3D7-4

構造類似性に基づく Active QSAR モデリング

Active QSAR Modeling Based on Structure Similarity

西野 達也 Tatsuya Nishino　藤島 悟志 Fujishima Satoshi　高橋 由雅 Yoshimasa Takahashi

豊橋技術科学大学 工学部 知識情報工学系
Department of Knowledge-based Information Engineering, Toyohashi University of Technology

When a training set of chemicals for QSAR modeling consists of structurally diverse compounds, in many cases, the model obtained don't give us good prediction for the corresponding prediction set. In this paper, in order to solve such a problem and explore higher performance in the prediction, we have investigated a technique for active QSAR Modeling that is based on active sampling with a query. In this method, when a query is specified, structurally similar compounds are searched and collected to make a local model around the query. Then a QSAR model for the prediction is explored with the subset of the training set, and used for the data prediction of the query. First, we validated the present method with artificial data. Next, to compare the predictive performances of the method with that of the conventional approach, structure-toxicity modeling was carried out with an aquatic toxicity database. Here the TFS (Topological Fragment Spectra) method is used for the numerical description of chemical structure to perform similar structure searching. For the QSAR modeling, a linear regression modeling is employed with substructure features of our interest. Computational experiment with the real database is resulted in that the present approach provides us better predictions for the data that have structurally diverse compounds.

1. 背景と目的

QSAR(Quantitative Structure-Activity Relationships：定量的構造活性相関)とは化学物質の構造と活性（毒性）との関係を定量的にモデル化し、これをもとに未知の化合物の活性をその化学構造から予測しようとするものである。しかし、構造的に多様な物質を対象とした毒性予測においては、対象とする化学物質の拡大に伴い、その近似精度が著しく低下してしまうという問題がある。このことから、本研究ではより精度の高い予測結果を得ることを目的として構造類似性検索による動的なモデルの生成と、その予測への応用について検討を行った。

2. 方法

2．1 構造類似性に基づく ActiveQSAR モデリング

従来の QSAR モデリングでは、図1に示す訓練集合が与えられた場合、与えられる訓練集合の全サンプルを用いて予測モデルを生成していた。すなわち、従来法では、予測対象となるクエリ(a)～(d)すべてに対して予測モデル D が適用されることになる。この予測モデル D はクエリ(d)に対しては十分な予測精度が期待できる。しかし、他の3つのクエリ(a)～(c)に対する予測モデルとしては十分な予測が成されるモデルとは言い難い。実際には、クエリ(a)には A、(b)には B、(c)には C の予測モデルが形成される事で最適な予測が期待できる。そこで、本研究では予測対象となるクエリごとに構造類似性検索を行い、その検索結果に基づいて動的に局所近傍を作成する QSAR モデリングを提案する。

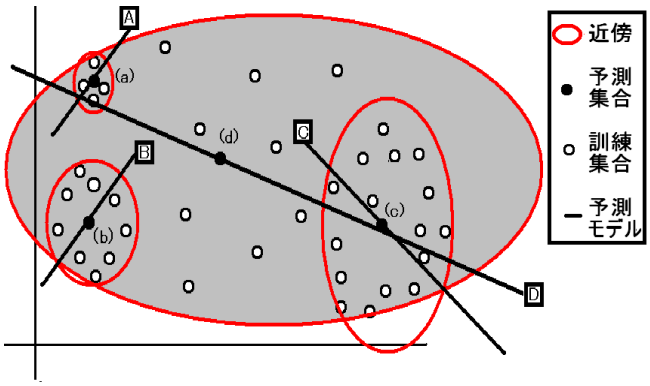

図1 最適な近傍作成に必要なサンプル

3. 合成データによる実験

本手法の提案の実効性を検証するため、図 1 に示した概要に従うクラスタ分布をする人工データを作成し、従来の一括モデリングと本手法の予測精度を比較した。実際に作成したデータの分布を図2に示す。

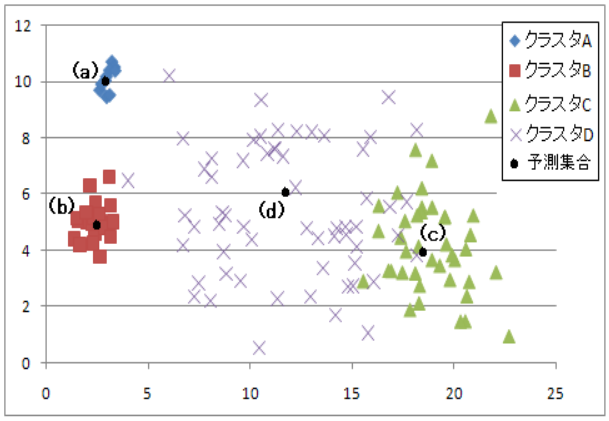

図 2 人工データの予測対象と訓練集合のクラスタ分布

図2の人工データにおいて、一括モデリングでの予測（従来法）と、訓練集合のクラスタ毎に動的に生成したモデルでの予測（提案法）、それぞれの予測誤差を表1に示す。

連絡先：高橋由雅，〒441-8580 愛知県豊橋市天伯町雲雀々丘 1-1 豊橋技術科学大学 知識情報工学系，Tel: 0532-44-6878, taka@mis.tutkie.tut.ac.jp

表１　人工データによる従来手法と本手法の誤差値

予測対象		(a)	(b)	(c)	(d)
分散(x,y)		(0.1,0.1)	(0.5,0.5)	(3.0,3.0)	(8.0,6.0)
クラスタのサンプル数		10	20	40	60
誤差	従来法	3.47	1.67	0.33	0.70
	提案法	0.07	0.03	0.15	0.64

表1の結果から、予測対象に対してクラスタ数が少なく、かつ予測対象からの分散が小さい予測対象(a)と(b)については従来法では大きな予測誤差を取ることになった。対して、提案手法では非常に小さい誤差で予測できていることが分かる。また、予測対象に対してクラスタ数が多く、かつ予測対象からの分散が大きい予測対象(c)と(d)については従来法と提案手法では予測対象(a)と(b)程は予測誤差に大きな差は見られず、特に予測対象(d)については従来法、提案法とも予測誤差はほぼ同じ値となった。この結果から、予測対象に最適な局所近傍を予測対象ごとに形成する事が出来れば、従来方法よりも精度の高い予測モデルを形成する事ができると言える。特に、予測対象に類似した(分散の少ない)訓練集合のクラスタを形成する事が出来ればその予測精度を顕著に上げることができることがわかった。

4. 化学物質の魚毒性 QSAR モデリングへの応用

ケーススタディとして水生毒性の指標のひとつである魚毒性を取り上げ、予測対象に最適な局所空間で予測する事で、その精度が向上するかを検討した。

4.1 データセット

データセットには毒性値として米国環境保護庁により収集された魚毒性データ 468 件を用いた[Admans 02]。毒性値は Fathead Minnow と呼ばれる小魚に対する 96h LC50[mol/l]である。

4.2 構造類似性評価

類似性評価のための構造特徴の記述には TFS (Topological Fragment Spectra) [Takahashi 98]を用いた。TFS は当研究室で考案された化学構造情報の定量的な記述手法である。その生成手順は、(1) 対象とする化学構造式から可能なフラグメントをすべて列挙し、(2) 列挙したそれぞれのフラグメントに対して数値的な特徴付けを行う。(3) その特徴付けの値と出現頻度のヒストグラムを生成する。このヒストグラムが TFS であり、これを多次元パターンベクトルとして用いることで、化学物質の構造特徴を数値的･定量的に表すことができる。本研究ではサイズ 5 までの部分構造を列挙し、特徴付けには各フラグメントの構造構成原子の質量数の和を用いた。

4.3 実験条件

予測実験に際して、データセットをランダムに 384 件(80%)の訓練集合および 94 件 (20%)の予測集合へ分割した。各予測集合に対する訓練集合の類似度を Cosine 係数で評価して類似度検索を行った。予測対象ごとに最適な近傍空間が存在するという本提案を検証するために、近傍作成に用いる訓練集合の件数を予測対象の構造類似性検索結果の上位 4～384 件まで変化させて、各近傍数で回帰分析による予測実験を行った。

4.4 実験結果

予測評価は各予測集合と予測モデルとの予測誤差をRMSで評価した。図3に近傍作成に用いる件数と RMS の推移を示す。

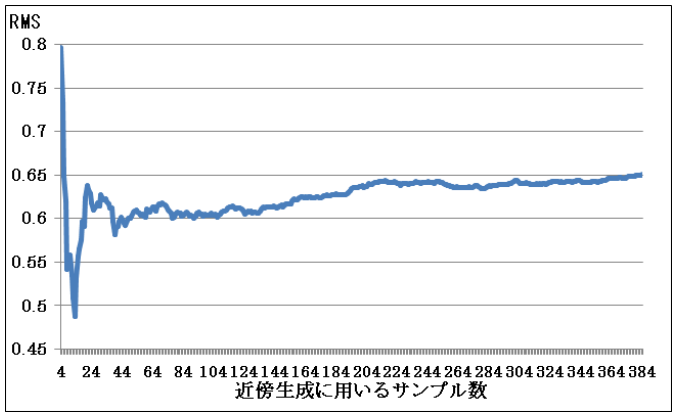

図３　近傍の件数と RMS の推移

図 3 から、予測精度の差が顕著に現れる近傍が作成できるサンプルの件数と RMS の推移を表2に示す。

表２　　近傍作成に用いるサンプルの件数と RMS

件数	4	5	6	7	…	13	…	384
RMS	0.80	0.73	0.65	0.62	…	0.49	…	0.65

この表2から、近傍に用いるサンプル数が 4～6 件の間は全訓練集合 384 件を用いて予測をした場合よりも RMS が同じかそれよりも高い値をとっている。しかし、それ以上の件数で近傍作成をした場合は、全訓練集合 384 件を用いて予測をした場合よりも低い RMS を取ることができた。つまり、サンプル数 7 件以上で作成された近傍での予測は、全訓練集合での予測に比べて高い予測結果を得ることができたといえる。特に 13 件のサンプルで作成された近傍での予測の RMS は、全訓練集合で予測した場合と比較して RMS が 25%も減少していることがわかる。この結果から、魚毒性に対する毒性予測では全訓練集合を用いてモデル生成をするよりも、構造的に類似した訓練集合の局所空間からモデル生成をした方がより高精度な予測モデルを生成することができることが示された。

5. まとめ

ケーススタディとして検証した魚毒性のモデルにおいて、予測対象に構造的に類似した化合物の局所近傍で予測をする事の有用性を示すことができた。今後は、今回検証に用いた魚毒性データ以外のデータセットについても本手法を適用し、その有効性を検討する。

参考文献

[Admans 02] Gary Admans: Artificial Neural Network for Predicting the Toxicity of Organic Molecules, Bull. Chem. Soc. Jpn., 74, 2451-2461.

[Takahashi 98] Y. Takahashi, H. Ohoka and Y. Ishiyama: Structural Similarity Analysis Based on Topological Fragment Spectra, Advances in Molecular Similarity, Vol.2, 93-104.