Classification of Pharmacological Activity of Drugs Using Support Vector Machines

Yoshimasa TAKAHASHI　Katsumi NISHIKOORI　and　Satoshi FUJISHIMA

Department of Knowledge-based Inforamtion Engineering, Toyohashi University of Technology
1-1 Hibarigaoka Tempaku-cho, Yoyohashi, 441-8580 Japan
E-mail: (taka, nishikoori, fujisima)@mis.tutkie.tut.ac.jp

Abstract 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 1,228 that interact with different type of receptors (D1, D2, D3 and D4) were used for training the SVM. For a prediction set of 136 drugs that were not contained in the training set, the SVM model classified 89.8% of the drugs into their own activity classes correctly.

Keyword Pattern Classification, SVM, Dopamine Antagonists, Topological Fragment Spectra, Structure-Activity Relationship, Risk Report

サポートベクタマシンを用いた薬物の活性クラス分類

高橋 由雅　錦織 克美　藤島 悟志

豊橋技術科学大学知識情報工学系 〒441-8580 豊橋市天伯町雲雀ヶ丘 1-1
E-mail: (taka, nishikoori, fujisima)@mis.tutkie.tut.ac.jp

あらまし　本研究では，化学構造情報にもとづく薬物活性クラス分類へのサポートベクタマシンの有用性について，実データを用いて検証した．個々の薬物の構造情報の数値的記述表現には先に筆者らが報告したトポロジカルフラグメントスペクトル法を用いた．実験に際しては市販の治験薬構造データベース（MDDR, Molecular Design Drug Data Report）より，異なる受容体（D1，D2，D3，D4）に作用する 1364 種のドーパミン拮抗薬を抽出し，用いた．全化合物の 90%（1228 化合物）を訓練集合として学習を行った後，残り 136 化合物を予測集合として実験を行ったところ 89.8%の化合物について活性クラスを正しく予測することができた．

キーワード　パタン分類，サポートベクタマシン，ドーパミン拮抗薬，トポロジカルフラグメントスペクトル，構造活性相関，リスクレポート

1. Introduction

For a half century, a lot of effort has been devoted to develop new drugs. It is true that such new drugs allow us to have better life. However, serious side effects of the drugs often have been reported and those raise a social problem. The aim of this research project is in establishing a basis of computer-aided risk report for chemicals on the basis of pattern recognition techniques and chemical similarity analysis.

The authors [1] proposed the Topological Fragment Spectral (TFS) method for a numerical vector representation of the topological structure profile of a molecule. The TFS provides us a useful tool for the evaluation of structural similarity between molecules. In our preceding work [2], we reported that an artificial neural network approach combined with input signals of the TFS allows us to successfully classify the type of activities for dopamine receptor antagonists, and it can be applied to the prediction of active class of unknown compounds. And we also suggested that similar structure searching on the basis of the TFS representation of molecules could provide us a good chance to discover some new insight or knowledge from a huge amount of data. It is clear that these approaches are quite useful for data mining and data discovery problems too.

On the other hand, in the past few years, support vector machines (SVM) have brought us a great interest in the area of machine learning due to its superior generalization ability in a wide variety of learning problems [3-5]. Support vector machine is originated from perceptron theory, but some classical problems such as multiple local minima, curse of dimensionality and over-fitting in artificial neural network little occur in this method. Here we investigate the utility of support vector machine combined with the TFS technique in classification of pharmacological activity of drugs.

2. Methods

2.1. Numerical representation of chemical structure

In the present work, to describe structural information of chemicals, Topological Fragment Spectra (TFS) method [1] was employed. The TFS is based on enumeration of all the possible substructures from a chemical structure and numerical characterization of them. A chemical structure can be regarded as a graph in terms of graph theory. For graph representation of chemical structures, hydrogen suppressed graph is often used. To get a TFS representation of a chemical structure, all the possible subgraphs with the specified number of edges are enumerated. Subsequently, every subgraph is characterized with a numerical quantity. For the characterization of a subgraph we used the overall sum of the mass numbers of the atoms corresponding to the vertexes of the subgraph. In this characterization process, suppressed hydrogen atoms are taken into account as augmented atoms. The histogram is defined as a TFS that is obtained from the frequency distribution of a set of individually characterized substructures (structural fragments) according to the value of their characterization index.

The TFS generated according to this manner is a representation of topological structural profile of a molecule. This is similar representation of mass spectra of chemicals. A schematic flow of the TFS creation is shown in Fig. 1.

The computational time required for the exhaustive enumeration of all possible substructures is often very large especially for the molecules that involve highly fused rings. To avoid such a problem the use of subspectrum was employed for the present work, in which each spectrum could be described with structural fragments up to a specified size in the number of edges (bonds).

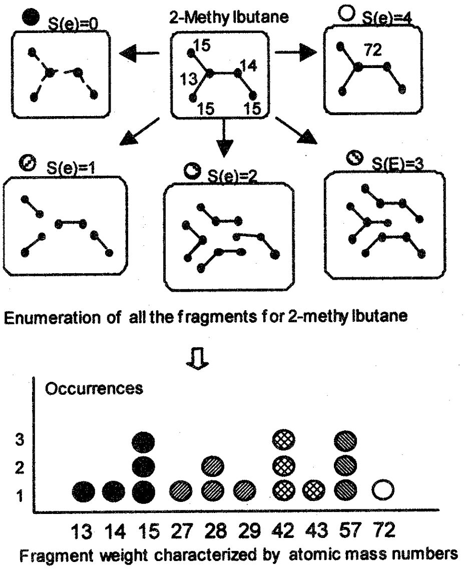
S(e)=0
2-Methy lbutane
S(e)=4
15
14
13
15
15
72
S(e)=1
S(e)=2
S(E)=3
Enumeration of all the fragments for 2-methy lbutane
Occurrences
3
2
1
13 14 15 27 28 29 42 43 57 72
Fragment weight characterized by atomic mass numbers

Figure 1. A schematic flow of TFS generation. S(e) is the number of edges (bonds) of the fragments to be generated.

2.2. Support Vector Machine

Support vector machine has been focused as a powerful nonlinear classifier due to introducing kernel function trick in the last decade. The SVM implements the following basic idea: it maps the input vectors $\mathbf{x}$ into a higher dimensional feature space $\mathbf{z}$ through some nonlinear mapping, chosen a priori. In this space, an optimal discriminant hyperplane with maximum margin is constructed. Given a training dataset represented by $\mathbf{X}(\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots,\mathbf{x}_n)$, $\mathbf{x}_i$ that are linearly separable with class labels $y_i \in \{-1,1\}, i=1,\ldots,n$, the discriminant function can be described as the following equation.

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

Where $\mathbf{w}$ is a weight vector, b is a bias. $f(\mathbf{x}_i)=0$ is the discriminant surface. The maximum margin plane can be found by minimizing

$$\|\mathbf{w}\|^2=\mathbf{w}^T\mathbf{w}=\sum_{i=1}^{d} w_i^2 \qquad (2)$$

With constraints,

$$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1 \quad (i=1,\ldots,n). \qquad (3)$$

The decision function takes the form $f(\mathbf{x})=\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x}+b)$, where sgn is simply a sign function which returns 1 for positive argument and -1 for a negative argument. This basic concept is generalized to a linearly inseparable case by introsucing slack variables ξ and minimizing the following quantity,

$$\frac{1}{2}w\cdot w+C\sum_{i=1}^{n}\xi_i \qquad (4)$$

which is subject to the constraints $y_i=(w\cdot x+b)\geq 1-\xi_i$ and $\xi_i\geq 0$. This primal information of the optimization problem reduces to the previous one for separable data when constant C is large enough.

This quadratic optimization problem with constraints can be reformulated by introducing Lagrangian multipliers α.

$$W(\alpha)=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\mathbf{x}_j-\sum_{i=1}^{n}\alpha_i \qquad (5)$$

with the constraints $0\leq\alpha_i\leq C$ and Since the training points $\mathbf{x}_i$ do appear in the final solution only via dot products, this formulation can be extended to general nonlinear functions by using the concepts of nonlinear mappings and kernels [6]. Given a mapping , $\mathbf{x}\rightarrow\phi(\mathbf{x})$, the dot product in the final space can be replaced by a Mercer kernel.

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

Here we used radial basis function as a kernel function for mapping the data into the higher dimensional space.

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

The TFS were submitted to the SVM as input feature vectors for the classification. All the SVM analyses were carried out using a computer program developed by the authors [3].

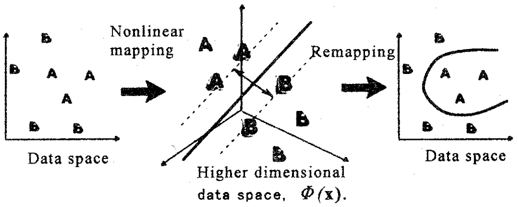

Figure 2. Illustrative scheme of nonlinear separation of two classes by the SVM with a kernel trick.

2.3. Data set

In this work we employed 1,364 dopamine antagonists that interact with four different types of receptors (D1, D2, D3 and D4). The data are a subset of MDDR [6] database. The data set was divided into two groups; training set and prediction set. The two include 1,228 compounds and 136 compounds respectively.

3. Results and Discussion

3.1. Classification and Prediction of pharmacological activity of dopamine receptor antagonists by SVM

The applicability of the SVM based on the TFS representation of chemicals was validated in discriminating active classes of pharmaceutical drugs. Here, Dopamine antagonists of 1,227 that interact with different type of receptors (Dl, D2, D3 and D4) were used for training a SVM with their TFS to classify the type of activity. The SVM model obtained classified 100% of the drugs into their own classes correctly. Then, the trained SVM model was used for the prediction of active class for unknown compounds. For 137 separately prepared in advance, the activity classes of 89.8% of the compounds were correctly predicted. All the results are summarized in Table 1. In the comparison between the results it is shown that the results for D4 antagonists are better than other classes in both cases of training and prediction. It is considered that the SVM model have got a set of well defined support vectors from the training set because the number of samples is considerably larger than those of the other classes. These results show that the SVM based on TFS provides us a very powerful tool for the classification and prediction of pharmaceutical drug activities, and to describe structural information of chemicals the TFS should be suitable as input signal to SVM in the case.

Table 1. Results of SVM analyses for 1364 dopamine antagonists

	Training		Prediction	
Class	Data	%Correct	Data	%Correct
ALL	**1228**	**100**	**136**	**89.8**
D1	156	100	17	83.3
D2	356	100	39	79.5
D3	216	100	24	91.7
D4	500	100	56	98.2

3.2. Comparison between SVM and ANN

In the preceding work, the authors reported that an artificial neural network based on the TFS gives us a successful tool for the discrimination of active classes of drugs. To evaluate the better performance of the SVM approach for the current problem, here, we tried to compare the results by SVM with those by artificial neural network (ANN). The data set of 1364 drugs used in the above section was used for the analysis as well. Ten-fold cross validation technique was employed for the computational trial. The results were summarized in Table 2.

Table 2. Comparison between SVM and ANN by ten-fold cross validation test.

	SVM		ANN	
Active class	Training	Prediction	Training	Prediction
	%correct	%correct	%correct	%correct
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

The results show that the TFS-based support vector machine could give us more successful results than TFS-base artificial neural network for the current problem.

4. Conclusions and Future Work

In this work, we investigated the utility 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. It is concluded that the TFS-based support vector machine can give us successful results for the prediction of type of activities of chemicals. Because many instances are required to establish predictive risk assessment and risk report of chemicals, it would be required to further test the performance of the support vector machine with various kinds of drugs. It would also be interesting to examine the support vectors chosen in the training phase and analyze them from the view point of structure-activity relationships of their drugs and data mining from a large scale real database of chemicals..

This work was supported by Grant-In-Aid for Scientific Research on Priority Areas (B) 13131210.

References

[1] 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)

[2] Y. Takahashi, S. Fujishima and K. Yokoe: Chemical Data Mining Based on Structural Similarity, International Workshop on Active Mining, The 2002 IEEE International Conference on Data Mining, pp.132-135, Maebashi (2002).

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

[4] C. J. Burges,. A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery **2**, 121-167 (1998).

[5] S. W. Lee and A. Verri, Eds, Support Vector Machines 2002, LNCS 2388, 2002.

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