4H1-OS-27a-2

生命医科学 RDF データの機械学習・人工知能への応用

Application of Machine Learning and Artificial Intelligence Methods for Biomedical RDF Data

片山 俊明[*1]
Toshiaki Katayama

川島 秀一[*1]
Shuichi Kawashima

[*1] ライフサイエンス統合データベースセンター
Database Center for Life Science

In the life sciences and biomedical domains, there are a number of databases have been developed. Recently, some of those databases are released as RDF data which enables researchers to apply machine learning and artificial intelligence techniques. Here we show an application which generates feature vectors on each node in the RDF graph and a pilot study on those vectors where a deep learning method is applied.

1. はじめに

ライフサイエンス統合データベースセンター(DBCLS)は、何千種もある生命科学・医科学のデータベースを統合的に利用するための技術開発を行っている。この分野では、遺伝子の配列やその注釈、ゲノムとその変異情報、タンパク質や化合物の分子間相互作用や代謝と制御、タンパク質と医薬品などの3次元立体構造、生物種ごとの表現型やヒトの疾患、マルチオミックスの計測データなど実に多種多様なデータが混在している。このため、近年では、セマンティック・ウェブの技術を用いてデータベースを RDF 化することによって、有機的にデータ統合を進めることを目指してきた。DBCLS でも、各ドメインの専門家と協力してオントロジーの開発と RDF データの創出を進めており、これらのデータを検索・可視化するツールやウェブアプリケーションも多数開発してきている。

一方で、これらのデータセットは、これまでそのフォーマットや内容の多様さから、統合的に利用されることはなかった。RDF としてデータが揃ってはじめて多様なデータを組み合わせて解析対象とすることができるようになってきた。そのため、これらの膨大なデータセットを統合的に活用するデータ科学が求められている。とくに、膨大で多様なデータに潜む様々な関連性を引き出したり、そこから学習した結果を新たなデータの予測に活用する、機械学習や人工知能への期待が高まっているといえる。ここでは、現在利用可能な RDF データの概観と、RDF データに基づく機械学習の例、それを元に著者らが適用した実験について報告する。

2. RDF データ

DBCLS では、毎年データベース統合に関わる国際的な技術開発会議 BioHackathon を主催しており、そこで 2010 年ごろから分散する多様なデータベースを統合するために、セマンティック・ウェブ技術を使う方向性が示された [Katayama 2010]。すでに、アミノ酸配列とその機能アノテーションで世界最大のデータベース UniProt [The UniProt Consortium 2013]が、その維持管理を RDF 化によって省力化し、タンパク質に対する多様な情報を一元的に扱っていた。そこで、国内のデータベースを統合するにあたり RDF 化を推奨していくとともに、海外の主要データベースの RDF 化も国際的に推進するようになってきた。

2.1 国内の生命医科学 RDF リソース

国内の成果としては、科学技術振興機構のバイオサイエンスデータベースセンター(NBDC)に NBDC RDF ポータルが構築され(https://integbio.jp/rdf/)、これまでに 16 以上のデータベースが収録されている。この中には、タンパク質立体構造データベースとして世界標準の PDB や、国際塩基配列データベースの世界標準 INSD、がんゲノム、医薬品と遺伝子発現、糖鎖のデータベースなどが含まれ、国際的に主要なデータベースを中心に分野的にも幅広く RDF として提供するイニシアチブを取っている。ガイドラインによってクオリティコントロールをしていること、できるだけデータベース構築者自身によって RDF が作られるよう支援していることも特徴である。

2.2 海外の生命医科学 RDF リソース

海外では、欧州バイオインフォマティクス研究所(EBI)が、先述の UniProt を含む、ゲノムの Ensembl、パスウェイの Reactome など主要なデータベース 7 つを EBI RDF Platform で公開しているほか(https://www.ebi.ac.uk/rdf/)、米国の国立生物工学情報センター(NCBI)でも、化合物の PubChem やシソーラスの MeSH などを RDF 化して提供している。

3. 機械学習

2016 年の BioHackathon で RDF データを元にした機械学習の手法開発が行われた [Alshahrani 2016]。まず、RDF グラフからノードをランダムに選択し、その前後に繋がる RDF のステートメントをランダムウォークによって探索することで、関連度の高いものが隣接すると期待される一連のサブグラフを抽出する。次に、それをグラフ中でのコンテクストをもつ文とみなして word2vec [Mikolov 2013-1, Mikolov 2013-2] を適用する。この結果、RDF のグラフに含まれる各ノードについて特徴ベクトルが得られる。これを用いて、グラフ中のノード間の関係を予測する検証実験が行われた。

そこで、この手法を応用し、データセットを変えて、ヒトの遺伝子と疾患を介して繋がる MeSH タームの関係を学習することで、機能未知遺伝子の疾患への関連を推測するアプリケーションの実験を行った。

連絡先：片山俊明，大学共同利用機関法人 情報・システム研究機構 データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター，〒277-0871 千葉県柏市若柴 178-4-4 東京大学柏の葉キャンパス駅前サテライト 6 階，電話 04-7135-5508，Fax 04-7135-5534，ktym@dbcls.jp

3.1 データセット一覧

ここでは UniProt の RDF、MeSH の RDF、JST の科学技術用語シソーラスの SPARQL 検索結果を用いることとした。

- UniProt RDF
- JST シソーラス
- MeSH RDF

UniProt については、ヒトのタンパク質をサブセットとして取得した。また UniProt に付随する pathway, go, enzyme のオントロジー情報も合わせて取得した。MeSH については 2016 年版を NCBI から取得して利用した。また、科学技術用語シソーラスについては、2016 年 12 月時点でトライアル公開されていた JST の J-GLOBAL knowledge から SPARQL 検索によって必要なサブセットを取得した。

3.2 特徴ベクトルの生成

UniProt のデータでは、各タンパク質から疾患のノードを介して MeSH タームにリンクされている。また、JST シソーラスにも対応する MeSH へのリンクがあるため、シソーラスの各用語間の関連などを通じて概念がより明確に反映されることが期待できる。

これらのデータセットを1つにまとめ、RDF データ中の各ノードに ID を振り、ID 間のエッジをリンクセットとした。この際、今回は RDF データの中で URI だけでなくリテラルもノードとみなすこととした。BioHackathon の手法では推論したグラフも追加されていたが、今回は省略している。

生成したリンクセットを DeepWalk [Perozzi 2014] によってランダムウォークすることで、RDF データ中 1856 万ノードについて、各 64 次元の特徴ベクトルを計算した。

4. 結果

生命科学ではタンパク質の類似性は主にアミノ酸配列の相同性から推測されてきた。まずは、結果の検証として今回使用した UniProt RDF のヒトサブセットに含まれる 238,503 タンパク質のうち、機能的に近いものが特徴ベクトルとしても近いかどうかと、配列相同性は低いが別の生物学的な観点から近い特徴ベクトルが拾えているかどうかを解釈した。

4.1 特徴ベクトルの類似性と深層学習

タンパク質間の類似性では、配列の相同性が高いものが特徴ベクトルでも類似している傾向が見られた。一方で、配列の類似性は低いものの、同じパスウェイに載っているなど、機能的な類似性が高いものについても特徴ベクトルが近接することがわかった(図 1)。

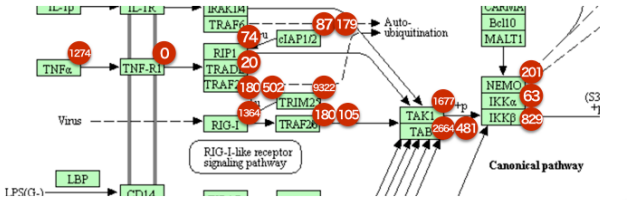

図 1: NF-kappa B signaling pathway において TNF receptor に対する特徴ベクトルの類似性の順位をパスウェイ上で周辺のタンパク質に対して表示している。

一方で、これらの特徴ベクトルのうち、タンパク質と MeSH タームの関係について、正解とされている組についてそれぞれの特徴ベクトルをペアにして深層学習を行い、ジャックナイフテストによって予測を行った結果 84%の再現性が得られた。これについては新規データでの予測が可能か、また他の特徴の組み合わせの学習と予測などを試行し、今後さらなる検証を進める必要がある。

4.2 今後の方向性

RDF を利用した機械学習・人工知能の研究はまだ始まったばかりで、どのような入力データを選別して何を学習すればよいか、その際のパラメータはどのようにするのが良いか、人工知能においてはどのように深層学習のレイヤーを組むのが良いかなど、まだまだ手探りの状況である。現在ヒトの遺伝子変異と疾患などの表現型に関する RDF データの構築を進めているところで、これらのデータを含めた解析が行えるようになることで、実際的なゲノム医科学へも貢献できると期待している。

参考文献

[Katayama 2010] Toshiaki Katayama: The 3rd DBCLS **BioHackathon**: improving life science data integration with Semantic Web technologies. J Biomed Semantics. 4(1):6, 2013

[Alshahrani 2016] Mona Alshahrani: Neuro-symbolic representation learning on biological knowledge graphs, arXiv:1612.04256, 2016

[Mikolov 2013-1] Tomas Mikolov: Efficient Estimation of Word Representations in Vector Space

[Mikolov 2013-2] Tomas Mikolov: Distributed Representations of Words and Phrases and Their Compositionality, The Proceedings of Neural Information Processing Systems 2013, 2013.

[The UniProt Consortium 2013] The UniProt Consortium: Update on activities at the Universal Protein Resource (UniProt) in 2013,Nucleic Acids Research Vol. 41(Database issue), pp D43-D47, 2013.

[Perozzi 2014] Bryan Perozzi: DeepWalk: Online Learning of Social Representations, Knowledge Discovery and Data Mining 2014, 2014