したときに，C=C=C が出力されるならば，その q は人工化合物としては出力しない．この問題は，包摂グラフ検索問題と呼ばれ，(1)を高速に動作させるためには，この問題を高速に解くための手法が必要になる．これが上記の(2)に相当する．

このようにして，人工化合物データベースが作成できれば，人工化合物データを G，機械学習分野でベンチマークデータとして使われている MUTAG データなどを利用し，G で事前学習し，それにより得られたパラメータで，MUTAG に対する学習を行えば，MUTAG のみの学習で得られる学習精度よりも高い性能を得られる可能性がある．

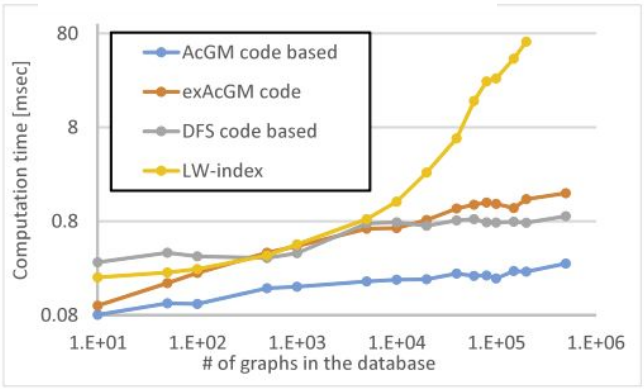
[1] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond: Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. Journal of Chemical Information and Modeling, 2864-2875, 2012.

４．研究成果

延長期間を含めた 4 年間の研究期間において，最も時間を割いたのは，研究方法の項目で述べた(2)に関する研究である．まずそれについて述べる．包摂グラフ検索問題では，データベース中のグラフがクエリグラフに含まれるかを判定する部分グラフ同型判定問題を解かなければならない．部分グラフ同型判定問題は NP 完全であることが知られているため，n 個（n は数万以上を想定）のグラフを含むデータベースに対して，単純な方法ではこの問題を実用的な計算時間で解くことは困難である．我々は，グラフをグラフコードと呼ばれる表現形式で表し，多数のグラフに対してグラフコードの接頭辞木による索引（インデックス）を作成することで，計算を効率化した．我々の手法の利点の 1 つは，接頭辞木を用いることで，データベース中の複数のグラフとクエリグラフの部分グラフ同型判定問題を同時に解けることである．また，データベース中のグラフ g とクエリグラフ q に対する部分グラフ同型判定問題を解く前に，g の部分グラフと q に対する部分グラフ同型判定問題を解くため，フィルタリングの効果が得られ，解とならないグラフを早期に，短い時間で判定することができる．

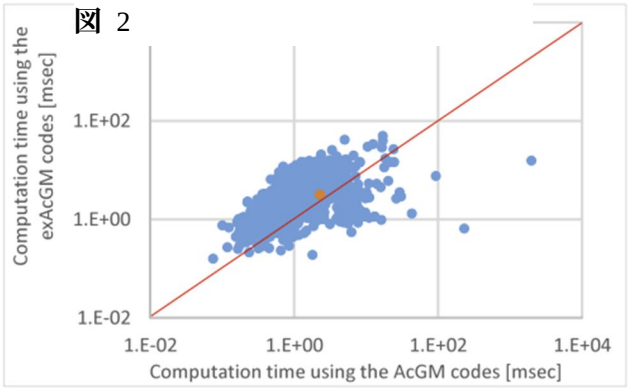
図 1 は，包摂グラフ検索問題を解く手法である LW-index と我々の手法 CodeTree の検索速度の比較である．横軸は，データベースに含まれるグラフの数であり，縦軸は 1 つのクエリグラフに対する計算時間（応答時間）である．データベース中のグラフ数を増やすと，LW-index の計算時間が大きく増える一方で，CodeTree の計算時間はそれほど増えない．LW-index はデータベースにグラフを追加したり，削除したり，変更したりすると索引の再構築に膨大な計算時間を要する．それに対して，CodeTree の索引再構築の計算時間は軽微であり，変更のないグラフに対して何もする必要がない．

図 1



CodeTree は，様々なグラフコードに対して適用可能な包摂グラフ検索のためのフレームワークである．図 2 は，AcGM コードと拡張 AcGM コードを CodeTree に組み込んだときの計算時間である．1 つの青点(x,y)は，AcGM コードを組み込んだ CodeTree での，あるクエリ q に対する計算時間 x と拡張 AcGM コードを組み込んだ CodeTree での，q に対する計算時間 y である．組み込むコードを変えることで，計算時間に関する性能を大きく変えることが分かった．最右の青点は，AcGM コード版の CodeTree だと 30 秒近くの計算時間がかかるのに対して，拡張 AcGM コードを使うことで，0.1 秒以下に抑えることができることを表している．今後，まだ世界中で開発されていない様々なグラフコードをこの CodeTree に組み込んで，その性能を評価する予定である．

図 2



続いて，研究方法の(3)で述べた成果について述べる．グラフに対する深層学習法においては，Graph Convolution Network（GCN）が注目を浴びており，事前学習法を含め様々な研究成果が世界中で報告されている．ただし，この GCN には，over smoothing と呼ばれる現象が報告されており，深層化することで性能が劣化することが分かっている．この性能劣化を緩和するために GCN に Densely Connection を組み込んだ．その結果，over smoothing を軽減できることを確認した．図 3 は，6 層 GCN を Cora データに適用したときの結果である．同じ色の点が塊になっているほど良い結果と言える可視化方法であるが，黒，赤，水色が混ざった塊が over smoothing によるものである．それに対して，我々の手法（図 4）は，上記の混ざりが緩和されていることが分かる．このように，GCN を多層化しても，Densely Connection を用いることで，over smoothing を緩和できることを確認した．

図 3

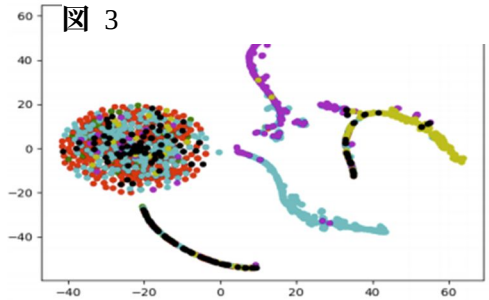
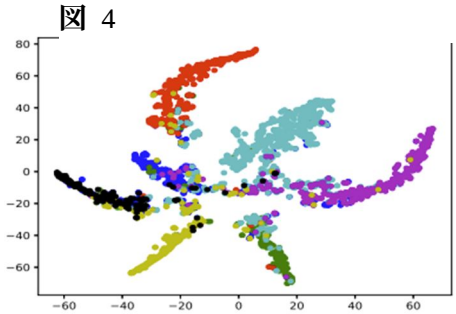


図 4



研究方法の(1)で述べた成果については，未公表の成果であるため，ここでは割愛し，今後，学会発表を通じて，公表する予定である．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Shun IMAI, Akihiro INOKUCHI	４．巻 E103.D 巻
２．論文標題 Efficient Supergraph Search Using Graph Coding	５．発行年 2020年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 130-141
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2019EDP7011	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8件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 谷岡豪，猪口明博
２．発表標題 アイテム密度に応じて分割されたオートエンコーダによる推薦システム
３．学会等名 電子情報通信学会 人工知能と知識処理
４．発表年 2019年

１．発表者名 木村至貴，猪口明博
２．発表標題 頂点併合と辺削除によるグラフ系列クラスタリングの効率化
３．学会等名 電子情報通信学会 人工知能と知識処理
４．発表年 2019年

１．発表者名 藤田将輝，猪口明博
２．発表標題 Generalized Multiple Kernel Learningを用いた近似グラフ編集距離の最適化による有機化合物の変異原性予測
３．学会等名 人工知能学会 第119回 知識ベースシステム研究会
４．発表年 2020年

１．発表者名
矢嶋 悠太，猪口 明博

２．発表標題
Attributed Network の高精度クラスタリングのための類似度行列の洗練

３．学会等名
人工知能学会 第119回 知識ベースシステム研究会

４．発表年
2020年

１．発表者名
桃田侑典，猪口明博

２．発表標題
人工化合物を用いたディープラーニングによる変異原性の予測

３．学会等名
人工知能学会第114回 知識ベースシステム研究会

４．発表年
2018年

１．発表者名
松井勇大，猪口明博

２．発表標題
交差検定を用いた説明属性の選択によるLocal SVMの正答率向上

３．学会等名
人工知能学会第114回 知識ベースシステム研究会

４．発表年
2018年

１．発表者名
Sousuke Takami and Akihiro Inokuchi

２．発表標題
Accurate and Fast Computation of Approximate Graph Edit Distance based on Graph Relabeling

３．学会等名
the 7th International Conference on Pattern Recognition Applications and Methods（国際学会）

４．発表年
2018年

１．発表者名 Xiaobo Xi and Akihiro Inokuchi
２．発表標題 Transition-based dependency parser with postponed determinations for Japanese sentences.
３．学会等名 International Conference on Asian Language Processing（国際学会）
４．発表年 2018年

〔図書〕　計2件

１．著者名 Maria De Marsico, Gabriella Sanniti di Baja, Ana Fred (編), Tetsuya Kataoka(著), Eimi Shiotsuki(著), Akihiro Inokuchi(著)	４．発行年 2018年
２．出版社 Springer	５．総ページ数 252
３．書名 Pattern Recognition Applications and Methods: 6th International Conference, ICPRAM 2017. Chapter 2, Graph Classification with Mapping Distance Graph Kernels	

１．著者名 Moamar Sayed-Mouchaweh (編), Sohei Okui, Kaho Osamura, and Akihiro Inokuchi	４．発行年 2018年
２．出版社 Springer	５．総ページ数 317
３．書名 Learning from Data Streams in Evolving Environments: Methods and Applications. Chapter 10, Detecting Smooth Cluster Changes in Evolving Graph	

〔出願〕　計1件

産業財産権の名称 人工化合物データを用いた化合物特性予測の深層学習法 および装置，並びに，化合物特性予測方法および装置	発明者 桃田侑典，猪口明博	権利者 学校法人関西学院大学
産業財産権の種類、番号 特許、2018-130287	出願年 2018年	国内・外国の別 国内

〔取得〕　計0件

〔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
研究分担者	田中　大輔 (Daisuke Tanaka) (60589399)	関西学院大学・理工学部・准教授 (34504)	

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