人やモノのつながりを表すグラフデータから新たな知見を導く新技術Deep Tensor

Deep Tensor: Eliciting New Insights from Graph Data that Express Relationships Between People and Things

● 丸橋弘治

あらまし

人やモノのつながりにより表現されるグラフデータを，つながりの形状に基づいて分類することが求められている。こうした分類は，例えば通信ログのIPアドレスやポート番号の間のつながりに基づく攻撃の発見や，銀行取引履歴の口座や支店の間のつながりに基づく不正取引の発見といった問題において，特に重要である。しかし，大量のグラフデータを分類の対象とする場合，専門家が人手で設計した部分グラフの一致に基づく従来のグラフ分類手法では，高精度な分類の実現には限界があった。そこで筆者らは，グラフデータを高精度に解析できる機械学習技術Deep Tensor(以下，DT)を開発した。DTは，テンソル分解と呼ばれる技術を応用することにより，専門家の設計に頼ることなく，グラフデータの特徴量の抽出方法をニューラルネットワークと同時に学習する。

本稿では，全く異なる三つの分野のデータに適用した実験により，DTがグラフデータを従来手法より高精度で分類できることを示す。また，DTによる予測結果は，ニューラルネットワークの活性に基づく解釈が可能であることも示す。

Abstract

An important problem in information and communications technology (ICT) is classifying graph data that expresses the relationships between people and things. For example, how can cyberattacks be detected by using network traffic logs showing the relationships between the source IP address and the destination IP addresses and ports, and how can fraudulent activities be detected by using banking transactions showing the relationships between senders and receivers and bank branches? When classifying large volumes of graph data, however, there are many yet-to-be-expressed features in the partial graphs used in conventional graph learning methods, so there are limits to achieving accurate classification. We propose using a novel tensor decomposition method called “Deep Tensor” for leveraging a deep neural network to enable it to automatically extract these features of graph data. Experiments in three different domains demonstrated that use of this decomposition method results in high accuracy for various types of graph data, enabling interpretation based on the activity of the neural network.

ま え が き

近年，通信の高速化やIoT（Internet of Things）の進展によって大量のデータが発生・蓄積されている。その中でも，人やモノのつながりを表すデータは，グラフデータとみなして解析できる。例えば通信ログは，送信・受信IPアドレス（以下，送信・受信IP）およびポート番号のつながりを表すグラフデータとみなすことができる。**図-1**では通信ログを値（送信IP，受信IP，ポート番号）を表す四角形のノードと，つながり（ログID）を表す楕円のノードとの間の接続によって表現している。更に，テンソルと呼ばれる数学上の概念を用いて，値の組み合わせに対してつながりが存在するか否かを表現している。筆者らの目的は，グラフデータg_iとその分類y_iが与えられたときに，グラフデータgを入力とし，その正しい分類yが精度良く出力されるような予測モデルμを学習することである（**図-2**）。これにより，例えば，通信ログのIPアドレスやポート番号の間のつながりに基づく攻撃の発見や，銀行取引履歴の口座や支店の間のつながりに基づく不正取引の発見といった，富士通のAI（人工知能）技術「FUJITSU Human Centric AI Zinrai」が対象とする，重要な問題の解決が期待できる。このような実データに基づくグラフデータは，複雑かつ大量になることが多い。例えば実際の通信ログは，ごく短い期間であっても，複雑なグラフデータとなる（**図-3**）。そして，このような複雑なグラフデータが，日々大量に蓄積され続けている。しかし，大量のグラフデータを分類の対象とする場合，専門家が人手で設計した部分グラフと一致するかどうかで判定する従来のグラフ分類手法では，高精度な分類の実現に限界があった。

筆者らは，近年注目されているDeep Learningを発展させ，グラフデータに対して高精度な解析を可能とする新たな機械学習技術「Deep Tensor」（以下，DT）を開発した。DTは，グラフデータをテンソルで表現し，テンソル分解の応用により，グラフデータの特徴量の抽出方法をデータから自動的に学習する。

本稿では，ネットワーク侵入検知，融資仲介サービス，およびコンピュータ創薬のデータを用いた評価実験により，富士通研究所が開発したDTが従来の手法よりも高い精度で多くのグラフデータを分類し，更に効果的に解析可能であることを示す。

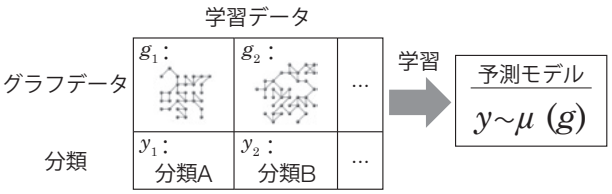


図-2　グラフデータの学習

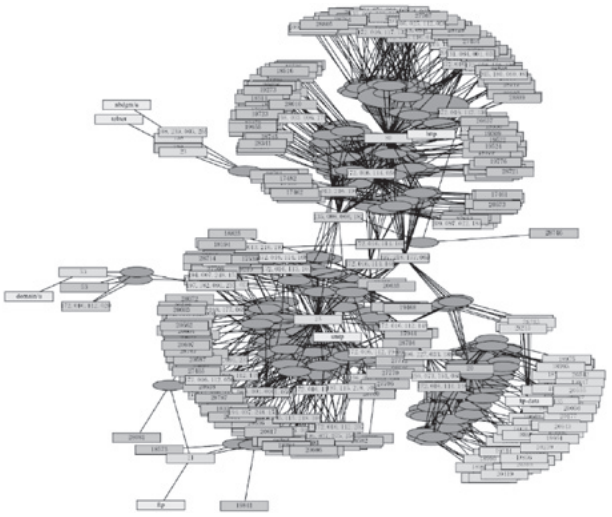


図-3　実際の通信ログに基づくグラフデータの例

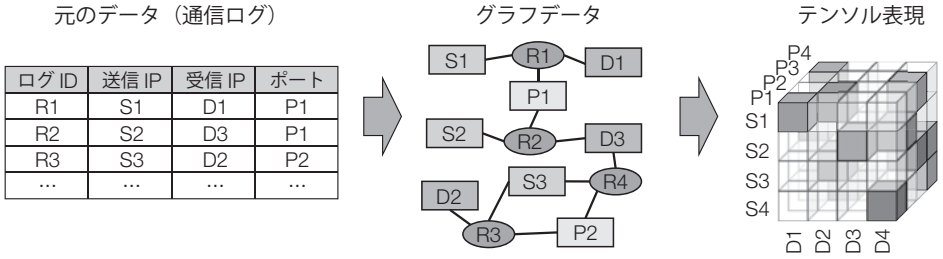


図-1　グラフデータとテンソル表現

グラフデータの特徴量抽出における課題

一般に機械学習では，データから特徴量を抽出する必要がある。本章では，その課題について述べる。

● 部分グラフの設計

従来，グラフデータの分類においては，専門家が設計した部分グラフが分類対象のグラフデータ中に含まれるかどうかを特徴量としていた（**図-4**）。グラフカーネルを用いたサポートベクターマシン（SVM）[(1)]においても，部分グラフを明に列挙することはないものの，基本的には同じ考え方である。しかし，大量かつ複雑なグラフデータを分類の対象とする場合，あらかじめ設計した部分グラフでは全ての特徴量を考慮することが困難なため，高精度な分類には限界がある。

● Deep Learningによる部分グラフの学習

専門家の設計に頼ることなくデータから特徴量を抽出する方法を学習可能な技術として，多層のニューラルネットワークを用いる機械学習技術であるDeep Learningが，画像や音声の認識において注目されている。[(2)] 一方，グラフデータに対しては，ノード間の接続関係を表す行列を画像とみなし，畳み込みニューラルネットワーク（CNN）を用いた分類方法が提案されている。[(3)] しかしこの方法では，行列の生成のためにノードの並び順を人手で設計する必要があり，分類に最適な並び順が自明ではない。

● テンソル分解による特徴量の抽出

ノードの並び順によらないグラフデータの変換方法として，テンソル分解の活用が考えられる。テンソル分解とは，コアテンソルと呼ばれるテンソルに要素行列と呼ばれる行列を掛け合わせて，得られたテンソルにより入力テンソルを近似する手法である（**図-5**）。[(4)] テンソル分解を使えば，グラフデータの主要なつながりの構造をコアテンソルとして取り出すことが可能である。[(5)] しかし，コアテンソルの要素をどのように並べ替えても表現されるテンソルは同じであるが，どれが分類に適したコアテンソルであるかは自明ではない。

この問題に対して筆者らが提案するDTは，分類に適したコアテンソルの抽出方法をデータから学習する。次章では，提案手法について述べる。

新技術Deep Tensor

DTの概要を**図-6**に示す。DTは，構造制約テンソル分解によりグラフデータをコアテンソルと要素行列に分解し，コアテンソルをニューラルネットワークに入力する。構造制約テンソル分解は，分類に重要な特徴量が表現されたターゲットコアテンソルにできる限り類似するように，コアテンソルを算出する。更に，従来のニューラルネットワークの学習で用いられる誤差逆伝搬法の拡張により，分類精度を高めるようにターゲットコアテンソルを最適化する。以下に各技術の要点を述べる。

● 構造制約テンソル分解

前述した従来のテンソル分解では，分類にとって重要な構造が，必ずしもコアテンソルの類似の

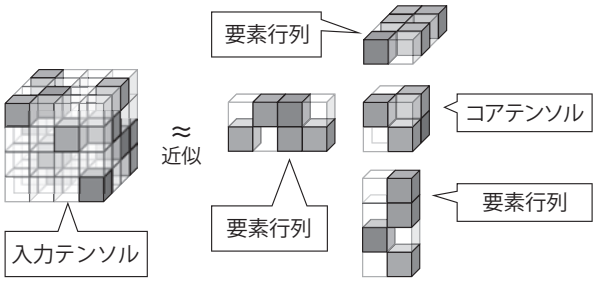


図-5　従来型のテンソル分解

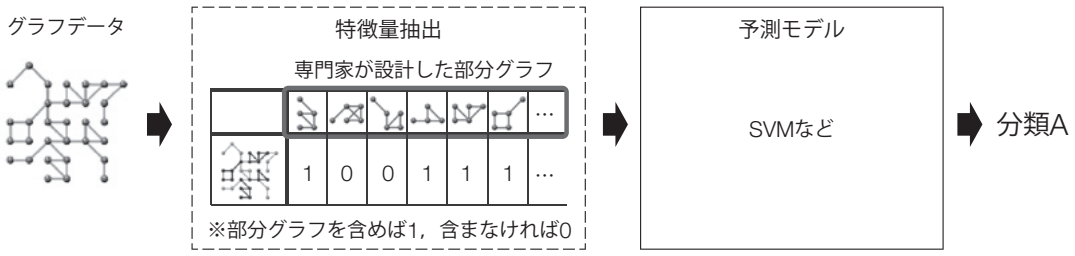


図-4　部分グラフの一致を用いる従来手法

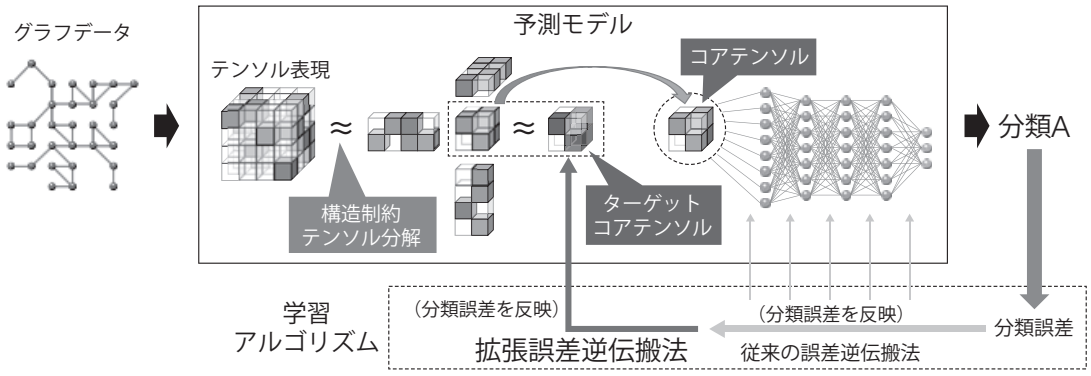


図-6　Deep Tensorの概要

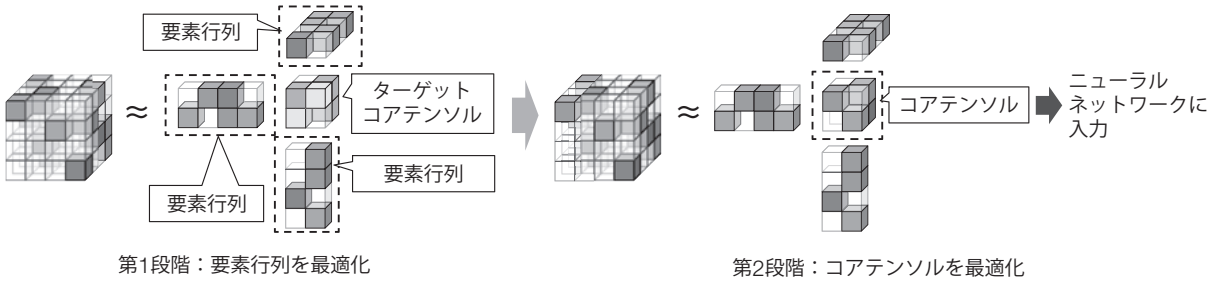


図-7　構造制約テンソル分解

位置に配置されるとは限らない。今回新たに開発した構造制約テンソル分解は，ターゲットコアテンソルに類似するようにコアテンソルを算出することにより，分類に重要な構造をコアテンソルの類似の位置に配置する。そして，このコアテンソルを用いてニューラルネットワークを学習することにより，精度が高い分類が可能となる。構造制約テンソル分解は，2段階の最適化により計算される（**図-7**）。第1段階では，与えられたターゲットコアテンソルを用いて，入力テンソルを最も良く近似するように要素行列のみ最適化する。第2段階では，第1段階で最適化された要素行列を用いて，入力テンソルを最も良く近似するようにコアテンソルを最適化する。

●　拡張誤差逆伝搬法

筆者らは，高い分類精度が得られるターゲットコアテンソルを算出するために，拡張誤差逆伝搬法を開発した。誤差逆伝搬法は，分類誤差を下層に伝搬させる形で，分類誤差を小さくするようなパラメーターの修正方向を算出するアルゴリズムである。拡張誤差逆伝搬法は，更にターゲットコアテンソルまで伝搬させ，ターゲットコアテンソルの修正方向を算出する。そして，確率的勾配法（SGD）により，ニューラルネットワークのパラメーターと同時にターゲットコアテンソルを更新する。

評　価　実　験

筆者らは，以下の三つの異なる分野のデータを用いて評価実験を行った。

●　データセット

（1）侵入検知

意図的に発生させた攻撃に由来するログが混入したネットワーク侵入検知ログ[6]から，7週間分の学習データと2週間分の評価データを取得した。この評価では，ログを10分ごとに区切り，各期間に攻撃が含まれているかどうかを，送信元IP，送信先IP，送信元ポート，送信先ポート（HTTPやFTPといったプロトコル名のラベル付き）の関係から予測する問題とした。予測モデルは，攻撃タイプ別に学習した。

（2）融資仲介サービス

融資仲介サービスの取引履歴[7]を使用する。融

資仲介サービスとは，貸し手と借り手をマッチングさせて個人間融資を仲介するサービスである。今回は，個人情報に基づいて10％以上の利率と判定された融資を高リスク融資とみなし，その融資に関連する貸し手（居住地ラベル付き）・借り手（居住地ラベル付き）・融資IDの間のつながりの構造から，高リスク融資を判定する問題とした。2012年のデータのうち，1月～ 11月のものを学習データ，12月のものを評価データとした。個人情報を全く用いることなく融資の構造からリスクを判定できるのであれば，フィンテックにおける新しいサービスの開拓につながる可能性がある。

（3）コンピュータ創薬

化合物分類のデータセット[8]を使用する。原子（元素記号ラベル付き）間の結合関係を表すグラフの構造に基づき，毒性や活性を予測する。

● **比較手法**

部分グラフの一致に基づく従来手法として，種々のグラフカーネルと組み合わせたSVMを用い，DTとの性能比較を行った。グラフカーネルとしては，Graphletカーネル（以下，GK），最短経路カーネル（以下，SP），Weisfeiler-Leman部分木カーネル（以下，WL）を用いた。また，構造制約テンソル分解の代わりに従来型テンソル分解を用いる手法（以下，Tucker）との比較も行った。更に，SGDがターゲットコアテンソルを更新しない手法（以下，noEBP）との比較も行った。

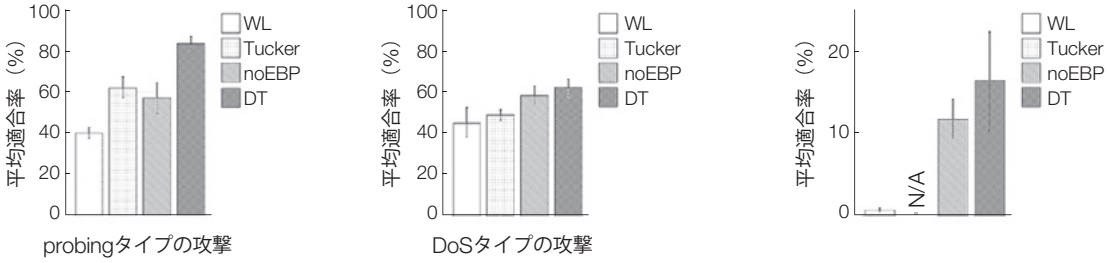
● **予測精度**

侵入検知と融資仲介サービスのデータによる評価結果を**図-8**に示す。侵入検知は，最もログ数の多いprobingタイプとDoS（Denial of Service）タイプにおける攻撃の検知精度を示す｛図-8（a)｝。なおGKとSPは，これらのデータでは値の数も関係の数も多すぎることから，膨大な計算時間とメモリ空間が必要となるため実行が不可能であった。またTuckerは，DoSタイプの攻撃の検知では同様の理由により実行が不可能であった。今回の評価では，全て平均適合率を用いて評価した。平均適合率とは，攻撃や高リスク融資の予測確率のしきい値を変化させたときの，適合率（検知数に対する正解数の割合）の平均値である。いずれの場合も，DTまたはnoEBPがWLやTuckerより精度が高く，またnoEBPよりDTの方が高い精度を示した。この結果は，構造制約テンソル分解が有効に働いており，更に拡張誤差逆伝搬法が多くの場合効果的であることを示している。特にprobingタイプの攻撃の検知では，WLやTuckerでは約半分が誤検知となるのに対し，DTは8割以上が正解となる実用的な検知精度を示した。一方，融資仲介サービスにおいては，DTの平均適合率は約16％と低い｛図-8（b)｝。しかし，DTは1,600以上ある評価用の融資の約2％を選択するだけで，10以下しかない高リスク融資の半数以上の検知に成功した。

化合物分類のデータセットは，10分割交差検証法により評価した。ほぼ全てのデータセットにおいて，WLが最も高い精度を示した。この結果は，化合物のデータに対して，このグラフカーネルがうまく設計されていることを示している。しかし，DTあるいはnoEBPは，いずれのデータセットに対しても全グラフカーネルの平均精度より高い精度を示しており，化合物データに対しても有効な手法であることが示唆された。

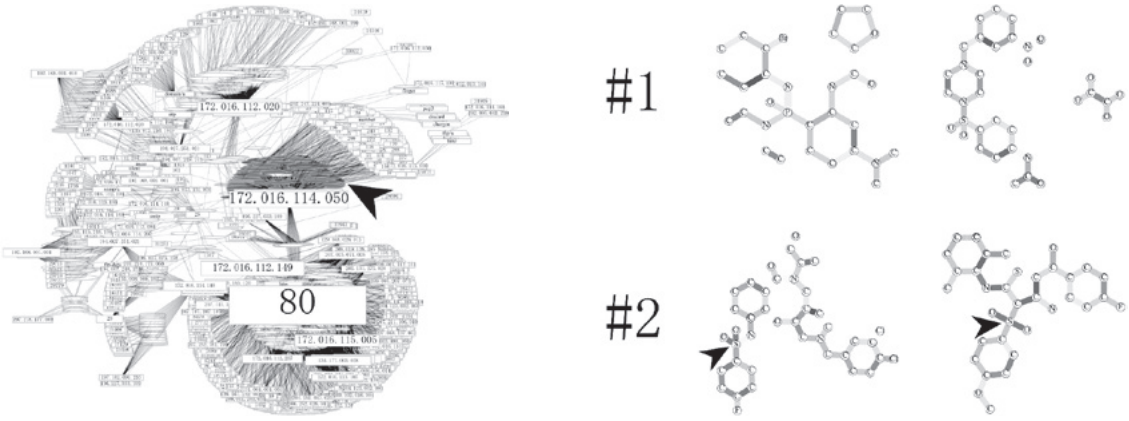
● **大規模データへの適用可能性**

筆者らは，大規模データへの適用可能性を検証



（a）侵入検知　　（b）融資仲介サービス

図-8　三つ以上の値のつながりのグラフデータにおける分類精度（上部のバーは標準誤差）



(a) 侵入検知（probingタイプの攻撃）　(b) コンピュータ創薬（PubChem BioAssay）

図-9　特定のニューロンの活性が大きいグラフデータ

するために，化合物のオープンなデータベースであるPubChem BioAssay[9] において，最も多くの活性化合物を含むデータセットへの適用実験を行った。10分割交差検証法による評価の結果，1,286個の学習データでは200回のSGDによる更新でも精度は約66％にとどまるのに対し，128,404個の学習データに対しては，わずか20回の更新にもかかわらず約75％の精度を示した。DTは，ほかのニューラルネットワークの手法と同様に，大量のデータを学習することにより，従来をはるかにしのぐ分類精度を実現する可能性があることが示された。

●　予測結果の解釈

予測結果は，ニューロンの活性に基づき解釈可能である。分類確率と最も相関の高い活性を持つニューロンを選択したときの，そのニューロンの活性が最も高い値を示したグラフデータを**図-9**に示す。侵入検知（probingタイプの攻撃）のデータにおいてニューロンの活性に最も寄与したログ｛図-9（a）の矢印部分｝は，HTTPサーバの一つが多数のポート番号によりアクセスされたログであり，意図的に行われたポートスキャン攻撃を示すログと一致していた。また，PubChem BioAssayのデータからは，上位二つのニューロンを選択した｛図-9（b）の#1と#2｝。これは，ニューロンごとに活性の高い二つの化合物を示したものである。＃2のニューロンでは，硫黄（S）と酸素（O）の2重結合の周辺の構造｛図-9（b）の矢印｝がニューロンの活性に強く寄与しており，これらの構造が化合物の活性・非活性の分類にとって重要であることを示唆している。

む　す　び

本稿では，グラフデータを分類するための新技術Deep Tensorについて解説した。Deep Tensorの特徴は，分類に重要なグラフデータの特徴量を抽出する構造制約テンソル分解を用いる点である。更に，構造制約テンソル分解で用いるターゲットコアテンソルを，ニューラルネットワークと同時に最適化する拡張誤差逆伝搬法を開発した。本手法を三つの異なる分野のデータを用いて評価し，高い分類精度が得られることを示した。また，予測結果の解釈可能性についても示した。

今回は比較的単純なモデルを用いて有効性を示したが，より大規模なニューラルネットワークを用いた場合や，既存の様々な形態のニューラルネットワークと組み合わせた場合，あるいはより多様なコアテンソルを用いた場合に，どの程度の性能が発揮できるかは未知数である。今後，多くの問題に適用しながら，本手法の可能性と限界を探っていくことが課題である。

本技術は今後，FUJITSU Human Centric AI Zinraiにおいて活用していく予定である。

参考文献

(1) N. Shervashidze et al.：Weisfeiler-Lehman

Graph Kernels. Journal of Machine Learning Research 12, p.2539-2561（2011）.
（2）A. Krizhevsky et al.：ImageNet Classification with Deep Convolutional Neural Networks. Proceedings of the 25th International Conference on Neural Information Processing Systems（NIPS'12）, p.1097-1105（2012）.
（3）M. Niepert et al.：Learning Convolutional Neural Networks for Graphs. Proceedings of the 33rd International Conference on Machine Learning（ICML'16）, p.2014-2023（2016）.
（4）T. G. Kolda et al.：Tensor Decompositions and Applications. SIAM Review Vol.51, Issue 3, p.455-500（2009）.
（5）Y. Lin et al.：MetaFac：Community Discovery via Relational Hypergraph Factorization. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining（KDD'09）, 527-536（2009）.
（6）DARPA Intrusion Detection Data Sets. http://www.ll.mit.edu/ideval/data/
（7）Show Me The Money. http://smtm.labs.theodi.org/
（8）ETH zurich. https://www.bsse.ethz.ch/mlcb/research/machine-learning/graph-kernels/weisfeiler-lehman-graph-kernels.html
（9）PubChem BioAssay. https://www.ncbi.nlm.nih.gov/pcassay

著者紹介

丸橋弘治（まるはし　こうじ）
人工知能研究所
人工知能基盤プロジェクト
機械学習によるデータマイニング，特にグラフデータからの知識抽出，アノマリ検知の研究に従事。