1F1-04

# 化学構造の TFS 表現を用いた SVM による薬物活性クラス分類

## Classification of Activity Classes of Drugs Using SVM and TFS Representation of Chemical Structure

中場 優佑　　高橋 由雅
Yusuke Nakaba　　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 by the authors. For the training we employed 49615 compounds that belong to ten different activity classes. For a prediction set of 4962 drugs that were never used in the training, the SVM model classified 86.8% of the drugs into their own activity classes correctly.

## 1. はじめに

当研究室では，活性クラス識別にもとづく薬物のリスク推定/リスクレポートの観点から，構造情報の Topological Fragment Spectra (TFS) 表現を入力パターンとした人工ニューラルネットワーク (ANN) およびサポートベクターマシーン (SVM) の応用の可能性について検討を進めてきた．その結果，ANN に比し，薬物活性クラス分類における SVM の優れた安定性を明らかにした．[Takahashi 03]

本研究では，TFS を基礎とした SVM による薬物活性クラス分類の問題に対し，薬物の対象活性種を拡大し，分類/予測性能の検討を行った．

## 2. サポートベクターマシーン

SVM は主にパターン認識の分野で卓越した性能を示しているパーセプトロン型学習モデル[Vapnik 95]である．SVM の基本的な構造は単純な線形識別関数であるにも関わらず，カーネル関数とマージン最大化といった工夫を加えることにより，高い識別性能を発揮できる．SVM は 2 クラス分類問題に適用でき，その優秀な学習モデルは，化学分野における活性識別の観点からも有用であると考える．SVM による入力ベクトル $\boldsymbol{x}=(x_1, ..., x_d)^T$ の識別関数は，(1)式の通りとなる．

$$f(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + b \qquad (1)$$

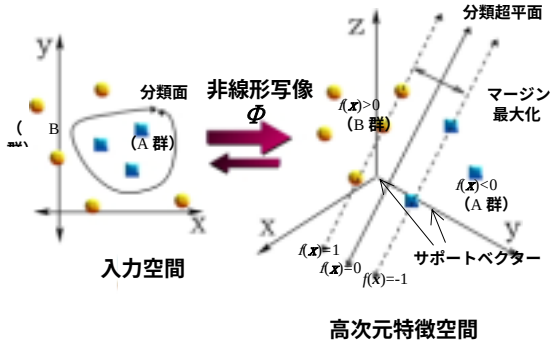


**図 1 サポートベクターマシン**

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(1)式の $f(\boldsymbol{x})=0$ を満たす点の集合 (識別面) は，$d$-1 次元の超平面 (hyperplane) となる．そして，超平面と訓練サンプルとの最小距離を最大化することで超平面を決定する．また，入力空間の高次元化とカーネルトリックにより，線形識別器は非線形に容易に拡張できる (図 1)．これは，図 1 に示すように非線形写像 $\Phi$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，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入力空間で直接解くことができる．

## 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

本研究で使用したデータセットは MDDR データベース[MDL 02]からデータ件数の多い上位 10 種類の薬理活性クラスに属する治験薬 49615 化合物 (表 1) を抽出したものである．また，今回の実験では単一の活性のクラスに属する化合物のみを選抜し，使用した．各々の化合物の構造特徴の記述には TFS [Takahashi 98]法を用いた．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たものである．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 225 次元のベクトルとして記述された．このようにして生成された TFS を入力シグナルとして，SVM による薬物活性クラスの分類/識別を試みた．

**表 1 データ件数の多い上位 10 種の治験薬**

| 活性クラス | | 化合物件数 (訓練集合/予測集合) |
|---|---|---|
| **Class** | **ALL** | **49615** |
| 1 | 抗癌剤 | 8994 |
| 2 | 抗高血圧剤 | 8467 |
| 3 | 抗アレルギー性/抗喘息物質 | 5446 |
| 4 | 認知障害改善薬 | 4859 |
| 5 | 抗関節炎薬 | 3656 |
| 6 | 抗高脂血症剤 | 4067 |
| 7 | 抗不安薬 | 4649 |
| 8 | 抗炎症薬 | 2560 |
| 9 | 神経細胞死阻害薬 | 3633 |
| 10 | 抗血小板凝集剤 | 3284 |

### 3.2 実験

活性クラスの分類には，当研究室で別途開発した SVM ツール SVMQsar[Nishikoori 03]を使用した．前述の治験薬 49615 化合物の内，予測検証用に 10%（4962 化合物）を除外し，残りの 44653 化合物を訓練集合として，SVM による多クラス分類モデルを作成した．SVM モデルの作成のためのカーネル関数には，広く良好な汎化性能を示すことが報告されている Gaussian カーネル（2）式を用いた．

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

また，ここでは分類の粒度を制御するσパラメータ，および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 つが調節可能なパラメータとして残る．これらの決定に際しては，パラメータの値を変化させながら，分類モデルを作成し，その識別/予測精度をもとに最適値を推定した．図 2 にパラメータσの変動による識別率と予測率を示す．また図 3 にパラメータ C の変動による識別率と予測率を示す．

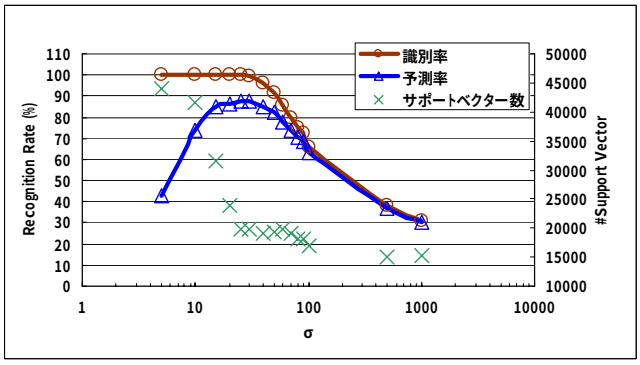


**図 2 パラメータσの変動による識別率と予測率**

図 2 より，Gaussian カーネルを用いた SVM ではσ値が小さな値のときは学習において 100%の識別率が得られ，σ値が大きくなるにつれて識別率が低下することが分かる．一方，予測率に注目すると，σには予測安定性を確保するための最適値が存在することがわかる．ここでは，σ=25 の時，識別率 99.8%，予測率 87.7%となり，最適な値を示すことが分かった．

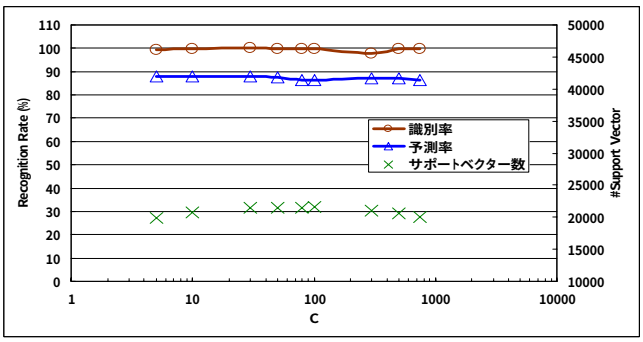


**図 3 パラメータ C の変動による識別率と予測率**

図 3 はパラメータ C の変動による学習および予測結果を示したものである．ここでは，上で得られた最適なσ値を用いて C の値のチューニングを行った．その結果，正規化値 C の値はここでの SVM 学習にはそれほど大きな影響を与えないことが示された．本研究では以下 C=5 としてすべての解析を行った．

### 3.3 結果

上記の学習パラメータを用い，SVM による学習･予測を行った．本研究で用いた SVM の学習条件を表 2 にまとめて示す．また，SVM の能力を統計的に評価するため，10-fold cross validation を用いた．その結果を表 3 に示す．表 3 から分かるように SVM において訓練集合 44653 化合物に対する正答率（識別率）は 99.4%，予測集合 4962 化合物に対する正答率（予測率）も 86.8%と良好な結果を得ることができた．

**表 2 学習条件の設定**

| | |
|---|---|
| **入力** | |
| データセット | 49615 件 |
| 訓練データ | 44653 件 (90%) |
| 予測データ | 4962 件 (10%) |
| 入力次元 | 225 の TFS パターンベクトル |
| 入力クラス数 | 10 クラス |
| **SVM 学習パラメータ** | |
| カーネル関数 | Gaussian カーネル |
| パラメータσ | 25 |
| パラメータ C | 5 |

**表 3 薬物活性データによるクラス分類結果**

| | 学習結果 | 予測結果 |
|---|---|---|
| Class | 正答数（識別率） | 正答数（予測率） |
| **平均** | **99.4%** | **86.8%** |
| 1 | 99.5% | 88.0% |
| 2 | 99.4% | 86.6% |
| 3 | 99.5% | 87.9% |
| 4 | 99.4% | 87.0% |
| 5 | 99.5% | 86.4% |
| 6 | 99.4% | 86.9% |
| 7 | 99.4% | 86.4% |
| 8 | 99.4% | 85.5% |
| 9 | 99.4% | 87.1% |
| 10 | 99.5% | 86.1% |

## 4. まとめ

以上，10 種のまったく異なる活性クラスに属する治験薬，約 5 万件に対しても SVM の良好な学習性能と高い予測安定性が示された．実用的な観点からは，何れの活性も持たないノイズデータ存在下での学習/予測についても合わせて考慮する必要があり，現在，鋭意検討を進めているところである．発表に際してはこれら結果についても合わせて報告したい．

## 参考文献

[Takahashi 03]Y. Takahashi, K. Nishikoori, S. Fujishima: Classification of Pharmacological Activity of Drugs Using Support Vector Machine, *Second International Workshop on Active Mining*, (2003) 152-158.

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

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

[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).

[Nishikoori 03] 錦織 克美，薬物活性クラス分類への Support Vector Machine の応用，豊橋技術科学大学 修士論文，(2003)