4J1-4

定量的構造活性相関予測における化合物特徴表現の実験的検証

Empirical Evaluation of Feature Representations of Chemical Compounds for QSAR

越野　沙耶佳 *1　岡崎　文哉 *1　瀧川　一学 *1*2
Sayaka Koshino　Fumiya Okazaki　Ichigaku Takigawa

*1北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido Universit

*2科学技術振興機構，さきがけ
JST,PRESTO

Quantitative Structure-Activity Relationship is a problem to predict its activity from chemical structure. In the field of drug discovery, it is common to derive feature quantities using chemical knowledge. Otherwise, in the field of computer science, machine learning methods for graphs have been proposed and applied to molecular graph data. But the performance comparison across the two fields has not been sufficiently investigated. In this research, machine learning approach using various feature representations is experimentally evaluated and discussed.

1. はじめに

創薬の分野において，薬の候補となりうる低分子化合物の活性を予測することは大きな課題である．そのため実験やシミュレーションと言った方法が取られてきた．しかし，突然変異誘発性やがん細胞の成長阻害性など，作用機序が複雑でモデル化が困難な活性も多いため，既知データに基づく統計的予測手法が創薬分野と情報科学分野においてそれぞれ研究されてきた．

創薬の分野では定量的構造活性相関として扱われ，活性値がわかっている化合物データから活性値が未知の化合物データの活性値を予測する教師有り学習の形で主に研究されている．その際の特徴ベクトルとしては化学的知識を用いて算出するのが一般的であり，そのために商用のものからオープンソフトウエアまで計算のためのツールが様々開発されている．そうやって算出された特徴量を用いて機械学習を行うことで活性値を予測しており，計算時間などのコストはシミュレーションに比べるとはるかに低い．

一方，情報科学分野ではグラフというデータ構造を用いて薬の化合物の活性予測が研究されている．グラフは幅広く用いられているデータ構造であり，低分子化合物の構造式の表現だけではなく自然言語処理の構文木表現などに様々用いられている．このように応用先が広いことからグラフを対象とした教師付き学習の研究が行われている [Saigo 09][Wale 08]．このように創薬分野，情報科学分野において化合物から活性を予測する問題は研究が行われているが，それぞれの手法との精度比較は十分になされていない．本研究では，両分野それぞれの特徴量を用いた機械学習を行い，実験的に精度を比較する．

2. 創薬分野における手法

化合物に対して活性を予測する問題を創薬分野では定量的構造活性相関予測として扱ってきた．定量的構造活性相関 (Quantitative Structure-Activity Relationship 以下 QSAR) とは医薬品などの化合物の構造と活性との間に対して相関関係を数値として定量的に求める方法である．化合物に対して化学的情報を用い特徴量を算出し，予測モデルを構築する教師有り学習の形で扱われている．QSAR において化合物の構造を反映する特徴量を記述子と呼び，一般な記述子としては部分構造の有無を判定するフィンガープリントや，化合物の物理化学的な性質の実測値や推算値が検討される．

2.1 Extended Connectivity Fingerprint

ECFP(Extended Connectivity Fingerprints)[Rogers 10] は Circular Fingerprint の一つで，現在広く使用されている．まず化合物に含まれるそれぞれの原子について，結合している原子や電気的特性に応じて atomic invariants と呼ばれる整数の識別子が与えられる．例えば原子は原子番号や原子の特性値が用いられる．各構成原子について 0 近傍，1 近傍，2 近傍... の部分構造を順に取り出していく．そしてここで得たすべての部分構造を Morgan 法によりシス，トランスといった立体化学も考慮された一意の整数に変換する．Morgan 法で得られた整数を特徴ベクトルの次元数 d で mod を取り，得られた値を j として x_{ij} を 1 とする．ただし x_{ij} は化合物 i の j 番目の特徴量を示す．本研究では商用記述子生成ソフトウエア DRAGON7 とオープンソース計算科学ソフトウエアである RDKit により標準値である，半径 2，1024 ビットの ECFP(DRAGON)，および，半径 2，2048 ビットの ECFP(RDkit) を用いた．これらで求められるフィンガープリントは ECFP とよく似たものとなっている．両者の違いとしては分子グラフの頂点にどのような化学情報を付与するかが挙げられる．

2.2 DRAGON 記述子

DRAGON*1 は最大 5270 種類の分子記述子を計算することができる分子記述子計算ソフトウェアである．本研究では DRAGON7.0 を使用した．求めることができる記述子として，官能基や環，原子タイプ，位相といった記述子がありその完全なリストはウェブサイトで見ることができる．例としては，分子量，原子ファンデルワールス量の和，原子分極率の和，第一イオン化ポテンシャルの合計などが挙げられ，与えられた化合物のこれらの特性値を特徴量として用いる．主に実数値をとる特徴量が多い．

3. 情報科学分野における手法

情報科学の分野では化学構造式をグラフとして表現し (分子グラフ)，グラフ分類問題として扱うことが多い．本節では訓練集合 S はグラフ集合 $\{G_i \mid i \in [n]\}$ とする．特徴ベクトルとしては例えば，データ中の部分グラフを調べ，その各々の有無を特徴量とする．

連絡先: 越野　沙耶佳，北海道大学大学院情報科学研究科，koshino@art.ist.hokudai.ac.jp

*1 http://www.affinity-science.com/_userdata/d7_desc_list.txt

本研究では頻出部分グラフ，辺数制約部分グラフ，gBoostを用いた．部分グラフは gSpan アルゴリズム [Yan 02] により列挙した．gSpan アルゴリズムでは部分グラフ列挙のために探索木というものを考える．探索木のノードはひとつの部分グラフを表し，親ノードが子ノードに包含されるようにエッジを張る．この探索木を深さ優先順にたどりながら部分グラフを列挙している．

この gSpan の探索木を利用すれば，有無を調べるべき部分グラフ特徴の探索とその特徴に基づく分類モデルを同時に学習することができる．このような方法としては今回用いた gBoostがある．

3.1 頻出部分グラフ

本稿においてグラフ分類問題とは，グラフに対する教師付き学習のことを指す．この時特徴ベクトルとして部分グラフの有無を用いる．しかしすべての部分グラフパターンの列挙は，出力数が組み合わせ爆発を起こすため，現実的に求めることは困難である．よって通常何かしらの方法で出力を打ち切る必要がある．今回は，ある閾値を与え，σ 個以上のグラフに出現するパターンを頻出部分グラフとして列挙した．この閾値 σ を最小支持度（以下，minsup）と呼ぶ．また部分グラフの大きさに関して maxpat という閾値を与えた．maxpat とは部分グラフに含まれる辺の数である．maxpat 以下の部分グラフのことを辺数制約部分グラフとして列挙した．

頻出部分グラフによるグラフ分類では，minsup によって精度が異なる．原理上は，minsup を下げることで，得られる部分グラフパターンの集合は単調に増大するので，用いる機械学習アルゴリズムが適切に対応できれば，精度も単調に向上する．しかし，minsup を下げれば出力数 (用いる特徴の数) は指数的に増加するため，計算時間や使用メモリの点で現実的には全列挙が困難となる．また maxpat も大きくするほど得られる部分グラフパターンの集合は増大する．本研究では，maxpat=7 に対して minsup=40，30，20，15，10 および，minsup=1 に対して maxpat=4, 5 で列挙される頻出部分グラフおよび辺数制約部分グラフの有無を特徴として用いた．

3.2 gBoost

gBoost アルゴリズム [Saigo 09] は線形計画問題で定式化した LPBoost[Demiriz 02] とグラフデータ上の特徴探索を組み合わせた手法である．特徴量としてすべての部分グラフ特徴の有無が使用され，弱学習器として Decision Stump(深さ 1 の決定木) が用いられる．列生成法により特徴探索を行い全部分グラフ特徴のうち分類に必要となる特徴のみを使用するため，すべての部分グラフの有無を調べずに済むことが特徴となっている．

4. 予測モデル

本研究では，予測モデルとしてロジスティック回帰，サポートベクトルマシン，k-近傍法，Random Forest[Breiman 01]，XGBoost[Chen 16] を使用した．サポートベクトルマシンに関してはカーネルとして Linear と RBF を使用しており，ロジスティック回帰とサポートベクトルマシンの Linear は線形モデルとなっている．Random Forest と XGBoost は弱学習器として決定木を使用したモデルとなっている．

5. 実験

化合物の活性を予測する問題に対して 4 種類の特徴量と 5 種類の予測モデルのそれぞれの組み合わせに加えて gBoost を比較するために実験を行った．特徴量としては創薬分野から ECFP2 種と DRAGON 記述子の 3 種類，情報科学分野から頻出部分グラフと辺数制約部分グラフを使用した．部分グラフは maxpat=7 に対して minsup=40，30，20，15，10 および，minsup=1 に対して maxpat=4, 5 で列挙される頻出部分グラフおよび辺数制約部分グラフの 7 種類である．予測モデルはそれぞれの特徴ベクトルに対してロジスティック回帰 (LR)，SVM，k-近傍法 (kNN)，Random Forest，XGBoost を使用し，加えて gBoost を用いて学習を行った．予測モデルのハイパーパラメータについては与えた範囲の値の組み合わせをすべて計算し正答率の最良値を結果とした．SVM においては正則化係数である $C = [0.1, 1, 10, 100]$，カーネル法の種類は linear と RBF，RBF カーネルのハイパーパラメータとして $\gamma = [0.1, 0.01]$ を使用した．k-近傍法では 近傍として k = [1,3,5,7,9] を使用した．Random Forest では，決定木の数として n_estimators = [10,50,100,300]，決定木の作成に使用する特徴量の数として max_features = [10,20,30,40] をパラメータとして使用した．以上のモデルは機械学習ライブラリ scikit-learn を用いて実験を行った．明示していないパラメータについてはデフォルト値を使用した．XGBoost では 4 種類のパラメータを使用した．木の深さの最大値の max_depth = [3,5,10]，各木においてランダムに抽出される標本の割合を示す subsample = [0.5,1] と特徴の割合を示す colsample_bytree = [0.5,0.7,1]，そしてデータの重み付けの合計値の最小値である min_child_weight = [1,3,5] である．gBoost においては特徴として考える部分グラフの大きさを示す maxpat = 0.8,0.6，正則化係数である $\nu = [8,6,4,3,0.9]$ で実験を行った．

データ	グラフ数	平均 ノード数	平均 エッジ数	DRAGON 記述子数
NCI1	4252	26.3985	28.4860	3818
NCI41	3292	26.8404	28.9958	3818
NCI47	4202	26.3660	28.4557	3818
NCI81	5006	26.3865	28.5756	3818
NCI83	4766	26.8767	29.0289	3818
NCI123	6474	26.4203	28.5201	3818
NCI145	4076	26.4172	28.5063	3818
NCI220	578	25.4605	27.1292	3818
NCI330	4920	22.3108	23.7577	3818
Mutag	167	17.9305	19.7968	3837
CPDB	601	13.7750	14.2217	3837

表 1: 使用したデータセット

5.1 データセット

本実験は表 1 に示す 13 種類のグラフデータを用いて，行った．データセット名「NCIx」は PubChem BioAssay の AID 番号 x のデータセットである [*2]．これらのデータセットのうちグラフ数が膨大なものは，正例の数に合わせて負例をランダムにサンプリングした．構造活性相関の実験の実データでは通常，負例数が極めて多く，この不均衡性 (imbalance) を考慮した評価尺度を用いる必要がある．本稿では単純に正例と負例の数を均衡化し，通常の正解率により特徴集合の良さを評価する．

5.2 実験結果と考察

本実験において，正解率はデータを 10-分割交差検証を行った際のテストデータに対する正解率を示す．正解率の結果を表 2，表 3 に示した．また各データにおける最高値を太字で示

*2 http://www.ncbi.nlm.nih.gov/pcassay

した．

まず各特徴量について見ていく．NCI データにおいては RDKit によって作成されたフィンガープリント，Mutag と CPDB においては DRAGON によって作成されたフィンガープリントで行った予測の精度が良いことがわかる．両者の違いとしてはまずはデータの特性がある．NCI データは抗がん作用に関するものでデータ数が 3000 から 6500 と多く，化合物自体も大きいデータセットである．一方 Mutag と CPDB のデータセットは突然変異誘発性に関するものでデータ数が Mutag 167 個，CPDB 601 個と少なく，化合物の大きさも NCI データと比べて小さいデータセットである．また RDKit によるフィンガープリントと DRAGON によるフィンガープリントでは付加している化学情報が異なる．これらの違いから最高値を示す特徴量の違いが生まれたのではないかと推測される．

次に各手法について見ていく．すると NCI データに関しては Random Forest が，Mutag と CPDB では XGBoost の精度が良いことがわかる．ここでもデータの特性に対する違いが現れている．今回の実験ではパラメータを細かくは調整しなかったが，細かく見ていくとパラメータの影響を受けやすいモデルの精度も変わってくるのではないかと思われる．

線形の予測モデルと非線形の予測モデルを比べるとフィンガープリントにおいては精度に大きく差が出ることはないのに対し，DRAGON の記述子を特徴ベクトルとしたときは線形モデルの精度が下がる．ここでフィンガープリントは 0,1 表現の特徴ベクトルであるのに対し，DRAGON の記述子による特徴ベクトルは主に実数値を取る．このことから実数値の特徴ベクトルは非線形性が強いと思われる．DRAGON の記述子による特徴ベクトルでは Random Forest や XGBoost よりも RBF カーネルを用いた SVM のほうが精度が出ているが，フィンガープリントにおいてあまり違いは見られない．Random Forest や XGBoost といった木構造を扱う予測モデルは 0,1 表現の特徴量を扱いに適していると考えられる．

また gBoost は創薬分野の特徴量を用いた線形手法と比べると精度は出るが，非線形の手法と比べるとやや精度が落ちる．gBoost は弱学習器に Decision Stump を使用している．この Decision Stump は実数値を取るデータに関しては非線形となるが部分グラフの有無のような 0,1 データに対しては線形のモデルとなっている．このことが非線形の手法に比べて精度が落ちたことに関係していると考えられる．よって部分グラフの有無に関して非線形な手法を構築することによって精度の向上が期待できる．

頻出部分グラフ・辺数制約部分グラフを特徴量とした時 minsup=1 で maxpat5 の時が一番よい精度が得られた．また頻出部分グラフにおいての予測モデルについてみたとき Random Forest が一番良い精度となっている．スペースの都合上，表 3 に線形モデル・非線形モデルのうち精度の良い SVM(Linear) および Random Forest の結果を示す．また LR や SVM(Linear) といったような線形手法と比べて非線形な手法のほうが精度は高い．この傾向は ECFP を用いた場合と共通しており，一般に低分子化合物の活性予測において部分構造の有無を特徴とした場合，データは非線形性の強い分布となっていると考えられる．また，maxpat=7 として，minsup を変化させた場合の結果（表 3）より，頻出パターンのみを特徴量とすると情報が十分ではないことが示唆される．この点は既存研究 [Wale 08] でも指摘されている．

6. 結論と今後の課題

本稿では化合物から活性を予測する問題に対する創薬分野と情報科学分野それぞれの手法について比較検討した．結果として化学情報を用いて予測を行う創薬分野とグラフに落とし込んで予測を行う情報科学分野において，精度の差としては少し創薬分野の方がよいという結果となった．今後の課題としては，まず情報科学分野の手法で用いた分子グラフでは，創薬分野の ECFP で考慮されているような化学情報を考慮していないため，同様の分子グラフ表現を用いた場合の精度を検討したい．また，今回は実験の対象や実験パラメタの組合せ数が大きいため，各パラメタの詳細なチューニングを検討したい．また今回情報科学分野における手法として gBoost と頻出部分グラフを特徴に用いたモデルを使用したが，この他にもグラフカーネル法のようなモデルを追加することも検討したい．

7. 謝辞

本研究は JSPS 科研費 26330242,16K13852 および JST さきがけの助成を受けたものです．

参考文献

[Breiman 01] Breiman, L.: Random Forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32 (2001)

[Chen 16] Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 785–794 (2016)

[Demiriz 02] Demiriz, A., Bennett, K. P., and Shawe-Taylor, J.: Linear Programming Boosting via Column Generation, *Machine Learning*, Vol. 46, No. 1-3, pp. 225–254 (2002)

[Rogers 10] Rogers, D. and Hahn, M.: Extended-Connectivity Fingerprints, *Journal of Chemical Information and Modeling*, Vol. 50, No. 5, pp. 742–754 (2010)

[Saigo 09] Saigo, H., Nowozin, S., Kadowaki, T., Kudo, T., and Tsuda, K.: gBoost: a mathematical programming approach to graph classification and regression, *Machine Learning*, Vol. 75, No. 1, pp. 69–89 (2009)

[Wale 08] Wale, N., Watson, I. A., and Karypis, G.: Comparison of descriptor spaces for chemical compound retrieval and classification, *Knowledge and Information Systems*, Vol. 14, No. 3, pp. 347–375 (2008)

[Yan 02] Yan, X. and Han, J.: gSpan: Graph-Based Substructure Pattern Mining, in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, pp. 721–724 (2002)

	SVM Linear(minsup 1)		SVM Linear(maxpat 7)			RF(minsup 1)		RF(maxpat 7)		
Data	maxpat4	maxpat5	minsup40	minsup20	minsup10	maxpat4	maxpat5	minsup40	minsup20	minsup10
NCI1	0.832	0.847	0.638	0.718	0.779	0.79	0.882	0.732	0.810	0.838
NCI41	0.772	0.806	0.698	0.765	0.791	0.834	0.845	0.771	0.823	0.838
NCI47	0.787	0.806	0.694	0.769	0.795	0.841	0.848	0.786	0.822	0.838
NCI81	0.788	0.818	0.687	0.765	0.801	0.839	0.850	0.776	0.827	0.837
NCI83	0.751	0.782	0.669	0.739	0.767	0.806	0.815	0.731	0.791	0.80
NCI123	0.745	0.768	0.654	0.729	0.753	0.803	0.808	0.732	0.783	0.795
NCI145	0.794	0.820	0.689	0.782	0.807	0.849	0.857	0.794	0.836	0.846
NCI220	0.529	0.546	0.576	0.510	0.529	0.510	0.531	0.541	0.529	0.515
NCI330	0.832	0.847	0.638	0.718	0.779	0.879	0.882	0.732	0.810	0.838
Mutag	0.850	0.855	0.818	0.802	0.845	0.850	0.829	0.834	0.818	0.861
CPDB	0.785	0.765	0.601	0.683	0.731	0.786	0.788	0.655	0.701	0.743

表 2: 部分グラフを用いた際の正答率

	Finger Print(RDKit)						Finger Print(DRAGON)						DRAGON						gBoost
Data	LR	SVM (Linear)	SVM (RBF)	kNN	RF	XGBoost	LR	SVM (Linear)	SVM (RBF)	kNN	RF	XGBoost	LR	SVM (Linear)	SVM (RBF)	kNN	RF	XGBoost	gBoost
NCI1	0.812	0.836	0.866	0.816	**0.868**	0.840	0.776	0.786	0.858	0.836	0.850	0.848	0.697	Time Out	0.866	0.679	0.837	0.846	0.831
NCI41	0.809	0.822	0.852	0.815	**0.865**	0.830	0.741	0.749	0.837	0.817	0.842	0.832	0.690	Time Out	0.852	0.662	0.827	0.837	0.819
NCI47	0.809	0.817	0.851	0.811	**0.856**	0.828	0.768	0.777	0.845	0.834	0.842	0.838	0.697	Time Out	0.851	0.678	0.832	0.849	0.797
NCI81	0.808	0.809	0.846	0.817	**0.856**	0.836	0.774	0.776	0.845	0.800	0.837	0.829	0.694	Time Out	0.846	0.663	0.827	0.844	0.809
NCI83	0.774	0.778	0.815	0.768	**0.825**	0.804	0.735	0.753	0.818	0.800	0.815	0.812	0.671	Time Out	0.815	0.649	0.797	0.810	0.774
NCI123	0.777	0.777	0.816	0.773	**0.818**	0.789	0.734	0.743	0.800	0.781	0.805	0.796	0.673	Time Out	0.520	0.643	0.786	0.799	0.765
NCI145	0.823	0.832	**0.869**	0.826	0.867	0.837	0.777	0.798	0.853	0.832	0.851	0.844	0.716	Time Out	0.519	0.683	0.836	0.845	0.827
NCI220	0.517	0.529	0.554	0.519	0.555	0.545	0.540	0.559	0.559	0.546	0.533	0.555	0.538	Time Out	0.507	0.562	0.543	0.557	**0.579**
NCI330	0.847	0.855	0.882	0.839	**0.888**	0.863	0.812	0.820	0.883	0.858	0.882	0.875	0.759	Time Out	0.544	0.710	0.862	0.882	0.845
Mutag	0.876	0.881	0.897	0.864	0.871	0.883	0.872	0.873	0.663	0.882	0.898	**0.910**	0.888	0.873	0.663	0.882	0.888	0.901	0.840
CPDB	0.760	0.881	0.897	0.864	0.871	0.787	0.745	0.873	0.663	0.882	0.875	**0.909**	0.745	0.873	0.663	0.622	0.870	0.890	0.778
Ave	0.783	0.802	0.831	0.792	0.804	0.831	0.752	0.775	0.803	0.806	0.821	0.823	0.706	0.000	0.695	0.676	0.810	0.824	0.708
SD	0.090	0.092	0.092	0.091	0.089	0.086	0.077	0.088	0.096	0.088	0.095	0.091	0.079	0.000	0.147	0.075	0.089	0.089	0.095

表 3: 正答率