1F5-01

薬物活性クラス分類へのサポートベクターマシン（SVM）の応用

Classification of Pharmacological Activity of Drugs Using Support Vector Machines

錦織 克美 Katsumi Nishikoori
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

In the present work, we investigated an applicability of Support Vector Machine (SVM) for classification of pharmacological activities of drugs. The numerical description of chemical structure of each drug was based the Topological Fragment Spectra (TFS) which was reported in our preceding work. Dopamine antagonists of 1,227 that interact with different type of receptors (D1, D2, D3 and D4) were used for training a SVM. For a prediction set of 137 drugs that were not contained in the training set, the SVM model classified 90.6% of the drugs into their own activity classes correctly.

1. はじめに

現在，医薬や農薬などの開発研究の場では既存薬物のデータを背景に，薬化合物の化学構造と生理活性など種々の作用（または特性）との間の関係を積極的に見出し，これらの情報を活用することによって新規有用物質の合理的な開発を進めようとする様々な試みが行われている．また，多大な努力によって生み出される新薬の登場の一方で，例えば大きな期待を背負って登場した医薬品に対して重篤な副作用が報告されたり，農薬においてはそのヒト健康や残留性による環境影響などが社会的な問題となっている．

本研究では，こうした化学物質のリスク評価の問題に関連し，薬物の化学構造情報のみからその活性クラスを識別・予測することをねらいとし，近年，分類学習モデルの一つとして注目を集めているサポートベクターマシン（Support Vector Machine [Vapnik 95]；以下 SVM）を適用し，実データを用いてその有用性を検証した．

2. サポートベクターマシン

SVM は 1990 年代に入ってカーネル関数の導入により実用的な非線形識別手法に拡張された．これによりチューニングを施した複雑な多層パーセプトロンに劣らない性能を有することが報告されたことから注目を集めている．

ある学習サンプル $\mathbf{x}_i \in R^d$ はクラス $y_i \in \{-1,1\}$ に属し，クラス毎に線形分離可能だとすると，その判別関数は重み $\mathbf{w}$ を用いて次式で表される．

$$f(\mathbf{x}_i) = (\mathbf{w}^T \mathbf{x}_i) + b \tag{1}$$

ここで b はバイアス項であり，$f(\mathbf{x}) = 0$ を満たす点の集合（識別面）が，$d-1$ 次元の分類超平面となる．また，この超平面が l 個全ての学習サンプルを分離可能として一意に定まるには，次の制約式を満たす必要がある．

$$y_i \cdot ((\mathbf{w}^T \mathbf{x}_i) + b) \geq 1 \qquad (i = 1,...,l) \tag{2}$$

この時，超平面に最も接近するサンプル（これをサポートベクターと呼ぶ）と超平面までの距離（マージン）は常に $1/\|\mathbf{w}\|$ となり，汎化能力の高い判別関数を推定するには，このマージンを最大化するような $\mathbf{w}$ を選べばよい．つまり線形 SVM の問題は(2)式の制約条件の下，$\|\mathbf{w}\|^2/2$ を最小化する凸二次計画問題に帰着する．しかし，一般に，実データに対しては完全な線形分離は困難な場合が多い．そこで若干の誤分類を許容し，その度合いを表す緩和変数 $\xi_i \geq 0$ と，誤分類とマージン最大化の関係を調節する係数 C を導入することによって，最小化問題は次式のように変更される．

$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \xi_i \tag{3}$$

$$\text{subject to} \quad y_i((\mathbf{w}^T \mathbf{x}_i) + b) \geq 1 - \xi_i \quad (i = 1,...,l) \tag{4}$$

これは緩和係数の和を小さく，かつ識別能力を高める $\mathbf{w}$ を求めるという問題となる．ここで係数 C を任意に定めることにより，そのトレードオフを決定できる． この最適化問題を扱いやすい形に変換するために，ラグランジュ乗数 $\alpha_i \geq 0$ を導入すると，最適解における条件として次式が導かれる．

$$\mathbf{w} = \sum_{i=0}^{l} \alpha_i y_i \mathbf{x}_i \tag{5}$$

結局 $\mathbf{w}$ は学習サンプルの展開式となり，次の双対問題に帰着できる．

$$\max \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j \tag{6}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \ (i = 1,...,l), \ \sum_{i=0}^{l} \alpha_i y_i = 0 \tag{7}$$

以上は線形分離の場合であるが，SVM はカーネルトリックにより非線形分離も可能とする．図1に示すように非線形変換 Φ を用いて高次元空間に写像し，その高次元特徴空間で線形分離を行うことで実質的な非線形分離を可能にする．ここで元の空間で定義され，Mercer の条件を満たすカーネル関数 $K(\mathbf{x}, \mathbf{x}')$ を導入することにより，写像空間での複雑な計算を避けて元の空間で直接解くことができる．一般的なカーネル関数として，次式で定義される Radial Basis Function（RBF）があり，本研究ではこれを用いている．

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2}\right) \tag{8}$$

こうして目的関数(6)式は次のように書き換えられる．

$$\max \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{9}$$

連絡先：高橋由雅，〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 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

この最適化問題を解いて得られる判別関数は，最終的に次式で表すことができる．

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \qquad (10)$$

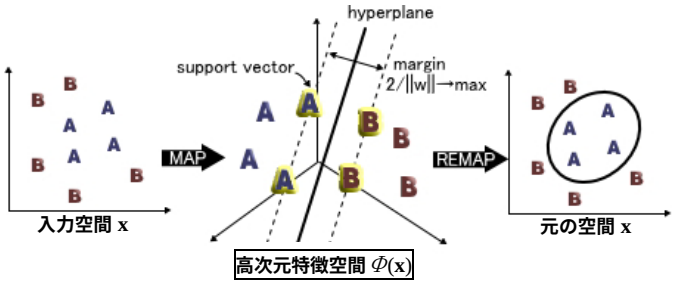

図1　非線形 SVM の例（非線形写像による分離性向上）

SVM は 2 クラス分類モデルであり，複数クラスに関する分類のためには，SVM を組み合わせる必要がある．本研究では，k クラスの分類問題を解く場合に，一般的な組み合わせ法である one-against-the-rest を利用した．クラスの判定には，k 個の SVM による k 回の判定が必要となる．判定の結果，複数のクラスが候補として残った場合，また候補が一つもない場合は，超平面からの距離によりクラスを決定した．

3. 実データによる薬物活性クラス分類

3.1 データセット

薬物活性クラス分類に際し，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，活性クラス分類に重要な構造的特徴を 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）[MDL 02]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 1364 化合物（D1：173, D2：395, D3：240, D4：556）を用いた．

化合物の構造特徴の記述子には筆者らの提案するトポロジカルフラグメントスペクトル（Topological Fragment Spectra; TFS[Takahashi 98]）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 163 次元のベクトルとして記述された．

3.2 実験

実験に際しては，SMO(Sequential Minimal Optimization [Platt 98])アルゴリズムに従って筆者らが作成した SVM ツールを用いた．カーネル関数の選択が可能であり，実験においては(8)式に示した RBF カーネルを使用している．この場合，チューニングの必要のあるパラメータは，分類の粒度を制御する分散値 σ，学習データに対するマージンの大きさと誤分類してもよいデータ数とのトレードオフである正規化値 C の 2 つである．これらの決定に関しては，先述のドーパミンアンタゴニストの内，学習用に 1227 件（90%）の化合物を選択し，残りの 137 件（10%）を検証用として，チューニングパラメータを変化させながら SVM による 4 クラス分類モデルを作成し，その分類精度を検証した上で最良の値を用いた．得られた最適なパラメータは $\sigma=10$, $C=1$ であり，以下の計算においては全てこれを使用した．

比較のため，従来手法として成果を挙げている人工ニューラルネットワーク（ANN）による分類も試みた．

3.3 結果

SVM と ANN の能力を統計的に評価･比較するため，cross validation を用いた．SVM による学習並びに予測結果をまとめて表1に示す．また，比較のため ANN による結果も合せて表 1 に示した．表 1 から解るように SVM においては，学習データによる認識率は 100%，予測集合 137 化合物に対する正答率（予測率）も 90.6%と極めて良好な結果を得ることができた．一方 ANN モデルにおいては，最も良好な結果を与えた学習時認識率は 86.5%，予測率は 81.1%であった．SVM は ANN に比べ高い予測精度を持ち，実に 9 割以上もの化合物の活性クラスを正しく予測するという結果を得た．このことは，ここでの TFS を利用した薬物活性クラス分類に対して，SVM が識別能力の高い安定した分類モデルを提供可能であることを示している．

表1　ドーパミンアンタゴニスト類の活性クラス分類

	SVM		ANN	
活性クラス	%学習	%予測	%学習	%予測
ALL	**100**	**90.6**	**87.5**	**81.1**
D1	100	87.5	76.0	70.7
D2	100	86.1	80.7	69.9
D3	100	88.3	90.9	85.8
D4	100	95.5	94.5	90.5

4. まとめ

SVM を薬物活性クラス分類のフィールドに適用し，実データを用いてその有効性を検証した．SVM による活性分類モデルは，その優秀な汎化性能から従来手法である ANN を大きく上回った．また，TFS による特徴量抽出と併用することによって，化学構造からの活性クラス識別における効果的な手順が確立できることを示唆している．一方，分類モデルにおいて対象の薬物の構造特徴がサポートベクター（他方のクラスとの境界事例）であるか否か，あるいは超平面からの距離を調べることにより，新たな知識発見への利用も期待できる．

SVM の利点は，比較的容易に実現可能であり，局所解の問題もなく，強力な性能を示せる点である．ただ，学習データ数が増えると，飛躍的に学習時間が増大する問題は SVM 学習の大きな問題のひとつであり，最適化手法の観点からは今も議論が交わされている．今後は，さらに広範な薬物構造データを用いながら引き続詳細な能力を検証したい．

参考文献

[MDL 02] MDL Drug Data Report, http://www.mdli.com/

[Platt 98] John C. Platt：Sequential Minimal Optimization：A Fast Algorithm for Training Support Vector Machines, *Microsoft Research Tech. Report* MSR-TR-98-14, Microsoft Research, (1998).

[Takahashi 98] 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).

[Vapnik 95] V.N. Vapnik : The Nature of Statistical Learning Theory, Springer, (1995).