2004－ICS－138　(29)
2004／12／6

Extended Study on Identification of Active Classes of Drugs by TFS-based Support Vector Machine with Large Data

YOSHIMASA TAKAHASHI[1], SATOSHI FUJISHIMA[1], KATSUMI NISHIKOORI[1], HIROAKI KATO[1], and TAKASHI OKADA[2]

In the preceding work, the author's investigated classification and prediction of dopamine D1 receptor agonists and antagonists with noises by TFS-based SVM. In this work, the data set was extended up to seven active classes (dopamine D1, D2 and auto-receptor agonists, D1, D2, D3, and D4 antagonists). And the noise ratio was also increased to ten times of those active compounds. Total number of compounds used in the present work is 16008 compounds for the training, and that for the prediction is 1779 compounds. The TFS-based SVM still gave us good and stable results in both classification and prediction, even in the case included ten times noise data. The model obtained resulted that it correctly predicted 97.2% of the prediction set of 1779 compounds.

大規模データセットを用いた TFS／SVM による 薬物活性クラス分類の再検討

高橋 由雅[1]　藤島 悟志[1]　錦織 克美[1]　加藤 博明[1]　岡田　孝[2]

著者らは先に、ドーパミン D1 受容体アゴニスト活性およびアンタゴニスト活性を有する化合物群に対し，ノイズデータ存在下でのクラス識別問題における SVM の有用性を示した．本研究では，活性クラスを７クラス（ドーパミン D1,D2auto-receptor アゴニスト、D1～D4 アンタゴニスト）に拡大し，合わせて大量のノイズ化合物存在を含む新たなデータセットに対して同様な検討を行った. その結果, TFS を入力シグナルとした SVM は 10 倍のノイズデータ存在下においても，依然として良好な識別結果を与えることが示された．全化合物の 90%（16008 化合物）を訓練集合として学習を行った後，残り 10%（1779 化合物）を予測集合として実験を行ったところその 97.2% (1730/1779）について活性クラスを正しく予測することができた.

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 machine learning techniques and chemical similarity analysis.

In the preceding works, the authors reported that an artificial neural network (ANN) approach combined with TFS (Topological Fragment Spectra) as input signals to ANN allowed us to successfully classify the type of activities for dopamine receptor antagonists that interact with four different types of dopamine receptors, and it could be applied to the prediction of active class of unknown compounds [1]. It was also shown that support vector machine (SVM) works for this problem much better [2].

[1] Department of Knowledge-based Information Engineering, Toyohashi University of Technology.
豊橋技術科学大学知識情報工学系
[2] Department of Informatics, School of Science and Technology, Kwansei Gakuin University.
関西学院大学理工学部情報科学科

Those were the results obtained with a set of chemicals that belong to any of typical activity classes without noise compounds. Generally, for risk estimation of drugs such as side effects, we have to treat a lot of chemicals that belong to particular active classes and much more chemicals that never belong to any of them of our interest. For this problem, classification and prediction for pharmacologically active classes of drugs under the presence of noise chemical compounds were also investigated by the TFS-based machine learning techniques [3]. The results suggest that the training by the TFS-based ANN (TFS/ANN) considerably depends on the sample size in each class. Thus the prediction ability tends to be less for the activity class that has smaller size of samples than others. On the other hand, the TFS-based SVM (TFS/SVM) works better than TFS/ANN in both of the training and the prediction. However, the data set was still small and it consisted of three classes: D1 agonists, antagonists and others. For predictive risk assessment and risk report, many instances are required in the classification modeling.

In the present work, we again investigated with further extended data sets including a large number of noise compounds from the practical viewpoint.

2. Data Set and Methods

2.1. Data set

In this work we employed 1617 drugs that interact with different dopamine receptors. They are assigned to seven different activity classes in the database: D1 receptor agonists, D2 receptor agonists, auto-receptor antagonists, and D1, D2, D3 and D4 antagonists. All the data were taken from MDL Drug Data Report (MDDR) [4] which is a structure database of investigative new drugs. In addition, ten times of those compounds (16170 compounds) were randomly chosen from the MDDR database. Those compounds are also drugs but they don't belong to any of these seven active classes. They were used as noise data against to the dopamine receptor active compounds. All the data used in the present analysis are summarized in Table 1.

Table 1. Data set used in the present work.

Activity class	Compounds
D1 receptor agonists	42
D2 receptor agonists	113
Auto-receptor Agonists	183
D1 receptor antagonists	132
D2 receptor antagonists	369
D3 receptor antagonists	223
D4 receptor antagonists	555
Others (noises)	16170
Total	17787

2.2. Numerical representation of structural features of chemicals

To describe structural features of the drug molecules, Topological Fragment Spectra (TFS) method [1] was employed. The TFS is based on enumeration of all the possible substructures from a given chemical structure and numerical characterization of them. In the way, a chemical structure can be regarded as a graph in terms of graph theory. All hydrogen atoms were suppressed in the representation.

First, all the possible subgraphs are enumerated. Subsequently, every subgraph is characterized with a numerical quantity given by a characterization scheme. For the characterization of a subgraph we used the overall sum of the mass numbers of the atoms corresponding to the vertexes of the subgraphs. 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 subgraphs (i.e. substructures or structural fragments) according to the value of their characterization.

The TFS generated along with this manner is a digital representation of topological structural profile of a drug molecule. In this work, the fragments up to the size of 5 were used. Obviously, the fragment spectrum obtained by the present method can be described as a multidimensional pattern vector. The number of dimensions of the TFS pattern description vector depends on chemical structures. Therefore, the different dimensionalities of the

spectra to be compared were adjusted to that of the highest ones by stuffing the values of 0.

2.3. Support Vector Machine

The SVM [5] 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 $\mathbf{X}(\mathbf{x_1},...,\mathbf{x_i},...,\mathbf{x_n})$, $\mathbf{x_i}$ that are linearly separable with class labels $y_i \in \{-1,1\}, i = 1,...,n$, 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, $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad (i = 1,...,n)$.

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 can be generalized to a linearly inseparable case by introsucing Lagrangian multipliers α and by using the concepts of nonlinear mappings and kernels [5]. The final decision function can be described as follows,

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (0 \leq \alpha_i \leq C). \qquad (3)$$

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

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2} \right) \qquad (4)$$

Basically, SVM is a binary classifier. For classification problem of three or more categorical data, plural discrimination functions are required for the current multi categorical classification. In this work, one-against-the-rest approach was used for the case. The TFS were used as input feature vectors to the SVM. All the SVM analyses were carried out using a computer program developed by the authors according to Platt's algorithm [6].

3. Results and Discussion

3.1. Classification and Prediction using the data set without noises

The classification and prediction abilities of the TFS/SVM were investigated for 1617 that belong to any of the seven activity classes: D1 receptor agonists, D2 receptor agonists, auto-receptor agonists, and D1, D2, D3 and D4 antagonists. The dataset was randomly divided into two groups, 90% of the data for the training set and 10% of them for the prediction set. The computational trial was carried out with a single set for each of the training and the prediction.

The TFS/SVM model that was obtained with the training set of 1455 compounds correctly classified 149 over 162 compounds of the current prediction set into their own classes in total. Then, the details of the classifications show that the recognition rate for individual class is considerably good for every class regardless of the sample size of individual classes. The results for these analyses are shown in Table 2.

Table 2. Results of the training and the prediction of active classes of dopamine receptor agonists and antagonists by TFS-based SVM

Class	Training	(%)	Prediction	(%)
D1 Ag	39/39	(100)	3/3	(100)
D2 Ag	101/101	(100)	37/38	(91.7)
AR Ag	163/164	(99.4)	19/19	(100)
D1 An	121/121	(100)	38/42	(81.8)
D2 An	333/334	(99.7)	31/35	(88.6)
D3 An	199/199	(100)	24/24	(100)
D4 An	497/497	(100)	52/58	(89.7)
Total	**1447/1455**	**(99.5)**	**149/162**	**(92.0)**

These results show that the TFS/SVM works better in the prediction too. The total prediction rates for the data sets with 50% noise, 100% noise and 300% noise are 91.4%, 93.6% and 97.8 % respectively. The results for individual classes also

are good and stable for all the classes.

It is concluded that the TFS/SVM works better in the training and it would be stable for the prediction even in the case with diverse size of samples for classes to be analyzed.

3.2. Classification and Prediction using the data set with noises

Another data set including a large number of noise compounds was prepared to test a stability of the current approach from the practical viewpoint. The dataset includes ten times noise compounds against the data set tested in the above section. The dataset was also randomly divided into two groups, 90% of the data for a training set and 10% of them for a prediction set. The computational trial was carried out with a single set for each of the training and the prediction. The results of the trials are summarized in Table 3.

Table 3. Results of the training and the prediction of active classes of dopamine receptor agonists, antagonists and noise compounds by TFS-based SVM

Class	Training	(%)	Prediction	(%)
D1 Ag	38/39	(97.4)	3/3	(100)
D2 Ag	94/101	(93.1)	11/12	(91.7)
AR Ag	157/164	(95.7)	17/19	(89.5)
D1 An	119/121	(98.3)	8/11	(72.7)
D2 An	314/334	(94.0)	24/35	(68.6)
D3 An	198/199	(99.5	22/24	(91.7)
D4 An	493/497	(99.2)	47/58	(81.0)
Noises	14542/14553	(99.9)	1598/1617	(98.8)
Total	**15955/16008**	**(99.7)**	**1730/1779**	**(97.2)**

In total, 16008 compounds were used for the training of the current TFS-based SVM model. The SVM model obtained correctly classified 99.7% of the compounds into their own active classes and the noises. Then, the SVM model trained was applied to the prediction set of 1779 compounds including noises. For the set, the prediction rates for individual classes were from 68.6% (for D2 antagonists) to 100% (D1 agonists). Totally, 1730 compounds (97.2%) were correctly classified into their own classes. All the results are summarized in Table 3.

The results show that the TFS-based support vector machine could give us successful results for the present problem.

4. Conclusion and Future Works

In this work, we investigated the utility of TFS-based Support Vector Machine (TFS/SVM) for classification of pharmacological activities of thousands of drugs. The present results show that the TFS-based SVM give us stable classification and prediction of pharmaceutical drug activities even in the case of existing relatively large noises. In the future works, further extended studies with more and more different kinds of activity classes are required to establish practical risk assessment and risk report of drugs. It would also be interesting to examine the support vector samples chosen in the training phase and analyze them from knowledge discovery on their structure-activity relationships.

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

References

[1] S. Fujishima, Y. Takahashi, Classification of Pharmacological Activity of Drugs using TFS-Based Artificial Neural Network, *J. Chem. Inf. Comput. Sci.*, Vol.**44**, 1006-1009 (2004).

[2] 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

[3] Y. Takahashi, S. Fujishima, K. Nishikoori, H. Kato, and T. Okada, Identification of Activity Classes of Drugs under Existing Noise Compounds by ANN and SVM, *Third International Workshop on Active Mining*, pp.93-101 (2004).

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

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

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