Vol.2017-BIO-49 No.5
2017/3/24

薬剤活性予測の改良のための化合物フィンガープリントの比較解析

松山 祐輔[1,a)]　石田 貴士

概要：
機械学習による薬剤活性予測問題において，どの特徴ベクトルを用いるかというのは重要な問題である。これまで多数の化合物特徴ベクトルが提案されてきたが，どの標的，どのタスクにおいても最良の性能を出すものは存在しない。本研究では，複数の化合物フィンガープリントについて，薬剤活性予測についての比較解析を行うとともに，特徴ベクトル同士の類似度や複数の特徴ベクトルを使った場合の予測性能の変化について詳細に解析した．その結果，それぞれのフィンガープリントは化合物の重要な特徴をある程度別々に拾っており，それらの組み合わせか性能向上に寄与てきるてあろうことか考察された．

キーワード：薬剤活性予測，化合物特徴量

Comparative analysis of molecular fingerpirnts toward improvement of drug activity prediction

Abstract: Selection of feature vector used for machine learning drug activity prediction is an important problem. A number of molecular fingerprints have been proposed until now, there is no fingerprint that provides the best performance in any target protein, and any task.
In this study, we conducted comparative analysis on drug activity prediction performance using many molecular fingerprints, and analyzed the similarity between molecular fingerprints. We also analyzed the change in prediction performance when using multiple fingerprints in detail. As a result, it is considered that their molecular fingerprints pick up dome important features of compounds separately to some extent, and that using multiple molecular fingerprints would contribute to performance improvement.

Keywords: Drug activity prediction, molecular fingerprints

1. はじめに

近年，製薬の現場においては創薬コストの増大が問題となっている．この解決策の一つとして，計算機を用いたヴァーチャルスクリーニング手法が注目されている．これは従来，主に試験管の中で行われていた，化合物が所望の薬剤としての活性を持つかどうかの試験を計算上で高速に行おうというものである．ヴァーチャルスクリーニングには，大きく分けてドッキングシミュレーションによる手法 (Structure Base virtual screening; SBVS) とリガンドベースによる手法 (Ligand Base virtual screening; LBVS) に分けることができる．

このうち，LBVS では，薬剤標的タンパク質に対する活性が既知の化合物の構造情報を入力として予測モデルを構築する手法であり，一般に機械学習の手法が用いられている．一般的な機械学習手法において，入力とする特徴ベクトルの次元は一定である必要がある一方，化合物の大きさはまちまちである．化合物の構造を一定の長さの特徴ベクトルに変換する手法は，長年多くの研究がなされてきた．数多くの構造特徴ベクトルが提案されて来たが，大まかに四つの手法に分けることができる．一つ目は，辞書的な手法で，

1 東京工業大学情報理工学研究科計算工学専攻
Tokyo Institute of Technology, Meguro, Tokyo, 1552-8660 Japan
2 東京工業大学情報理工学院情報理工学系
Tokyo Institute of Technology, Meguro, Tokyo, 1552-8660 Japan
3 情報生命博士教育院
a) matsuyama@cb. cs. titech. ac. jp

ⓒ 2017 Information Processing Society of Japan

あらかじめキーとなる構造とその構造に対応するベクトルの番地を決めておき，それらの構造を化合物が持っていれば 1，もっていなければ 0 を立てることによる手法である．これには MACCS key[1] などが存在する．二つ目は，化合物が持つ結合に着目した手法である．これは，化合物が持つ結合の長さや種類，両端の原子の情報を用いて符号化する手法である．これには atom pair[2] などが存在する．三つ目には，化合物の各原子を中心とした半径 r 結合以内の原子を含めた構造を捉え，これを Morgan[3] 法などにより一意の数値に置き換え，ハッシュ関数を通して得られた数字に当たる番地にビットを立てていく手法である．これには Extended Connectivity Fingerprint (ECFP)[4] などがある．最後に，化合物が持つ標的タンパク質との結合について重要と見られる構造 (ファーマコフォア) の候補をリストアップし，この情報を符号化する手法である．これには 2-Dimentioanl Pharmacophore Fingerprint (2DPF)[5] などが存在する．

このように多数の特徴ベクトル化手法が提案されていることから，それらを比較した研究もいくつか存在する．例えば文献 [6][7] では，多数のターゲットを用いて類似性検索のタスクにおいて特徴ベクトル化手法の比較を行っており，平均的に ECFP4 が最も高い精度をもたらすことが報告されている．一方で，特徴ベクトル化手法の違いによる予測精度の差は小さく，どのタスク，どの標的タンパク質においても常に良い性能を返す特徴ベクトル化手法は存在しなかった．これは，それぞれの手法が化合物について異なる点に着目しているからと考えられる．このことから，それぞれの特徴ベクトルの性質の違いや，得手不得手を理解することはこれらを利用する上で非常に重要である．

一方，このように得手不得手のある特徴ベクトルを複数同時に予測に使うとどうなるだろうか．もし同時に使用される特徴ベクトルが化合物に関する情報を補完し合えるのであれば，複数の特徴ベクトルを用いることで一つの特徴ベクトルを用いるよりも予測精度の向上を見込むことができる．一方で，もし特徴ベクトルが化合物について同じような情報しか持たない場合は予測精度の向上は期待できず，逆に多重共線性や次元の呪いにより 予測精度が低下することも十分に考えられる．

本研究では，機械学習による薬剤活性予測タスクにおける各特徴ベクトル化手法の比較や考察をまず行ない，その後複数の特徴ベクトルを同時に使った場合の予測精度について検証を行った．その結果，特徴ベクトル単体としては先行の文献の通り ECFP 族が平均的に最も良いという結果になったが，加えてそれぞれの特徴ベクトルは別々に化合物の重要な特徴を拾っており，それらの組み合わせが性能向上に寄与できるであろうことが示された．

表 1 本研究で用いたアッセイデータおよびデータ個数の一覧

Pubchem AID	Target or goal	#Active	#Inactive
1915	Group A Streptokinase expression inhibition, streptkinase	5266	7751
2358	Protein phosphatase 1	1006	934
463213	Identify small molecule inhibitors of tim10-1 yeast	4141	3235
463215	Identify small molecule inhibitors of tim10 yeast	2941	1695
488912	Sentrin-specific protease 8	2491	3705
488915	Sentrin-specific protease 6	3568	2628
488917	Sentrin-specific protease 7	4283	1913
488918	Sentrin-specific proteases	3691	2505
492992	two pole domain potassium channel	2094	2820
504607	Mdm2/MdmX interaction	4830	1412
624504	mitochndrial permeablity transition pore	3944	1090
651739	Inhibition of Trypanosoma cruzi	4051	1324
651744	NIH，3T3 toxicity	3102	2306
652065	bind r(CAG) RNA repeats	2966	1287

表 2 本研究で用いた化合物特徴ベクトルの一覧

Feature	length	Ref
ECFP4/6	2048	David rogers(2010)
2DPF	3348	A. Gobbi(1998)
MACCS key	166	Symyx Technologies
Daylight	2048	Daylight Chemical Information Systems Inc.
Topological Torsion	2048	R. Nilakantan(1987)
atompair	2048	R. E. Carhart(1985)
Chemical/Structure descriptors	193	Using RDKIT

2. データセット

本研究では，データセットとして PubChem[8] より，**表 1** に示すターゲットに対するアッセイデータを利用した．これらのデータセットは，ニューラルネットを用いた転移学習についての研究 [9] で使われていたもので，各データセットにはあらかじめ活性の有無がラベルとして付与されている．それらの数についても表 1 に示す．

また，比較解析の対象となる特徴ベクトルとして，**表 2** に示す 7 つの特徴化手法によるものを用いた．いずれも特徴ベクトルの生成には RDkit[10] を用いた．物理化学的特徴ベクトル (Chemical/Structure descriptors) に関しては，RDkit がサポートしている特徴ベクトルを用い，0 から 1 の範囲で正規化した．また，本研究で用いた MACCS key は構造が公開されている 166 次元分のみを用いている．

3. 特徴ベクトルの違いによる予測精度の比較

まずはじめに，用意したデータセット及びを用いて，各特徴量ごとに薬剤活性予測モデルを機械学習を用いて構築し，それらの間の予測精度の比較を行った．学習には RandomForest[11] を用いた．これは後の解析で特徴量の重要度を利用可能であり，性能的にも優秀なためである．予測精度の評価には Receiver Operating Charactaristics (ROC) 曲線下の面積（Area Under the Curve，AUC）を用い，4-fold cross validation (CV) を行い，その平均を用いた．RandomForest のパラメータとして，決定木の数 ($n_estimators$)，および決定木で使う特徴数の上限 ($max_feature$) があるが，これらは以下の範囲においてグリッドサーチを行って決定した．

- $n_estimators$: $50, 100, 200, 300, 400$
- $max_features$: $50, 100, 200, 300, 400$

結果を**表 3** に示す．平均的に見て，Sereina らの文献 [6] と同じように，ECFP4/6 を用いた場合に予測結果が優れていることがわかる．物理化学的特徴ベクトルは難しい問題 (総じて AUC が低くなるような標的) においては構造的な特徴ベクトルよりも有効であったが，簡単な問題に対してはあまり高い精度を出せなかった．また，構造的な特徴ベクトルについては，その予測精度にそれなりの差異はあれど非常に大きな違いはみられなかった．

4. 特徴ベクトル同士の “類似度”

3 節の結果，平均的に見れば今回用いたどの特徴ベクトル化手法でも予測精度に大きな差はみられなかった．このような結果が生じた理由として，どの特徴ベクトルも結局化合物について同じような特徴のみを拾っており，その結果予測精度には大きな差が出なかったということが考えられる．そこで，特徴ベクトル同士の類似度がどのようになっているかを解析を行った．特徴ベクトルの類似性を解析には文献 [12] で用いられた手法を用いた．

この解析手法では，まず，各データセットよりランダムに 500 化合物を抜き出し，それらを各特徴ベクトル化手法を用いて符号化しておく．次に，それらの全組み合わせ約 25 万通りについて Tanimoto 係数を計算し，相関係数に従いそれらを並び替えリスト化する．その後，8 つの特徴ベクトルの各組み合わせごとにそのリストのスピアマン順位相関係数を計算する．この相関係数が高いほど，その特徴ベクトルのペアは似ていることになる．

結果を**図 1** に示す．各特徴ベクトル間の相関を見ていくと，まず ECFP4 と ECFP6 では相関係数が 0．95 以上と似ているが，これらは生成の際のパラメータのみが違うものであるため妥当な結果と考えられる．ところが，その他の組み合わせについては，多少の相関はみられるものの，そこまで相関が高いわけではないことがわかった．特に辞書ベースの特徴ベクトル (MACCS key) とその他の組み合わせについては相関が非常に小さいと言える．そのため「各特徴ベクトルは化合物の異なる特徴を捉えているにもかかわらず，それらの予測性能には大きな違いがない」という興味深い結論が得られた．

図 1 各特徴ベクトル間の類似度の比較

しかし，この解析では各特徴ベクトルのすべての要素を用いて化合物の比較を行っている．そのため，もし各特徴ベクトルの予測に重要な特徴はその長さ (ECFP なら 2048) の一部のみであって，その部分は各特徴ベクトルで共有されているにも関わらず，残りの部分がノイズとなって相関が得られていないという可能性も考えられる．また，どの特徴ベクトルにも共通して予測に重要な特徴が共通して存在していた場合，この部分を解析すれば普遍的な化合物の特徴を得ることができ，非常に有用であると考えられる．

4.1 予測において重要な特徴の比較

予測において重要な特徴が各特徴ベクトルで共通となっているか確認するため，各特徴の予測における重要度を計算し，重要な特徴のみでの特徴ベクトル間の相関を確認した..

本研究では予測に RandomForest を用いているため，各特徴の重要度を計算することが可能である．RandomForest の特徴ベクトルの重要度を測る手法には，他に OOB エラーを用いた手法やエントロピー誤差を用いた手法もあるが，今回は最もよく使われているジニ係数による手法を用いた．RandomForest の各決定木の学習において，データを分割する閾値を決めるのにジニ係数が用いられる．ジニ係数は，不純度とも呼ばれ，標本のがバラバラである程度を示す．RandomForest では，このジニ係数ができるだけ少なくなるような閾値を探し，新たな木のノードとする．よって，決定木を構成する際にそれぞれの特徴がどの程度ジニ係数を減少させるかといった観点により，各特徴の重要度を測ることができる．scikit-learn が出力する各特徴の重要度は，0 から 1 までの間の値を取るようになっている．各特徴ベクトルについて，この重要度が平均値*1．25 を上回

表 3 単独特徴ベクトルでの薬剤活性予測性能の比較．各アッセイに対し，平均+標準偏差以上の場合に*印をつけた．

PubChem AID	2dpf	atompair	daylight	ecfp4	ecfp6	maccskey	topo tor	descriptor	Average	StdDev
1915	0. 730	0. 753	0. 755	0. 752	0. 737	0. 765*	0. 764	0. 766*	0. 753	0. 012
2358	0. 746	0. 774	0. 773	0. 773	0. 760	0. 764	0. 775	0. 781*	0. 768	0. 011
463213	0. 654	0. 691	0. 697	0. 706*	0. 700	0. 691	0. 694	0. 686	0. 690	0. 015
463215	0. 599	0. 635	0. 637	0. 641	0. 646*	0. 628	0. 642	0. 628	0. 632	0. 014
488912	0. 637	0. 697	0. 682	0. 691	0. 685	0. 670	0. 692	0. 688	0. 680	0. 018
488915	0. 693	0. 737*	0. 719	0. 734	0. 733	0. 705	0. 732	0. 722	0. 722	0. 015
488917	0. 775	0. 820*	0. 803	0. 809	0. 806	0. 788	0. 818*	0. 804	0. 803	0. 014
488918	0. 742	0. 785	0. 770	0. 787	0. 783	0. 748	0. 787	0. 775	0. 772	0. 017
492992	0. 837	0. 850	0. 854	0. 863*	0. 858	0. 838	0. 861*	0. 844	0. 851	0. 010
504607	0. 653	0. 694*	0. 667	0. 670	0. 654	0. 676	0. 680	0. 683	0. 672	0. 013
624504	0. 888	0. 883	0. 898	0. 897	0. 883	0. 897	0. 895	0. 898	0. 892	0. 006
651739	0. 792	0. 800	0. 800	0. 802	0. 799	0. 785	0. 798	0. 807*	0. 798	0. 006
651744	0. 873	0. 879	0. 880	0. 881	0. 879	0. 862	0. 879	0. 884*	0. 877	0. 006
652065	0. 764	0. 789*	0. 762	0. 756	0. 742	0. 764	0. 753	0. 801*	0. 766	0. 018
Average	0. 742	0. 770	0. 764	0. 769	0. 762	0. 756	0. 769	0. 769		
Std. Dev	0. 085	0. 072	0. 076	0. 075	0. 075	0. 074	0. 074	0. 077		

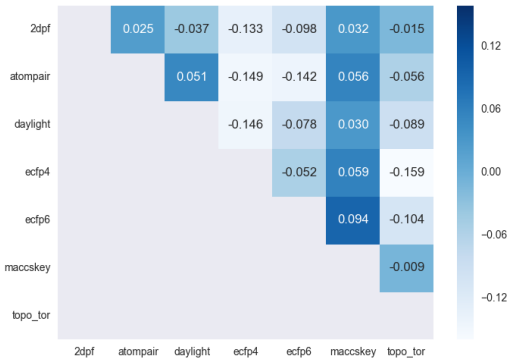

図 2 重要な特徴のみ取り出した場合の比較．抽出前 (図 1) から抽出後の数値を引いたもの．

る特徴だけを取り出し，前節と同じように 500 個の化合物を取り出し，全組み合わせでの Tanimoto 係数のリストの相関係数を計算した．仮に各特徴ベクトルが予測に重要な特徴を共有しているのであれば，これにより得られる相関係数は元の相関係数よりも高くなることが期待される．

結果を**図 2** 表に示す．この数値は重要な特徴を抽出する前の相関から抽出後の値を引いたものであり，プラスであれば相関係数が増加した＝予測に重要な特徴を共有していると考えられる．

相関係数の変化を見る限り，MACCS key 以外の特徴ベクトルのペアにおいては，重要な特徴に絞って計算しても相関が大きく上がることはなく，逆に下がることが多いことが読み取れる．このことから，それぞれの特徴ベクトルは予測に重要な特徴を共有してはおらず，それにも関わらず予測精度は似たようなものになるということがわかった．

4.2 特徴ベクトル空間における tSNE プロット

ここまでの解析から，それぞれの特徴量は異なる側面から化合物の情報を抽出していると考えられるが，各特徴がカバーする化合物の特徴とそれらの関係を理解するため，高次元データの圧縮アルゴリズムでもある t-distributed neighbor embedding(t-SNE)[13] を用いて可視化を行った．

tSNE は多様体学習の一種で，高次元空間に広がる特徴を低次元空間上で表現することができる．この手法では，高次元上での二つの点の距離を，1 次元の t 分布で表す．ある基準となる点 x_i を考える．このとき，この点を中心とした正規分布を考える．そして，低次元上に配置したい点 x_j と点 x_i の距離を，この正規分布から点 x_j が生成される確率密度と考える．このとき，同時確率 p_{ij} は

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} \exp(-||x_k - x_l||^2/2\sigma^2)}$$

と書ける．

ここで，点 x_i と点 x_j を次元圧縮後の低次元空間上へ写像したものをそれぞれ y_i，y_j とする．このとき，この二点においても同様に近さを考える．この時，確率分布を正規分布ではなく t 分布を用いることにより，’遠い’ データを正規分布に比べより遠くに配置することができる．ここで y_i と y_j の近さ q_{ij} を

$$q_{i|j} = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k \neq l} \exp(-||y_k - y_l||^2)}$$

と書くことができる．tSNE では，訓練データに基づいてこの p_{ij}，q_{ij} が似た分布となるようにパラメータを最適化する．最適化関数としては，カルバックライブラー (KL) 情報量を用いる．このコスト関数が最小となるように求めた y_i を x_i を次元圧縮後の点として表現する．

可視化の対象は AID1915 のデータとし，データセット

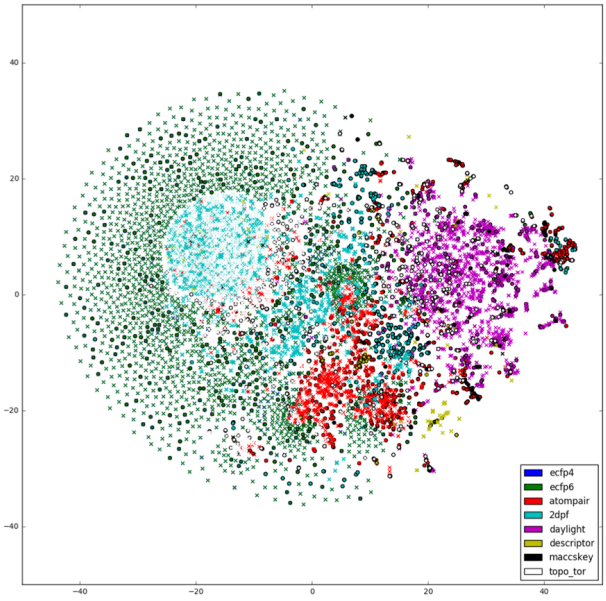

図 3　各特徴ベクトルの t-SNE プロット．

中に含まれる約 3000 の化合物を特徴ベクトル化したうえで一つの行列につなげた．次に，t-SNE により化合物の軸に従って 2 次元に圧縮し，これをプロットした．これを**図 4.2** に示す．なお，ECFP4(青色) の点が少ないが，これは ECFP6(緑色) の点にその多くが上書きされているためである．また図中の○印はその特徴の重要度が高い (平均の重要度*1．25 以上であった) ことを表し，×印はそうでないことを表す．この図から，

- ECFP 系，atompair，daylight がカバーしている領域は異なり，この三つで全体の大きな領域をカバーできる
- 「予測に重要な特徴」は特に偏在しているわけではないことがわかる．

図の中心部に Topology Torsion および 2DPF の×印の塊があるが，これは 1 がほとんど立っていないビットの集合と考えられる．この図を見るにあたって一つ注意することとして，各特徴ベクトルで 1 が立っているビットの数の分布の違いが挙げられる．**図 4** に AID1915 のデータセット中に含まれる化合物について，各特徴ベクトルで 1 が立っている個数の分布をヒストグラムにしたものを載せる．特に Daylight に関して，1 が立っているビット数が他の特徴ベクトルに比べて多いことが，tSNE プロットにおける領域が他と比べて分離している理由である可能性も考えられる．

5. 特徴ベクトルの組み合わせによる予測の改善

これまでの解析から，これまでに提案されてきた様々な特徴ベクトル化手法には予測精度の観点から大きな性能差は存在しないが，それらが捉えている化合物の特徴は大き

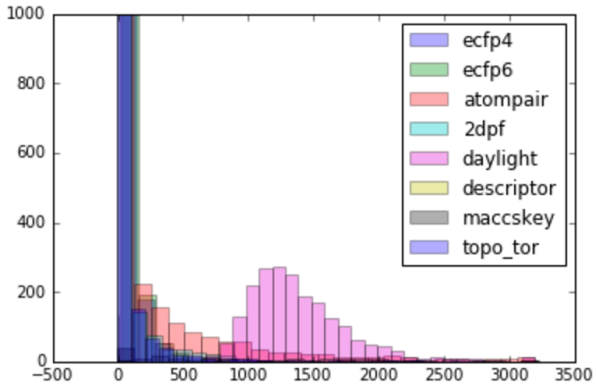

図 4　各特徴ベクトルにおける，1 が立っているビット数の分布

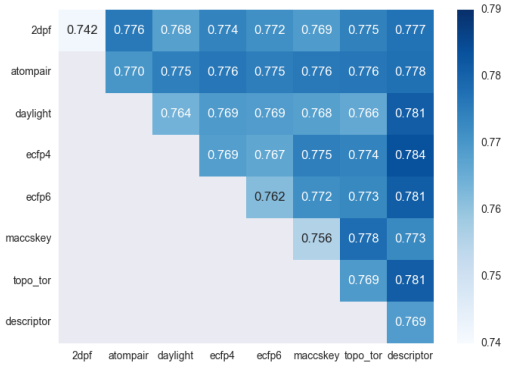

図 5　二つの特徴ベクトルを用いることによるの予測精度の変化．全データの平均値を記す．対角線上の平均 AUC の値は単独特徴量の時の性能 (表 3 に同じ)

く異なることがわかった．そのため，これらの複数の特徴ベクトルを組み合わせて予測モデルを構築することで，予測モデルの精度向上が見込めると考え，2 つの特徴ベクトルを単純につなげて一つの特徴ベクトルとしたものを用いて，2 節と同じ方法で学習を行い，予測性能を評価した．評価指標としては前の解析と同様に AUC を用い，4-fold CV を行った．結果を**図 5** に示す

多くの場合において，単独の特徴ベクトルによる予測性能よりも，2 つの特徴を用いた場合の予測性能の方が上がっていることがわかる．一方，予測精度の向上が見られなかった組み合わせも存在し，ECFP4 と ECFP6 を用いた場合の予測精度は ECFP4 単独の 0．768 から 0．765 に低下している．

6. 考察

6.1 特徴ベクトル間の類似度と組み合わせによる予測精度の向上

特徴ベクトルを組み合わせた予測を行った際，ECFP4 と ECFP6 の組み合わせでは予測精度は逆に低下した．先の解析結果で示された様に ECFP4 と ECFP6 は非常に似ている特徴ベクトルであり，そのようなものの組み合わせ

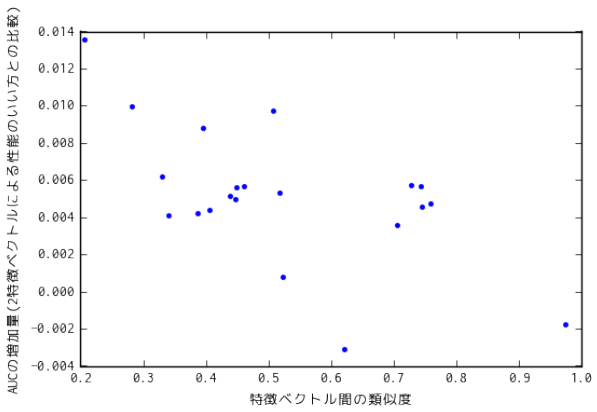

図 6　特徴ベクトル間の類似度と AUC の増加量との関係．図は特徴ベクトル A 及び B を用いた予測性能から，A および B 単独で用いた時の性能のうち良いものを引いたものである．

では予測精度の向上は期待できないという可能性が考えられる．そのため，特徴ベクトル間の類似度と組み合わせによる予測精度の向上の関係について解析と考察を行った．**図 6.1** に，横軸を特徴ベクトルペアの類似度，縦軸を AUC の増加量としてプロットしたものを示す．増加量は特徴ベクトル A および B を用いた時の予測性能から，特徴ベクトル A と特徴ベクトル B を単独で用いた時の予測性能のうち性能のいい方のものを引いたものである．この図から，類似度が小さい特徴ベクトルの組み合わせでは，類似度が大きい特徴ベクトルの組み合わせよりも予測性能の向上幅が大きい傾向にあることが読み取れる（ピアソン相関係数:-0．595）．

6.2　各特徴ベクトルの予測結果の比較

今までの解析から特徴ベクトル間の類似度が低いことが明らかとなったが，更なる確認のため予測結果の詳細についても解析を行った．各特徴ベクトルについて，それぞれで学習させた予測モデルの予測結果についてそれらがどの程度共有されているか確認した．AID1915 を対象とし，そこに含まれるデータを用いて 8 つの特徴ベクトル単独で学習させた場合の，テストセットにおける活性があると強く予測された上位 20 パーセント (162 個) の化合物についてベン図を描画した．**図 6.2** に，その一部を載せる．ベン図内の数字については，左側が実際に活性のある (Active) 化合物の個数であり，右側の数字が予測された個数を表している．

これらのベン図から，それぞれの予測結果の重なる部分は全体のおおよそ 2/3 程度であり，それぞれの特徴ベクトルにより学習された予測モデルで上位に予測された化合物の少なくない部分が重なっていないことがわかる．このことから複数の特徴ベクトルを用いることで，予測される化合物の多様性が広がる可能性が考えられる．また，ベン図の和集合の部分，つまり二つの特徴ベクトルによる予測モデルでどちらも上位だと予測した化合物の領域では，そうでない部分と比べると全体的に正解率が高くなっている．これは 化合物について複数の尺度から既知の活性化合物に似ていると判定された化合物は，実際に活性を持つ化合物である可能性が高いということであり，複数の特徴ベクトルの組み合わせが予測精度の向上に有効であることが示唆されている．これらのことから，複数の特徴ベクトルを用いることにより，片方の特徴では間違えてしまうような化合物を他の特徴では正しく捉えることができる場合があり，そのためにお互いが補うことにより予測精度の向上が見られたのではないかと考えられる．

図 7　それぞれの特徴ベクトル単独の予測モデルによって強く活性があると予測された上位 20 %以内に入った化合物の重なり具合．図中．/ の右側は予測された数，左側はそのうち Active である化合物の数

7.　まとめと今後の課題

本研究では，まずこれまでに多数提案されている特徴ベクトルについて，機械学習による薬剤活性予測の文脈において比較，解析を行った．その結果，先行研究で示されたとおり，各特徴ベクトル化手法には予測精度の観点から大きな性能差はないことが確認された．その一方で，各特徴ベクトルは化合物の予測に重要な特徴をある程度は別々に拾っており，それらの間の類似性，相関は大きくないとい

うことがわかった．さらに，これら特徴ベクトルを組み合わせた上で学習することにより，予測性能の向上が見込まれることを示した．

一方，ECFP4 と ECFP6 を組み合わせた場合の様に，単に特徴ベクトルの次元を増やすと，一般に「次元の呪い」と呼ばれる問題に直面することになる．今後の課題としては，各特徴ベクトルから予測に重要な特徴をうまく抽出した特徴ベクトルを構築することや，複数の学習器を各特徴ベクトルによって構成しその結果を組み合わせるアンサンブル学習の手法を用いることなどが挙げられる．

8. 参考文献

参考文献

[1] Joseph L. Durant et al., “Reoptimization of MDL Keys for Use in Drug Discovery”, J. Chem. Inf. Comput. Sci. 2002, 42, 1273-1280

[2] Raymond E. Carhart, Dennis H. Smith, R. Venkataraghavan, “Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications”, J. Chem. Inf. Comp. Sci. 25:64-73 (1985)

[3] H. L. Morgan “The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.”, J. Chem. Doc., 1965, 5(2), pp 107-113

[4] David Rogers and Mathew Hahn, “Extended-Connectivity Fingerprints”, Journal of Chemical Information and Modeling, 2010, vol.50(5), pp.742-754

[5] Malcolm J. McGregor and Steven M. Muskal, “Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design”, Journal of Chemical Information and Modeling, 1999, vol.39(3), pp 569-574

[6] Sereina Rinikerand Gregory A Landrum, “Open-source platform to benchmark fingerprints for ligand-based virtual screening”, Journal of Cheminformatics, 2013, 5:26

[7] Noel M. O’Boyle and Roger A. Sayle, “Comparing structural ngerprints using a literature-based similarity benchmark”, Journal of Cheminformatics, 2016, 8:36

[8] Sunghwan Kim et al., “PubChem Substance and Compound databases”, Nucleic Acids Research.i, 2016, 44, D1202-D1213

[9] George E. Dahl and Navdeep Jaitly, “Multi-task Neural Networks for QSAR Predictions” , arXiv 入手先 ⟨http://arxiv.org/abs/1406.1231⟩, 2014

[10] RDKit: Open-source cheminformatics; 入手先 ⟨http://www.rdkit.org⟩

[11] Breiman, L. “Randon Forests”, Machine Learning (2001) 45: 5., doi:10.1023/A:1010933404324

[12] David Duvenaud et al., “Convolutional Networks on Graphs for Learning Molecular Fingerprints”, NIPS 2015

[13] Laurens van der Maaten and Geoffrey Hinton, “Visualizing Data using t-SNE”, Journal of Machine Learning Research 1 (2008) 1-48

ⓒ 2017 Information Processing Society of Japan