Vol.2016-MPS-108 No.26
Vol.2016-BIO-46 No.26
2016/7/5

活性値情報のグループ化とランク学習による化合物スクリーニング

鈴木 翔吾[1,4,a)]　大上 雅史[2,3,b)]　秋山 泰[2,3,4,c)]

概要：新薬開発の支援手法に，まだ実験されていない化合物の活性値を機械学習等の技術によって予測する化合物スクリーニングがある．化合物スクリーニングのうち，活性値そのものではなく活性値の大小の順序関係を予測する問題としてランク学習で扱う手法が提案されている．既存手法では生化学実験データに含まれる化合物間の全ての順序関係について学習・評価を行っているが，非活性化合物間の順序関係や活性値が類似した化合物間の順序関係といった生化学実験上の誤差による無意味な順序関係も考慮しているという問題がある．そこで本研究では，活性値のグループ化によってこれらの無意味な順序関係を無視し，ランク学習に適用する手法を提案する．評価実験には PubChem BioAssay に登録されている Luciferase，Scp-1，sEH を標的とした 3 つのハイスループットスクリーニングデータを用いた．評価実験の結果，活性値のグループ化とランク学習を組み合わせて学習する提案手法によって，グループ化をせずに学習する既存手法よりも予測精度が有意に向上することを確認した．

キーワード：化合物スクリーニング，機械学習，ランク学習

A Compound Screening with Grouping Activity Score and Learning to Rank

Shogo D. Suzuki[1,4,a)]　Masahito Ohue[2,3,b)]　Yutaka Akiyama[2,3,4,c)]

Abstract: Virtual screening which predicts activity of untested compounds by machine learning techniques is widely used in the process of a drug discovery research. Recent studies have proposed some methods of virtual screening with learning to rank with the order of activity scores (e.g. %Inhibition, IC_{50}, etc.). However, these methods have a problem that nonsense order (i.e. the order between inactive compounds and that between compounds which have a similar activity score) is considered. In this study, we present a new method to ignore nonsense order with grouping activity score and apply to learning to rank. With three high-throughput screening data (each of the target is Luciferase, Scp-1 and sEH) registered in PubChem BioAssay, we confirmed that our methods of grouping and learning to rank significantly improves prediction accuracy compared with the existing method of without grouping.

Keywords: Compound Screening, Machine Learning, Learning to Rank

1. 導入

新薬の開発には，有効性や安全性を試験するために 1 つ

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[2] 東京工業大学 情報理工学院 情報工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology
[3] 東京工業大学 科学技術創成研究院 スマート創薬研究ユニット
Advanced Computational Drug Discovery Unit, Institute of Innovative Research, Tokyo Institute of Technology
[4] 東京工業大学 情報生命博士教育院
Education Academy of Computational Life Sciences, Tokyo Institute of Technology
[a)] s_suzuki@bi.cs.titech.ac.jp
[b)] ohue@c.titech.ac.jp
[c)] akiyama@c.titech.ac.jp

の医薬品あたり十数年といった年月と約一千億円の費用が必要である [1]．新薬開発の初期段階では，創薬ターゲットに有効な活性を持つ化合物を化合物ライブラリーから探索する化合物スクリーニングが行われる．しかし，化合物ライブラリーに含まれる化合物は数が多く，生化学実験による化合物スクリーニングはきわめて大きなコストがかかる．そのため，計算機を用いて化合物の活性を予測するバーチャルスクリーニングという技術が用いられている．バーチャルスクリーニングによる活性予測によって生化学実験にかける順番を決定することで，新薬開発に必要なコストの削減が期待できる．

バーチャルスクリーニングの手法の 1 つとして，創薬ターゲットに対して既に実験された化合物の活性情報を用いるリガンドベースの手法があり，様々な機械学習技術が用いられている [2]．リガンドベースの手法によるバーチャルスクリーニングは，創薬ターゲットに対する活性の有無を予測する分類問題や，活性値そのものを予測する回帰問題として従来扱われてきた．一方で，活性値そのものではなく活性値の大小の順序関係を予測する問題として扱い，学習手法としてランク学習を用いる手法が近年提案されている [3–5]．分類問題ではなく順序を予測する問題として扱うことで，創薬ターゲットに対して活性を持つ化合物間の活性強弱を扱えるというメリットがある．また，活性強弱を絶対的な値でなく相対的に扱うことで，生化学実験によって得られた活性値に含まれる誤差の影響を低減することができる．

ランク学習を初めてバーチャルスクリーニングに導入したのは S. Agarwal ら [3] である．S. Agarwal らは Support Vector Machine (SVM)，Support Vector Regression (SVR) およびランク学習手法の一つである RankSVM [6] の 3 つの手法による予測精度を比較し，RankSVM による予測精度が最も高かったと報告している．F. Rathke ら [4] はランキング上位を正しくランク付けすることを重視する StructRank という手法を提案し，SVR と RankSVM を比較して予測精度の向上を報告している．W. Zhang ら [5] は複数のアッセイデータに対し様々なランク学習手法を適用し予測精度を比較した．さらに，標的タンパク質に対するアッセイデータだけでなく他のタンパク質に対するアッセイデータも学習に用いることで，予測精度をさらに向上させることができたと報告している [5]．

ランク学習を用いたバーチャルスクリーニングの先行研究では，阻害率や IC_{50} といった活性値の大小の順序関係を用いていた．一方で，活性値の大小関係の中には学習・評価の上で大きな意味を持たない 2 つの順序関係が含まれていると考えられる．

(1) 非活性化合物間の順序
標的タンパク質に対して活性を持たない化合物にも生化学実験によって得られた活性値が報告されるが，活性を持たない化合物の活性値はきわめて誤差が大きく，本質的な意味は無い．

(2) 活性値が類似した化合物間の順序
活性値には生化学実験による誤差が含まれる．そのため，活性値が近い化合物間の順序は容易に入れ替わる．

先行研究ではこれらの順序関係も含めて学習・評価しているという問題がある．

そこで，本研究では活性値のグループ化によって，無意味な大小関係を無視し，ランク学習に適用する手法を提案する．つまり，各化合物に紐づく活性値の大小関係ではなく，各化合物が属するグループに紐づくグループ評価値の大小関係をランク学習に適用する．

グループ化の際には，非活性化合物は全て同じグループに属するように，活性化合物は活性値が類似しているものは同じグループに属するようにグループ化を行った．そして，活性値の高い順に各グループにグループ評価値を割り振り，グループ評価値の大小関係をランク学習に適用した．上述の本研究の概要を**図 1** に示す．

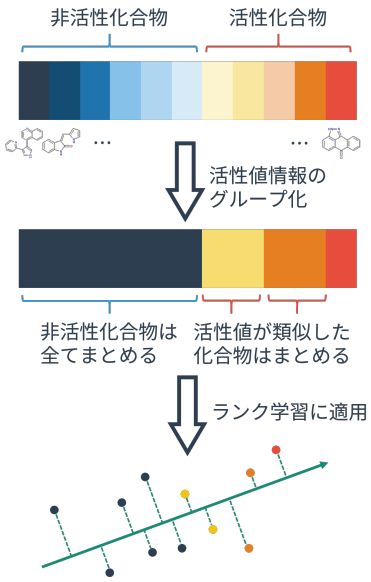

図 1 本研究の概要

2. 手法

2.1 ランク学習

ランク学習は情報検索，特にウェブ検索の分野で広く用いられている機械学習手法である．ランク学習は特徴ベクトルの集合を入力とし各特徴ベクトルに対応する順位を出力する [7]．ランク学習における訓練データは，n 個のクエリ q_i $(i = 1, \ldots, n)$ とそれに紐づく要素の集合 $\mathbf{x}^{(i)} = \{x_j^{(i)}\}_{j=1}^{m^{(i)}}$ ($m^{(i)}$ はクエリ q_i に紐づく要素の個数)

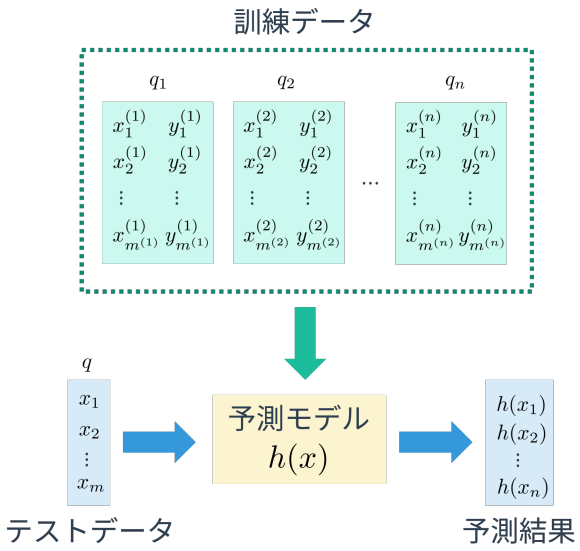

図 2 ランク学習の枠組み

と各要素の関連度 $\mathbf{y}^{(i)} = \{y_j^{(i)}\}_{j=1}^{m^{(i)}}$ で表される．与えられた訓練データを学習することで予測モデル $h(x)$ を得て，新たな入力 $(q, \mathbf{x})$ に対して各要素の予測関連度 $\{h(x_j)\}_{j=1}^{m}$ を出力する．上述のランク学習の枠組みを**図 2** に示す．

W. Zhang ら [5] は，いくつかのランク学習手法をバーチャルスクリーニングに適用し予測精度の比較を行い，RankSVM [6] による予測精度が最も高かったと報告している．そこで本研究ではランク学習の手法として RankSVM を用いた．RankSVM は様々な実装が存在するが，本研究では C. P. Lee ら [8] によって実装されたものを用いた．C. P. Lee らの実装は，線形カーネルを用いた RankSVM の解がその他の実装よりも高速に収束するものである．

2.2 活性値のグループ化

先行研究では各化合物の評価値として標的タンパク質に対する活性値を用いていたが，非活性化合物間の順序関係や活性値が類似した化合物間の順序関係も考慮してしまうという問題がある．標的タンパク質に対して活性を持たない化合物に対して生化学実験によって得られる活性値は無意味な値であり，生化学実験による誤差の影響を考えると活性値が非常に近い化合物間の順序関係も無意味なものである．本研究ではこれらの無意味な順序関係をグループ化によって無視することで解決する．つまり，化合物の評価値 y として活性値 a を用いるのではなく，グループ評価値 $g(a)$ を用いてランク学習に適用する．

本研究で用いた生化学実験データに含まれる各化合物には，実験者によって活性が認められるかどうかのラベルが付与されている．生化学実験の誤差は活性化合物よりも非活性化合物の方が大きいと考えられるため，それぞれ異なるグループ化の基準が必要である．以下に非活性化合物のグループ化および活性化合物のグループ化の方法について述べる．

2.2.1 非活性化合物のグループ化

実験者によって非活性だと判断された化合物群について，非活性化合物間の順序関係を全て無視するため，非活性化合物は全て同一のグループとしてまとめる．つまり，非活性化合物の活性値 $a^{(-)}$ に対し，グループ評価値 $g(a^{(-)}) = 0$ として定義する．非活性化合物のグループ化の様子を**図 3** に示す．

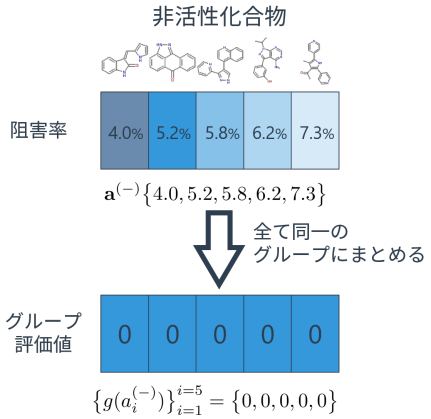

図 3 非活性化合物のグループ化の様子

2.2.2 活性化合物のグループ化

活性化合物の中で活性値の類似した化合物をグループ化するために，本研究では階層型クラスタリングによってグループ化を行った．具体的には，データセットに含まれる活性化合物の活性値を並べたベクトル $\mathbf{a}^{(+)}$ に対し，活性値間のユークリッド距離を用いた Ward 法による階層型クラスタリングを適用する．クラスタ数については，生化学実験の精度を考慮しながらユーザーが指定する．こうして得られたグループには順序が付いていることから，各グループに含まれる化合物についてグループ評価値 $g(a^{(+)})$ を 1 から順に振っていく．活性化合物のグループ化の例を**図 4** に示す．

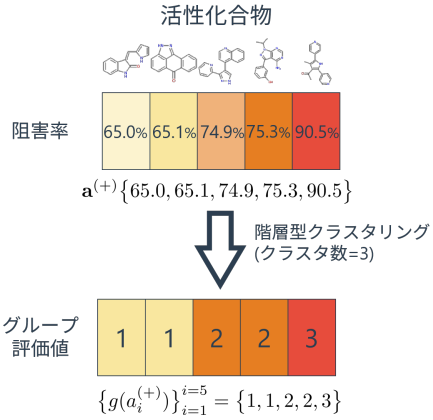

図 4 阻害率が {65.0%, 65.1%, 74.9%, 75.3%, 90.5%} として与えられた活性化合物をクラスタ数 3 でグループ化する例．

3. 評価実験

3.1 データセット

本研究では PubChem BioAssay [9] に含まれている，Luciferase, Scp-1, sEH をそれぞれ標的タンパク質としたハイスループットスクリーニングのデータを用いて実験を行った．Luciferase を標的タンパク質としたハイスループットスクリーニングのデータは Q. Li ら [10] が，Scp-1 および sEH を標的タンパク質としたハイスループットスクリーニングのデータは M. Lindh ら [11] がバーチャルスクリーニングの研究に用いていたものである．これらのハイスループットスクリーニングのデータは，化合物をある濃度で固定した際の阻害率が報告されており，実験者によって各化合物に活性有無のラベルが付与されている．本研究で用いたデータセットの概要を**表 1** に示す．

特徴量には RDKit [12] で計算される Extended-connectivity fingerprints (ECFP) [13] を用いた．ECFP は化合物の部分構造の有無をビット列で表現したものであり，化合物を機械学習で扱う際の特徴量として広く用いられている．ECFP にはパラメータとして化合物をグラフとして見た際の部分構造の直径および次元数があるが，本研究では直径を 4，次元数を 2048 とした ECFP を用いた．

3.2 評価基準

本研究では予測精度を 5-fold Cross Validation によって評価した．予測精度の評価基準としては, Discounted Cumulative Gain (DCG) に基づく NDCG@100 および NDCG@All を用いた．

DCG とは関連度付きランキング出力結果の評価に用いられる評価指標であり，真のランキングに近い予測ほど高い値となる [14]．関連度を並べたベクトルを $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ とすると，DCG は式 (1) で与えられる．

$$\mathrm{DCG@}k = y_1 + \sum_{i=2}^{k} \frac{y_i}{\log_2(i)} \tag{1}$$

ここで，k は順位付けを行う個数である．

Normalized DCG (NDCG) とは関連度の順に正しく順位付けを行った際に得られる理想的な DCG で予測順位から得られる DCG を正規化した値であり，式 (2) で表される．

$$\mathrm{NDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{idealDCG}} \tag{2}$$

NDCG は 0 から 1 の範囲の値を取り，予測順位が真の順位と一致するとき 1 となる．DCG はデータセットによって取りうる値の範囲が異なるが NDCG は 0 から 1 の値を取るため，ランク学習を用いた研究の予測精度の評価には NDCG が広く用いられている．

3.3 ハイパーパラメータ探索

SVM と同様に RankSVM もカーネルを指定することができるが，本研究では線形カーネルの RankSVM を用いた．線形カーネルの RankSVM にはコストパラメータ C がハイパーパラメータとして存在する．本研究では 5-fold Cross Validation によって得られる予測精度の平均値が最も高くなるようなコストパラメータ C を，$\{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$ から探索した．

4. 実験結果

4.1 非活性化合物のグループ化の効果

まず，訓練データに含まれる非活性化合物のグループ化による予測精度の変化について検証する．データセットを 5-fold Cross Validation で訓練データとテストデータに分割した後，訓練データは「グループ化なし (No Grouping)」および「非活性化合物のみグループ化 (**Inactive Grouping**)」，テストデータは「非活性化合物のみグループ化 (**Inactive Grouping**)」としたときの予測精度を評価する．この実験の概要を**図 5** に，得られた予測精度を**表 2** に示す．

4.2 活性値の類似した化合物のグループ化の効果

次に，訓練データに含まれる活性値の類似した化合物のグループ化による予測精度の変化について検証する．データセットを 5-fold Cross Validation で訓練データとテストデータに分割した後，訓練データは「非活性化合物のみグループ化 (**Inactive Grouping**)」および「非活性化合物＋活性値の類似した化合物グループ化 (**All Grouping**)」，テストデータは「非活性化合物＋活性値の類似した化合物グループ化 (**All Grouping**)」としたときの予測精度を評価する．訓練データに含まれる活性化合物のグループの個数として $\{10, 100\}$ を実験した．テストデータに含まれる活性化合物のグループ化は次の手順で行った．

(1) 訓練データに含まれる活性化合物について階層型クラスタリングを行い，各グループの活性値の範囲を求める．
(2) 上で求められた各グループの活性値の範囲を用いて，テストデータに含まれる各活性化合物についてそれぞれ属するグループを決定する．

この実験の概要を**図 6** に，得られた予測精度を**表 3** および**表 4** に示す．

表 1 本研究で用いたデータセットの概要

Target	PubChem AID	#Active	#Inactive	Activity
Luciferase	1006	2,976	192,588	%Inhibition at 10 μM
Scp-1	493091	2,984	337,701	%Inhibition at 20 μM
sEH	707	310	90,339	%Inhibition at 2 μM

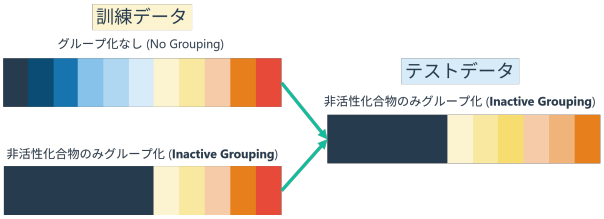

図 5 非活性化合物グループ化の効果の検証実験の概要

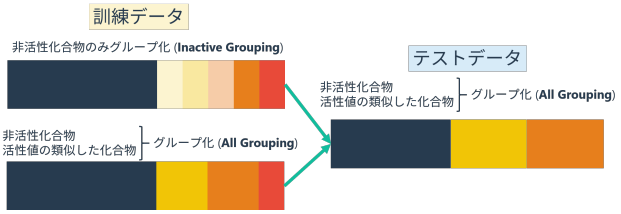

図 6 活性値の類似した化合物のグループ化の効果の検証実験の概要

表 2 テストデータ：「非活性化合物のみグループ化」に対する予測精度．No Grouping はグループ化なしを，**Inactive Grouping** は非活性化合物のみグループ化したことを表す．No Grouping と **Inactive Grouping** について対応のある t 検定によって得られた P 値を括弧内に示し，$P < 0.05$ となった項目を*をつけて表す．

ターゲット	手法	NDCG@100	NDCG@All
Luciferase	No Grouping	0.070	0.700
	Inactive Grouping	0.132	0.755
		$*(6.48 \times 10^{-3})$	$*(1.02 \times 10^{-3})$
Scp-1	No Grouping	0.041	0.571
	Inactive Grouping	0.090	0.644
		$*(2.36 \times 10^{-3})$	$*(2.30 \times 10^{-5})$
sEH	No Grouping	0.332	0.605
	Inactive Grouping	0.560	0.771
		$*(7.77 \times 10^{-3})$	$*(4.30 \times 10^{-3})$

表 3 テストデータ：「非活性化合物＋活性値の類似した化合物グループ化」に対する予測精度．訓練データに含まれる活性化合物のグループの個数を 10 とした．**Inactive Grouping** は非活性化合物のみグループ化したことを，**All Grouping** は非活性化合物および活性値の類似した化合物をグループ化したことを表す．**Inactive Grouping** と **All Grouping** について対応のある t 検定によって得られた P 値を括弧内に示し，$P < 0.05$ となった項目を*をつけて表す．

ターゲット	手法	NDCG@100	NDCG@All
Luciferase	**Inactive Grouping**	0.149	0.722
	All Grouping	0.150	0.723
		(3.35×10^{-1})	(6.70×10^{-2})
Scp-1	**Inactive Grouping**	0.090	0.605
	All Grouping	0.090	0.606
		(2.75×10^{-1})	$*(3.06 \times 10^{-3})$
sEH	**Inactive Grouping**	0.558	0.721
	All Grouping	0.556	0.721
		(1.96×10^{-1})	(4.74×10^{-1})

表 4 テストデータ：「非活性化合物＋活性値の類似した化合物のグループ化」に対する予測精度．訓練データに含まれる活性化合物のグループの個数を 100 とした．**Inactive Grouping** は非活性化合物のみグループ化したことを，**All Grouping** は非活性化合物および活性値の類似した化合物をグループ化したことを表す．**Inactive Grouping** と **All Grouping** について対応のある t 検定によって得られた P 値を括弧内に示し，$P < 0.05$ となった項目を*をつけて表す．

ターゲット	手法	NDCG@100	NDCG@All
Luciferase	**Inactive Grouping**	0.151	0.719
	All Grouping	0.151	0.719
		(4.69×10^{-1})	(1.03×10^{-1})
Scp-1	**Inactive Grouping**	0.091	0.596
	All Grouping	0.091	0.596
		(4.20×10^{-1})	$*(1.70 \times 10^{-2})$
sEH	**Inactive Grouping**	0.559	0.718
	All Grouping	0.559	0.718
		(1.97×10^{-1})	(9.72×10^{-1})

Vol.2016-MPS-108 No.26
Vol.2016-BIO-46 No.26
2016/7/5

5. 考察

5.1 非活性化合物のグループ化の効果

表 2 によると，3 つのデータセット全てにおいて非活性化合物をグループ化することで NDCG@100 および NDCG@All が向上していることがわかる．これは，グループ化をせずに学習して得られる解は無意味な非活性化合物間の順序も正しく並ぶように学習されて得られるものであるため，汎化性能が低下しているからだと考えられる．このことを**図 7** を用いて直感的に説明する．

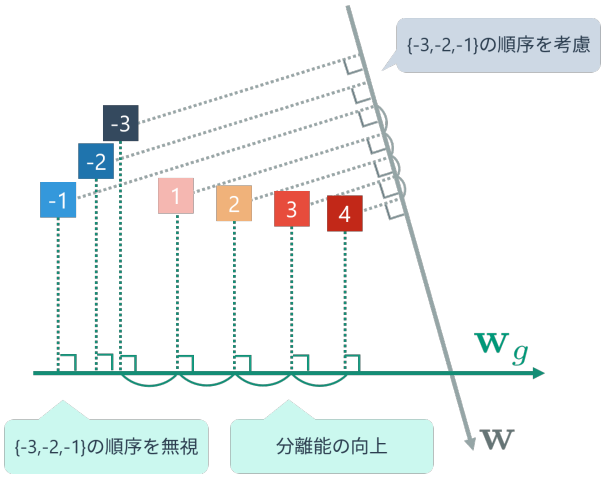

図 7 非活性化合物のグループ化による汎化性能の向上．赤色で示した点 $\{1, 2, 3, 4\}$ は活性化合物を，青色で示した点 $\{-3, -2, -1\}$ は非活性化合物を表している．青点の順序を無視して得られた解 $\mathbf{w}_g$ は，青点の順序を考慮して得られた解 $\mathbf{w}$ よりも，射影後の赤点間の間隔および赤点と青点の間隔が大きくなっている．

図 7 において，$\mathbf{w}$ は $\{-3, -2, -1, 1, 2, 3, 4\}$ の順に並ぶように学習して得られた解を，$\mathbf{w}_g$ は青点 $\{-3, -2, -1\}$ の順序を無視し学習して得られた解を表している．各点を $\mathbf{w}$ および $\mathbf{w}_g$ へ射影すると，$\mathbf{w}_g$ に射影した方が赤点間の間隔および赤点と青点の間隔が大きくなることがわかる．これは，青点の順序を無視すれば，射影後の赤点間の間隔および赤点と青点の間隔がより大きくなるような解が存在するためである．赤点を活性化合物，青点を非活性化合物と捉えると，活性化合物間の射影後の距離および活性化合物と非活性化合物の射影後の距離が大きくなることから，汎化性能が向上したのだと考えられる．

5.2 活性値の類似した化合物のグループ化の効果

表 3 および表 4 によると，非活性化合物のみグループ化をした場合と，非活性化合物＋活性値の類似した化合物のグループ化をした場合とで NDCG@100 および NDCG@All の差はほとんど見られなかった．本研究では 2048 次元という比較的次元数の大きい特徴量を用いたため RankSVM のカーネルには非線形カーネルではなく線形カーネルを用いたが，非線形カーネルと比べて線形カーネルを用いた場合は訓練データにアンダーフィッティングする傾向がある．そのため，活性化合物に含まれる生化学実験由来の誤差の影響をあまり受けなかったことから，予測精度の差がほとんど見られなかったのではないかと考えられる．

5.3 階層型クラスタリングにおけるグループ数

本研究では活性値の類似した化合物のグループ化の際に，グループの個数を $\{10, 100\}$ として実験を行った．グループの個数を小さくすると生化学実験の誤差の影響を小さくすることができるが，活性化合物間の順序関係を無視することになる．そのため，生化学実験による誤差がどの程度含まれているのかを確認しながらグループの個数を決定する必要がある．

6. まとめ

6.1 結論

先行研究では活性値そのものの大小の順序関係を用いてランク学習によるバーチャルスクリーニングを行っていたが，本研究では活性値をグループ化することで無意味な順序関係を無視しランク学習に適用する手法を提案した．無意味な順序関係として，(1) 非活性化合物間の順序と (2) 活性値の類似した化合物間の順序の 2 種類を考慮した．非活性化合物は全て一つのグループとしてまとめ，活性値の類似した化合物は階層型クラスタリングを用いてグループ化を行った．NDCG@100 および NDCG@All を用いて予測精度の評価をした結果，活性値のグループ化による予測精度の有意な向上が確認できた．一方で，「非活性化合物のグループ化 (**Inactive Grouping**)」と「非活性化合物＋活性値の類似した化合物のグループ化 (**All Grouping**)」の予測精度に差は見られなかったが，より生化学実験の誤差が大きいデータセットを用いる場合は，活性値の類似した化合物のグループ化による効果が現れると考えられる．

6.2 今後の課題

6.2.1 非線形カーネルを用いた RankSVM の適用

本研究では活性値の類似した化合物のグループ化の効果は確認できなかった．これは，線形カーネルを用いた RankSVM は訓練データにアンダーフィッティングし，活性化合物に含まれる誤差の影響をあまり受けなかったのではないかと考えられる．誤差の影響が大きくなるような非線形カーネルを用いた場合では，活性値の類似した化合物のグループ化による効果が大きくなるのではないかと予想される．非線形カーネルを用いた RankSVM を適用し，活性値の類似した化合物のグループ化の効果を確認したい．

6.2.2 生化学実験の誤差の大小を考慮したグループ化

本研究では活性値の類似した化合物のグループ化を階層型クラスタリングによって行った．これは，活性化合物は

全て同等の生化学実験による誤差を含むという状況を想定している．しかし，創薬ターゲットに強く結合する化合物の実験誤差は小さくなることや，弱く結合する化合物の実験誤差は大きくなることが予想される．そのため，活性値の大小に応じてグループ化の基準を変更する手法が求められる．

謝辞 本研究の一部は JSPS 科研費 基盤研究 (A) (24240044)，および JST CREST 「EBD: 次世代の年ヨッタバイト処理に向けたエクストリームビッグデータの基盤技術」の支援を受けて行われた．

参考文献

[1] A. Mullard, New drugs cost US$2.6 billion to develop., *Nature reviews. Drug Discovery*, 13(12), 877, 2014.

[2] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications., *Drug Discovery Today*, 20(3), 318–331, 2015.

[3] S. Agarwal, *et al.*, Ranking Chemical Structures for Drug Discovery : A New Machine Learning Approach., *Journal of Chemical Information and Modeling*, 50(5), 716–731, 2010.

[4] F. Rathke, *et al.*, Structrank: A new approach for ligand-based virtual Screening, *Journal of Chemical Information and Modeling*, 51(1), 83–92, 2011.

[5] W. Zhang, *et al.*, When drug discovery meets web search: Learning to Rank for ligand-based virtual screening., *Journal of Cheminformatics*, 7(5), 2015.

[6] T. Joachims, Optimizing search engines using clickthrough data., In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, 2002.

[7] T. Liu, *Learning to Rank for Information Retrieval*, Springer, 2011.

[8] C. P. Lee, *et al.*, Large-scale linear rankSVM., *Neural Computation*, 26(4), 781–817, 2014.

[9] Y. Wang, *et al.*, PubChem BioAssay: 2014 update., *Nucleic Acids Research*, 42(D1), 1–8, 2013.

[10] Q. Li, *et al.*, A novel method for mining highly imbalanced high-throughput screening data in PubChem., *Bioinformatics*, 25(24), 3310–3316, 2009.

[11] M. Lindh, *et al.*, Toward a benchmarking data set able to evaluate ligand- and structure-based virtual screening using public HTS data., *Journal of Chemical Information and Modeling*, 55(2), 343–353, 2015.

[12] G. Landrum, RDKit: Open-source cheminformatics., http://www.rdkit.org

[13] D. Rogers, *et al.*, Extended-connectivity fingerprints., *Journal of Chemical Information and Modeling*, 50(5), 742–754, 2010.

[14] K. Jarvelin, *et al.*, IR evaluation methods for retrieving highly relevant documents., In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 41–48, 2000.