

適用領域を考慮した薬剤標的親和性予測

杉田 駿也

大上 雅史

東京工業大学 情報理工学院 情報工学系

1 序論

創薬研究において、薬剤の標的親和性を計算によって予測することは、初期段階の化合物スクリーニングや未知の標的への結合可能性の検証等に有用である。特に薬剤と標的の両方の情報を活用した機械学習による予測法は、迅速に薬剤-標的間の網羅的な親和性予測が可能である。

2017年に提案されたSimBoost[1]は、薬剤同士および標的の類似性から得られる特徴量を用いて勾配ブースティング法により薬剤標的親和性を予測する手法であり、高い予測精度を持つ。しかし、化合物空間は一般に広範であり、訓練データセットがカバーできないサンプルの存在が考えられ、それらは予測精度に影響する。適用領域 (applicability domain, AD) は予測モデルが妥当な性能を発揮できるデータの範囲のことであり、ADを求めることにより妥当な性能を発揮できるサンプルを事前に推定することが可能となる。

本研究では、訓練データのデータ密度により決定されるADを考慮した薬剤標的親和性予測手法を提案する。ADを求める手法としてk最近傍法(k-NN)および1-クラスSupport Vector Machine(OCSVM)を用い、あらかじめ入力サンプルがAD内となるかAD外となるかを推定することで予測精度が改善できるかどうかを検証した。

2 手法

2.1 データセット

提案手法の精度を評価するために、SimBoostの文献[1]と同様に表1に示す3つのデータセット(Davis[3], Metz[4], KIBA[5])を使用した。これらはIC₅₀などの標的活性値(実数値)のラベルを持つ。各データセットは、4:1の割合で訓練データとテストデータに分割し、訓練データは予測モデルの学習およびADの学習を行うために利用した。

2.2 予測モデルの学習

薬剤標的親和性の予測モデルの学習は、SimBoost[1]を用いた。実装は[2]の公開コードを利用した。特徴量も文献[1]と同一とし、ハイパーパラメータは文献[1]で最適化された値を用いた。

表1 本研究で使った薬剤標的活性のデータセット

データセット	薬剤数	標的数	ラベル付データ件数
Davis [3]	68	442	30,056 (100%)
Metz [4]	1,421	156	93,326 (42.1%)
KIBA [5]	2,116	229	118,234 (24.4%)

% 表記は薬剤数 × 標的数の数に対する割合を示す。

2.3 適用領域 (AD)

テストデータセットのサンプルが予測モデルのカバー範囲にあたるか否かを事前に推定するため、本研究ではADを活用した。ADを求める手法として、本研究ではk-NNとOCSVMを検討した。それぞれ、AD外となるデータの割合をパラメータ ν として事前に与えた。k-NNの ν はテストサンプルに占めるAD外の割合、OCSVMの ν は訓練サンプルのうちAD外となる割合を示す。

k最近傍法 (k-NN)

k-NNは、テストサンプルと近傍の訓練サンプル k 個との距離によってADを定める手法である。サンプル間の距離は、特徴ベクトル $\mathbf{x}$ に重み $\mathbf{w}$ をかけた $\mathbf{w}^T \mathbf{x}$ を用い、ユークリッド距離により算出した。重み $\mathbf{w}$ の要素はSimBoostによって算出した。なお、各特徴量はあらかじめ訓練データの分布に基づいて標準化を行ったものを用いた。

1-クラスSVM (OCSVM)

OCSVMは、全てのサンプルを同一クラスとみなして識別平面を構築するSVMである。学習した識別平面をADとみなし、識別平面から外れたサンプルはAD外と推定される。カーネル関数としてRBFカーネルによる非線形写像を用いるため、特徴量は正規化せずにそのまま用いた。RBFカーネルのパラメータ γ は、文献[6]に従いグラム行列の分散を最大にするように選択した。

2.4 評価方法

評価指標として、親和性予測値の順序関係が正しいかどうかを評価するconcordance index (CI) と、親和性予測値そのものの誤差を評価する根平均二乗誤差 (root mean squared error, RMSE) を使用する。

概要を図1に示す。テストデータ全体での精度評価の他、ADによってAD内かAD外かを分類し、それぞれのデータでの精度評価も行った。

3 結果と考察

各データセットでの予測精度を表2に示した。各データセットの結果の1行目はSimBoostで学習した予測モデルをそのままテストデータセットに適用した場合の、ベースラインとなる予測精度である。

Drug-target affinity prediction with applicability domain

Shunya Sugita & Masahito Ohue, Department of Computer Science, School of Computing, Tokyo Institute of Technology

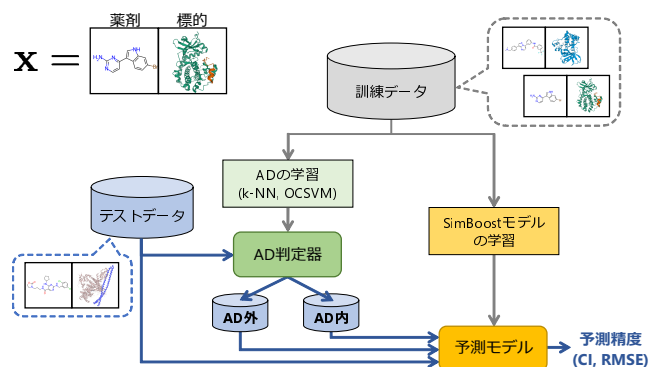


図1 適用領域 (AD) を考慮した予測手法と評価方法

いずれのデータセットでも、OCSVM より k-NN の方が上手く AD を決定できたと考えられる。パラメータ ν を大きく (適用領域を狭く) するほど RMSE 値は改善され、 $k = 10$ より $k = 5$ の方が高い精度を示した。また、これらの AD 外のデータについての RMSE はベースラインと比べて悪い値となっており、予測が難しいサンプルが AD 外と判定されたことが推察される。サンプルが AD に入るかどうかを提示することが、より信頼性の高い予測を提供できることを示した。CI 値については AD 内はいずれもベースラインと同等であった。

パラメータ ν による RMSE 値の変化は、データセットのラベル付きデータの割合 (密度) に依存すると考えられる。薬剤-標的の全てのペアにラベルが存在する Davis では AD の狭い $\nu = 0.03$ でも良い予測が可能であるが、ラベル有りサンプルの割合が小さい Metz と KIBA ではサンプル空間が疎となっている可能性があるため、AD を広めに (AD 外とする判定を厳し目に) する $\nu = 0.01$ の方が AD 外サンプルの予測が悪くなる結果となった。実用上は、訓練データセットとして使えるサンプルの薬剤数 $\times$ 標的数に対する割合と、予測を適用したいサンプルが訓練データセットとどの程度近いか (薬剤もしくは標的の片方が訓練データと共通するか否か) で AD の範囲を決定すると良いと考えられる。

4 結論

本論文では薬剤標的親和性の予測の精度と信頼性の向上ために、AD を併用した予測手法を提案した。k-NN と OCSVM で構築した AD に対して3つのデータセットで予測精度を検証した結果、k-NN によって構築された AD がサンプルを適切に分類できることを示した。ただし、k-NN ではテストデータに対する AD 外サンプルの割合を事前に決める必要があり、例えば予測したいサンプルを1個だけ与えられた場合に AD 内かどうかを判定する基準がない。k-NN のユークリッド距離の閾値を利用するなど、一定の基準を訓練データセットから求める方法が必要であり、これは今後の課題である。

薬剤標的親和性予測の最近の手法では深層学習が用いられている [7]。AD は深層学習にも適用可能であり、今後同様に検証を進めていきたい。

表2 3つのデータセットにおける予測精度

データセット	AD	ν	判定	CI	RMSE
Davis	-	-	-	0.901	0.488
	k-NN ($k = 5$)	0.01	AD 内	0.902	0.485
			AD 外	0.842	0.738
	k-NN ($k = 5$)	0.03	AD 内	0.900	0.476
			AD 外	0.861	0.771
	k-NN ($k = 10$)	0.01	AD 内	0.902	0.485
			AD 外	0.832	0.690
	k-NN ($k = 10$)	0.03	AD 内	0.899	0.479
			AD 外	0.878	0.709
	OCSVM	0.01	AD 内	0.902	0.488
			AD 外	0.852	0.429
	OCSVM	0.03	AD 内	0.902	0.489
Metz				0.873	0.436
	-	-	-	0.849	0.413
	k-NN ($k = 5$)	0.01	AD 内	0.849	0.410
			AD 外	0.805	0.700
	k-NN ($k = 5$)	0.03	AD 内	0.848	0.407
			AD 外	0.842	0.585
	k-NN ($k = 10$)	0.01	AD 内	0.849	0.411
			AD 外	0.828	0.638
	k-NN ($k = 10$)	0.03	AD 内	0.848	0.408
			AD 外	0.854	0.569
	OCSVM	0.01	AD 内	0.850	0.411
			AD 外	0.784	0.599
KIBA	OCSVM	0.03	AD 内	0.851	0.410
			AD 外	0.806	0.521
	-	-	-	0.835	0.452
	k-NN ($k = 5$)	0.01	AD 内	0.835	0.443
			AD 外	0.812	1.027
	k-NN ($k = 5$)	0.03	AD 内	0.835	0.433
			AD 外	0.798	0.866
	k-NN ($k = 10$)	0.01	AD 内	0.835	0.443
			AD 外	0.814	1.015
	k-NN ($k = 10$)	0.03	AD 内	0.835	0.433
			AD 外	0.818	0.871
	OCSVM	0.01	AD 内	0.835	0.452
			AD 外	0.815	0.481
	OCSVM	0.03	AD 内	0.834	0.449
			AD 外	0.832	0.536

※太字は各データセット (AD 内) で最も良い RMSE 値を示す。

謝辞 本研究は、JSPS 科研費 基盤研究 (B) (No. 20H04280)、JST ACT-X (No. JPMJAX20A3)、上原記念生命科学財団の支援を受けて行われた。

参考文献

- [1] He T, et al. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines, *J Cheminform*, 9:24, 2017.
- [2] <https://github.com/hetong007/SimBoost/>
- [3] Davis MI, et al. Comprehensive analysis of kinase inhibitor selectivity, *Nat Biotechnol*, 29:1046–1051, 2011.
- [4] Metz JT, et al. Navigating the kinome, *Nat Chem Biol*, 7:200–202, 2011.
- [5] Tang J, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis, *J Chem Inf Model*, 24:735–743, 2014.
- [6] Kaneko H & Funatsu K. Fast optimization of hyperparameters for support vector regression models with highly predictive ability, *Chemom Intell Lab Syst*, 142:64–69, 2015.
- [7] Jiang M, et al. Drug-target affinity prediction using graph neural network and contact maps, *RSC Adv*, 10:20701–20712, 2020.