3D7-5

多クラス SVM を用いた薬物の活性予測システム

Biological Activity Prediction System for Chemicals using Multi-class SVM

河合 健太郎 藤島 悟志 高橋 由雅
Kentaro Kawai Satoshi Fujishima 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

We have investigated classification and prediction of pharmacological activity classes of chemical compounds from their chemical structures using multi-class support vector machine (SVM). For the input to the SVM, every chemical structure was represented by a multidimensional pattern vector that was obtained by the Topological Fragment Spectra (TFS) method. In this work, we adopted the "One against the Rest" method, which combine two-class SVMs to solve the multiclass classification problem. For the computational trial, we employed 98,634 compounds that belong to 100 different activity classes. The data set was divided into two groups: a training set of 88,770 compounds and a prediction set of 9,864 compounds. The SVM model was trained and validated by three-fold cross validation procedure using the training set. For the prediction set that consists of 100 classes, the best model correctly classified 80.8% of the drugs into their own active classes in average. The resulted classifier would be useful in drug-discovery and also helpful in risk estimation of adverse effects.

1. はじめに

当研究室では，これまでに薬物構造の Topological Fragment Spectra (TFS)表現を入力パターンとしたサポートベクターマシン(SVM)の応用可能性について検討を進め，薬物活性クラス分類における SVM の有効性を明らかにしてきた [Takahashi 05]. その一方で創薬研究への応用を考慮した場合， (1)より多様な活性クラスに対する有用性の検証，(2)複数の活性を示す薬物への対応，(3)どの活性も示さないノイズ化合物への対応，などが課題として残された[中場 05].

本研究では，これらの課題の解決と実用的なシステム構築の観点から，対象活性クラスの数を大幅に拡大し，100 種の活性クラスを対象とした薬物活性クラスの分類･予測システムを構築し，その予測安定性について検討を行った.

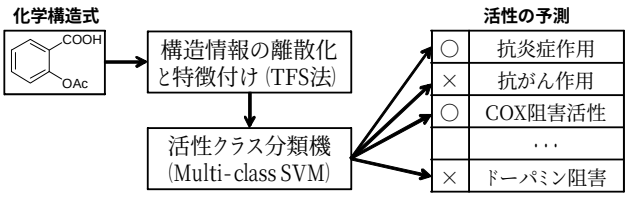

図1. 活性予測システムの概要.

2. 多クラス分類問題と SVM

SVM は本来 2 クラス分類機であるが，これを複数組み合わせることで多クラス分類が可能となる．その方法論は，これまでに幾つか知られているが，主なものは One against the Rest 方式と One against One 方式である．今回は One against the Rest 方式を用いて多クラス分類問題への対応を行った.

One against the Rest 方式は，k クラスの分類問題を解くため，注目する 1 個のクラスと残りの k－1 個のクラスとの識別面を，k 個のクラス全てについて構築する．そして予測時には，求められた k 個の 2 クラス識別関数 $fs_1,\cdots,fs_k$ の出力に従い，クラス分類を行う.

この場合，クラス識別関数 $fs_1,\cdots,fs_k$ の最大出力に従ってクラス分類を行うと，複数の活性クラスに属する薬物に対応できない．そこで本研究では「各クラス識別関数の出力値が正ならばそのクラスに属し，負であればそのクラスには属さない」事とした．つまりクラス識別関数の出力値が全て正であれば全てのクラスに属し，出力値が全て負であればどのクラスにも属さない事になる．この定義に従えば，複数の活性を示す薬物やノイズ化合物に対応する事が可能である.

今回採用した One against the Rest 方式は，各活性クラスに対して識別面を構築する必要がある．本研究では 100 クラスの分類問題を扱うことから，100 個の識別面を構築することになる.

3. 活性予測システムの構築

(1) データセット

本研究では，治験薬構造データベース (MDDR) [MDL 01]に収載された登録件数上位 100 種の活性クラスに属する薬物を用いた．ここでは分子量の大きなペプチドや天然物を除き，重原子数が 50 以下の低分子化合物 98634 化合物を対象に，分類モデルの作成と予測を試みた． 本研究で用いたデータセットの概要を表 1 に示す.

表1. 本研究で用いた薬物の活性クラスと化合物数 (抜粋).

	活性クラス名	化合物数
1	抗癌剤	10402
2	抗高血圧剤	9503
3	抗アレルギー性/抗喘息物質	8180
4	認知障害改善薬	6096
5	抗関節炎薬	5786
・・・	・・・ 中略 ・・・	・・・
97	5HT1D 作動薬	589
98	逆転写酵素阻害薬	539
99	代謝拮抗剤	554
100	ドーパミン D4 阻害薬	574
	ALL (100 クラス)	**98634**

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

本データセットのうち，複数の活性クラスを示す薬物は，過半数の 52684 件であった．また，最高で 8 種の活性クラスに属する薬物が存在した．これらの事実からも，複数の活性クラスに対する分類問題の考慮が必要であった．

(2) 構造記述子

薬物の構造表現は，TFS 法[Takahashi 98]を用いた．TFS とは化合物の構造式から部分構造を列挙し，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，その質量数で特徴付けた．結果として各化合物は，257 次元の構造特徴ベクトルとして記述された．この TFS を入力シグナルとして，SVM による薬物活性クラスの識別を試みた．

(3) 学習と評価

本データセットのうち，10%を検証用の「**外部集合**」(9864 件)とし，残りの 88770 件を用いて交差検証 (cross validation) を行った．一般には 10-fold cross validation を行う事が多いが，今回はデータ数が多い事から 3-fold cross validation を実施した．交差検証用のデータは，「**訓練集合**」と「**予測集合**」から構成される．以後，交差検証用の 3 つのデータ集合を「交差検証セット 1」，「交差検証セット 2」，「交差検証セット 3」として表現する．

SVM のカーネルは，過去の検討結果から Gaussian カーネルを利用した．分類の粒度を制御するパラメータσ，ソフトマージンに関連するパラメータ C の最適値は，グリッドサーチにより求められた．

学習結果は，Sensitivity や Specificity，サポートベクトルの数などの指標を使って評価する事ができる．「Sensitivity」は「正例」を「正例」であると正しく予測できた割合を示し，「Specificity」は「負例」を「負例」であると正しく予測できた割合を示す．サポートベクトルの数については，少ない方が汎化能力が高いと考えられる．予測システムの性能評価のため，Sensitivity，Specificity，サポートベクトル数を計算し，今回はその中の Sensitivity を基準にパラメータの最適値を探索した．

4. 結果

(1) 予測モデルの特徴とその性能

「交差検証セット1」を使って得られた活性予測モデルの予測性能を表 2 に示す．予測性能は，「交差検証セット 1」の「予測集合」に対する Sensitivity により評価した．ここでは紙面の都合上，Sensitivity が上位と下位の 5 クラスのみを記載した．活性クラスと Sensitivity との関係について特徴を纏めると，抗炎症などの疾患名を表す活性クラスは予測が困難であり，レニン阻害薬などの作用メカニズムを表す活性クラスは予測し易いという傾向が確認された．

100 クラスの平均 Sensitivity は 82.7%と高い予測性能を示し，Specificity は平均 99.2%であった．これらの結果から，本予測モデルが正例と負例を正しく識別できている事が確認できた．

表 2. 100 クラスの学習結果 (上位と下位 5 クラスのみ記載).

予測困難なクラス	Sensitivity	予測容易なクラス	Sensitivity
(局所) 抗炎症剤	60.95	カルバペネム	99.74
血管再狭窄	62.69	セファロスポリン	99.07
免疫抑制剤	67.44	キノロン	95.73
抗パーキンソン病薬	69.38	レニン阻害薬	95.52
神経保護薬	69.52	H+/K+-ATPase 阻害薬	95.19

一方，「交差検証セット 2」と「交差検証セット 3」については，「交差検証セット 1」で取得した最適パラメータをそのまま用いて予測モデルを構築した．その結果，「交差検証セット 2」の予測集合に対する平均 Sensitivity は 76.2%，「交差検証セット 3」の平均 Sensitivity は 74.4%であった．「交差検証セット 1」と比較すると，Sensitivity がやや低下する結果となった．

(2) 外部集合による予測性能の評価

「交差検証セット 1」で得た予測モデルを用いて，外部集合 (9864 件) の活性クラスを予測した．その結果，外部集合の 100 クラスの平均 Sensitivity は 80.8%であり，高い予測安定性を示した．

一方，「交差検証セット 2」および「交差検証セット 3」を使った場合，外部集合に対する Sensitivity の平均値はそれぞれ 74.8%と 72.4%であった．「交差検証セット 1」によるモデルと比較すると Sensitivity がやや低下したが，交差検証の結果から考えると，予想された範囲内であった．

5. まとめ

本研究では，(1)100 クラスの多クラス分類，(2)複数のクラスに属する薬物，(3)どのクラスにも属さない薬物，の分類問題を考慮し，薬物の活性クラス分類問題に対して SVM を適応した．「交差検証セット 1」を用いて最適な学習パラメータを探索した結果，100 クラスの平均 Sensitivity が 82.7%という良好な予測モデルを得る事ができた．また，外部集合に対する Sensitivity は 80.8%であり，予測安定性の面でも良い結果が得られた．

一方で，「交差検証セット 2」と「交差検証セット 3」を用いた場合，Sensitivity がやや低下する事が確認された．Sensitivity の低下については，「交差検証セット 1」で取得したパラメータをそのまま用いた事が大きな原因であると思われる．つまり，交差検証セット毎に最適パラメータを探索して，モデルを構築する必要があるのではないかと考えられた．

結論として，この活性予測システムは高いクラス識別能力を有しており，実際の創薬研究に対しても応用可能であると考えられる．例えば，薬物は副作用 (望まない活性) を伴う場合があり，そのリスクを早期に見出す事は医薬品の研究開発において極めて重要である．本システムにより，未知化合物の活性予測だけでなく，作用メカニズムの推定や，副作用等のリスク推定に関する応用が期待される．

参考文献

[MDL 01] MDL, MDL Drug Data Report, 2001.1, (2001).

[中場 05] 中場優佑, 高橋由雅: “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”, 2005 年度人工知能学会全国大会 (第 19 回) 論文集, 1F1-04 (2005).

[Takahashi 98] Y. Takahashi, H. Ohoka, and Y. Ishiyama, Structural Similarity Analysis Based on Topological Fragment Spectra, In “Advances in Molecular Similarity”, 2, (Eds. R. Carbo & P. Mezey), JAI Press, Greenwich, CT, 1998, pp.93-104 (1998).

[Takahashi 05] Y. Takahashi, S. Fujishima, K. Nishikoori, H. Kato, and T. Okada, Identification of dopamine D1 receptor agonists and antagonists under existing noise compounds by TFS-based ANN and SVM, *J. Comput. Chem. Jpn.*, 4, 43-48 (2005).