KP21

索引層の導入による SOM の学習高速化と薬物構造類似性マッピングへの応用

（豊橋技術科学大）○佐々木 英史、高橋 由雅*

1. はじめに

当研究室では、”類似性”の概念を基礎としたよりやわらかな化学構造情報の取り扱いとその化学データマイニングへの応用に関する研究を行ってきた。本研究は、その要素技術の1つとして、SOM による TFS 空間の可視化のための学習高速化アルゴリズムを提案する。SOM の学習におけるボトルネックは、BMU 探索時間である。そこで競合層上の BMU 出現域を推定するために索引層を導入した学習高速化アルゴリズムを提案する。また、改良法の性能を評価するためにドーパミンアンタゴニスト 1354 化合物の構造類似性マッピングへの応用を試みた。

2. SOM の改良

2.1. 従来の基本学習アルゴリズム

自己組織化マップ（Self-Organizing Map、SOM[2]）は、教師無し学習によって、多次元データを、実データ空間中の近接関係を保持したまま可視空間に非線形写像することができる可視化技術の1つである。従来の基本学習アルゴリズムは、以下の通りである。

(1) 学習データの提示
(2) 競合層上のユニット全てに対して類似度の計算を行い、類似度が最大となるユニット BMU (Best Match Unit)を探索
(3) BMU の近傍空間に位置する競合層ユニットの重みを学習データに近づける。

*taka@mis.tutkie.tut.ac.jp

最大学習回数あるいは RMS などの学習条件を満たすまで、各データに対して(1)～(3)の処理を繰り返すことで学習が行なわれる。

従来法では、1つの学習データに対し全ての競合層ユニットに対して類似度計算を求め、最大類似度を持つ BMU を探索する。そのため、重み修正処理に比べると計算量が大きく、訓練時間の 90%以上を BMU 探索時間が占める。

また、サンプル数やデータの次元数が増大により、データ空間が複雑になった場合、競合層ユニット数も合わせて増やす必要がある。これは更に計算時間を大きなものにさせる。そのため、大規模データセットを対象に計算を行うには、計算量を削減し高速化するためになんらかの工夫が必要となる。

2.2. 索引層を用いた学習高速化アルゴリズム

本研究では、こうした大容量多変量データ空間の可視化への応用を念頭に置き、競合層上での BMU の出現域を大まかに推定するための索引層の利用を基礎とした学習高速化アルゴリズムを提案する。その概念を図 1 に示す。

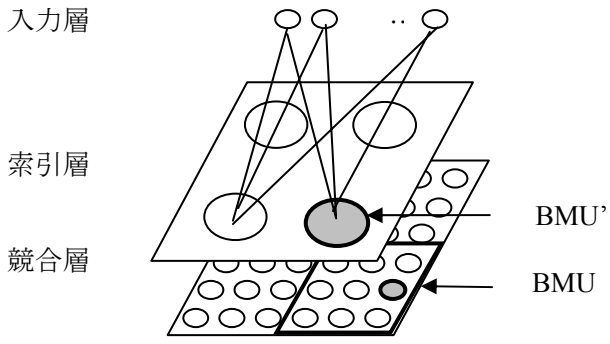

図 1 索引層を用いた学習高速化アルゴリズム

索引層を導入した場合の学習アルゴリズムは以下の通りである。改良法も、従来法と同様に、学習条件を満たすまで以下の処理を繰り返すことで学習する。

(1) 学習データの提示
(2) 索引層上のユニット全てに対して類似度の計算を行い、BMU'を探索
(3) BMU'の直下の競合層空間をBMU出現域と限定し、BMUを探索
(4) BMUの近傍空間に位置するユニットの重みを学習データに近づける。同時に、索引層ユニットの重みを修正した競合層ユニットの重みに近づける

図1を例に改良法を説明する。図1では、競合層サイズ6×6、索引層サイズ2×2である。

従来法の場合では、1つのデータが提示されると全競合層ユニットに対し類似度を計算しBMUを探索する。一方、改良法の場合では、まず索引層ユニット上でBMU'を探索する。この直下に位置する(一部の)競合層ユニットと類似度を計算しBMUを探索する。その結果、索引層を用いることで類似度計算回数を36回から13回に抑えることができる。

ここで重要となるのは、索引層ユニットが直下に位置する競合層ユニット群の代表となるかという点である。今回は、索引層ユニットの重みを学習データそのものに近づけるのではなく、修正した競合層ユニットの重みに近づけるアルゴリズムを採用した。

3. 薬物構造類似性マッピングへの応用

3.1. 実験方法

実データを用いて、索引層を用いた学習高速化アルゴリズムを適用したSOMの性能を評価するために、薬物構造類似性マッピングへの応用を試みた。

実験に際し、4種のドーパミン受容体のアンタゴニスト活性を取り上げ、構造的特徴をTFSによって記述した。データセットにはMDL社より市販されている治験薬データベース(MDDR[2])から抽出した、4種の異なる受容体(D1、D2、D3、D4)に作用するドーパミンアンタゴニスト1354化合物(D1:169、D2:429、D3:250、D4:573、複数活性を有す化合物含む)を用いた。ただし、SOMは教師無し学習であるため、学習にクラス情報は使用していない。

薬物の構造特徴記述子には、筆者らの提案するTFS(Topological Fragment Spectra[3])を用いた。TFSとは化学物質の構造式から可能な部分構造を列挙し、その数値的な特徴付けに基づいて化学物質のトポロジカルな構造プロフィールを多次元数値ベクトルとして表現しようとするものである。ここでは、結合サイズ5までの部分構造を列挙し、特徴付けには各部分構造の質量数を用いた。結果として、各化合物の構造特徴は164次元のベクトルとして記述された。

実験に用いたSOMのその他の学習パラメータを表1に示す。

表 1 学習に用いたパラメータ

競合層サイズ	49×49
索引層サイズ	7×7
学習回数	1000
類似度関数	ユークリッド距離
初期学習率	0.4
初期近傍半径	14
初期マップ	主成分平面

3.2. マッピング結果

以上の条件で学習を行った結果、得られた薬物構造類似性マッピングの結果を図2に示す。

図2(a)に競合層での薬物構造類似性マッピングの結果を示す。図中の色はそれぞれドーパミンアンタゴニストの活性分布を示す(図中の色はそれぞれ□D1、▒D2、▓D3、■D4を示す)。これから同じ活性クラスに属する薬物群が写像空間上にクラスターを形成していることが分かる。

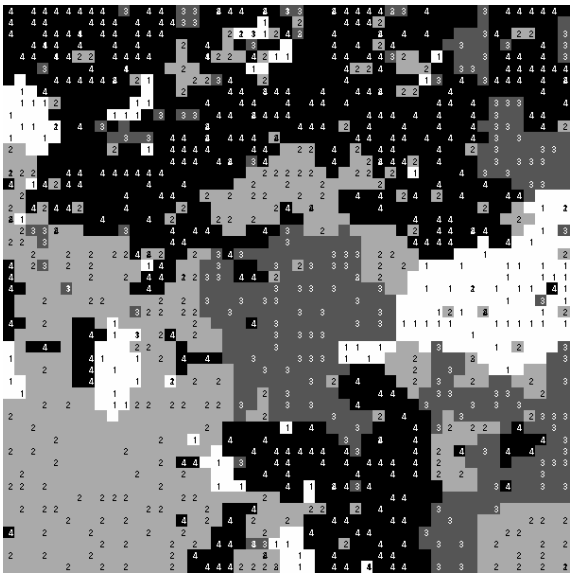

図 2(a) 構造類似性マッピング(競合層)

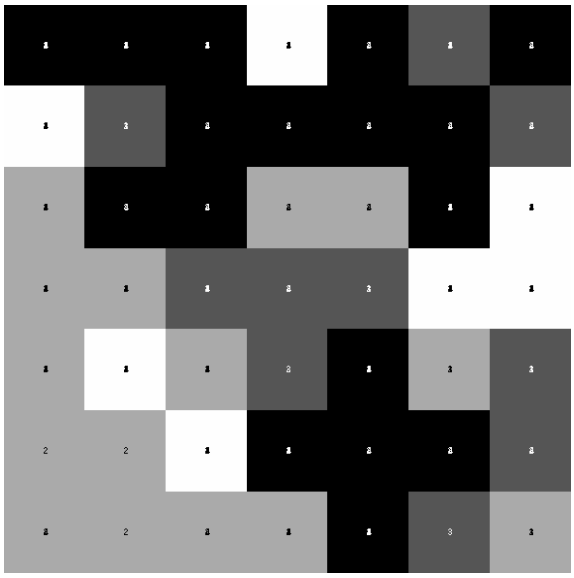

図 2(b) 構造類似性マッピング(索引層)

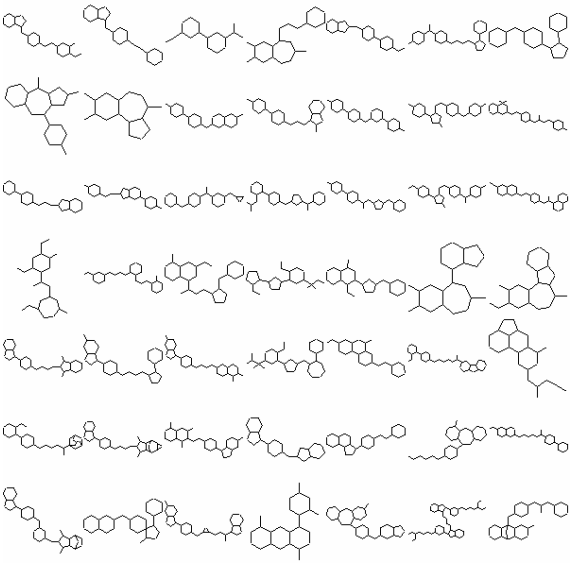

図 3 各索引層ユニットの代表構造

47	55	50	35	54	17	44
20	16	55	49	45	23	28
49	21	19	32	12	27	18
16	42	29	15	39	25	37
8	28	21	29	11	19	15
17	6	11	25	31	24	26
13	23	31	29	34	7	27

図 4 各索引層ユニットで BMU’として参照されたデータ件数

図 2(b)は同学習結果の索引層上での活性分布を示す。これにより、競合層と索引層で活性分布が類似しており、索引層が競合層ユニットの代表を成していることが確認できる。

ここで、索引層が学習に与える影響について考察する。図 3 に索引層上での薬物構造類似性マッピングの結果を示した。ここでは、簡略化のため原子、結合の情報は省略した。また、索引層ユニットの重みに最も近い薬物構造を代表として表した。これより、共通する部分構造を持つ類似した構造群が近傍に配置されていることが分かる。

また、図 4 は各索引層ユニットを BMU’とするデータ件数の分布を示したものである。1354 件のデータに対し索引層ユニット数が 49 であるため、1ユニットが平均で 27.6 件のデータに対する BMU’となることを期待した。しかし、この学習結果では各ユニットの分布が必ずしも一様ではなく、6～55 件とばらつきが見られ、そのため、競合層上での分布にも偏りが生じている。

次に、従来法、改良法の計算時間に対する RMS の比較結果を図 5 に示す。

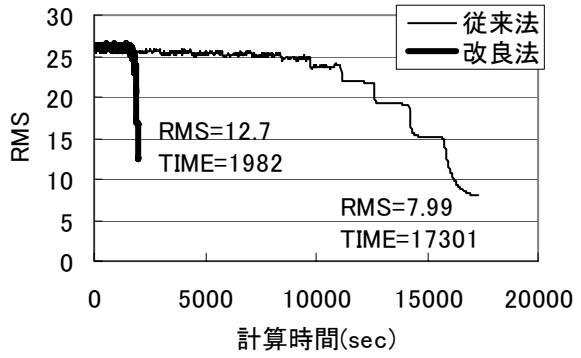

図 5 従来法・改良法の学習曲線の比較

学習に要する時間を比較すると、従来法の 17301 秒に対し、索引層を用いることで 1982 秒に短縮ができた。これは約 9 倍の高速化を実現したことになる。特に、従来法でのボトルネックとなる BMU 探索時間では、15965 秒から 497 秒にまで短縮でき、32 倍の高速化となった。

しかし、学習終了時の RMS を比較すると、従来法の 7.99 に対し改良法は 12.7 と増加した。これは先に述べたデータ分布の偏りによって、データが密集している箇所で RMS の値が増加したものと考えられる。

4. まとめ

SOM の学習のボトルネックとなる BMU 探索プロセスの改善手法として、索引層を導入する学習高速化アルゴリズムを提案し、実データを用いてその有用性を検証してきた。その結果、約 9 倍の高速化が可能であることを示した。

しかし、最適な索引層サイズはデータセットの複雑さに依存していると考えられ、ユーザーが設定しなければならないパラメータが増えるという問題点もある。

今後は、更なる高速化のためにアルゴリズムの改良し、登録件数 10000 件以上の大容量データベースに本手法を適用し実用性を確認したい。

参考文献

[1] MDL: Drug Data Report, MDL, ver. 2001.1, (2001).

[2]自己組織化マップ,コホーネン著, 得高平蔵他訳, シュプリンガー・フェアラーク東京(1996)

[3] Y. Takahashi, H. Ohoka, and Y. Ishiyama: Structural Similarity Analysis Based on Topological Fragment Spectra, In: R. Carbo and P. Mezey (Eds), Advances in Molecular Similarity 2, pp.93-104, JAI Press, Stamford CT,(1998).