

Feature Selection and Activity Prediction in Chinese Medicine Research Using a Hybrid Model GA-SVM

Zhang Shaojie, Zhao Yannan, Song Yixu, Wang Jiabin, Yang Zehong
Computer Science & Technology Department
Tsinghua University
FIT1-509, Tsinghua University, Beijing, China

Abstract - A new hybrid method called GA-SVM was proposed which combines GA (Genetic Algorithm) as a feature selection model and SVM (Support Vector Machine) as a regression model. With some modifications to the general GA and SVM models, this method can implement feature selection and activity prediction simultaneously, and its performance can be improved. Two experiments are carried out which indicate its better performance than traditional models such as BP net in small sample sets. Important features could be selected by GA-SVM which are validated by MLR. By applying the new model in Chinese medicine research, the QSAR of COX-2 and PGE2 inhibitors were found out and useful conclusions to instruct practical pharmaceuticals were drawn.

Key Words: SVM (Support Vector Machine), GA (Genetic Algorithm), Chinese medicine research, QSAR(Quantitative Structure-Activity Relationship)

1.0 Introduction

Chinese medicine has a long history and bearing very good effect in dealing with certain diseases, however, in the word of theory, this domain remains almost a mystery castle because it was built on the base of experience instead of theory. In this paper we investigated the relationship between the chemical structure and the activity for thermoregulation of the components of a certain Chinese medicine, 'Guizhi Decoction', which is proved to be effective in curing fever and other inflammation caused by flu or other diseases. COX-1 and COX-2 are two cyclooxygenase (COX) isoformates [1], whose principal pharmacological effect is inhibiting prostaglandin E2 (PGE2) synthesis, which directly causes the temperature rise of human body. COX-2 is an inducible enzyme which is mainly produced during inflammation processes [2]. The study of selective inhibition of COX-2 led to a new class of anti-inflammatory, analgesic and antipyretic drugs with significantly reduced side effects. [3] The process how COX-2 affects the production of PGE2, which is a reason of temperature rise, is show in Fig. 1. In order to find out how the components in 'Guizhi Decoction' control human body temperature, their influence on the activity of COX-2 and the secretion of PGE-2 should be investigated first. A large number of studies in finding selective COX-2 inhibitors in the domain of biology have been carried out [4-6], but none of them pointed out the quantitative relationship, which is generally called Quantitative Structure-Activity Relationship (QSAR) between the structure parameters of COX-2 and PGE2 inhibitors and their activity. In this article we aim at finding the quantitative relationship and predicting the activity according to the given structural formula of certain group of compounds which have similar structures.

In the reported four types of selective COX-2 inhibitors [3], the skeleton structure shown in Fig. 2(a) can be extracted, and the skeleton of PGE2 inhibitors extracted from the biology experiments in our project is shown in

Fig. 2(b). As is suggested in some molecular biology or chemistry research studies[3, 15], ten factors (the hydrophobic parameter, the constant of gram-molecular refraction, the solid and electric parameters of substituent X1, X2, Y, Z in Fig. 2(a) or X, Y, Z, W in Fig. 2(b) respectively) might be important in influencing the activity. In this article the ten features are labeled as F1~F10. Only 98 compounds with the same skeleton structure as showed in Fig. 2(a) and only 18 for Fig. 2(b) can be acquired, making this problem a small-sample-set one.

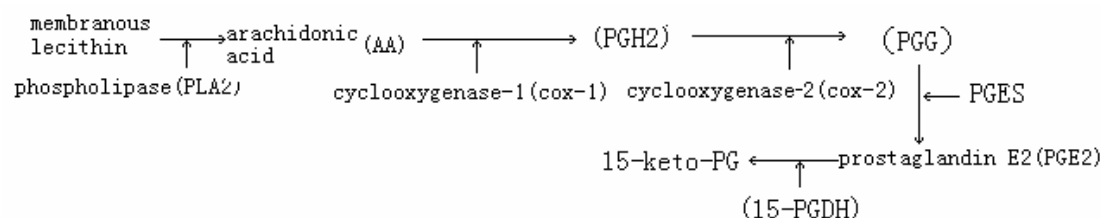


Fig.1 pathway of thermoregulation

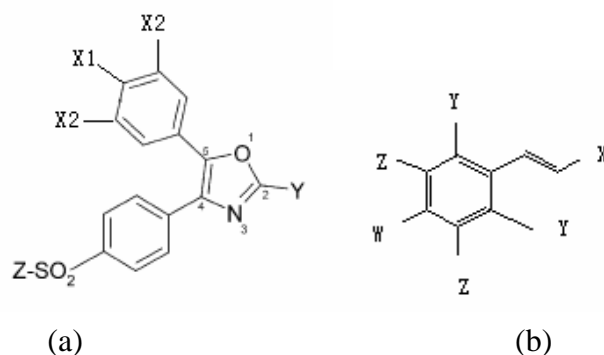


Fig. 2 skeleton structure of COX-2 and PGE2 inhibitors

It's a huge project for experts to determine the QSAR from the values of the ten features mentioned above while it becomes easier for computer. The mainly methods for medicine modeling and predicting are PCA (Primary Component Analysis), neural network (e.g. BP net), fuzzy clustering, SVM (Support Vector Machine), etc [7-12], among which SVM has been extensively used as a classification tool with satisfying success in a variety of areas [13-14], and it has both quantitative and qualitative advantages and outperforms other methods in classification for small gene subsets.[11] As is known, in many supervised learning problems, feature selection is important, and for SVM, it also performs badly when there are many irrelevant features [12]. In order to improve its performance, suitable feature selection algorithm, such as MLR (Multiple Linear Regression), GA, should be adopted. In this article a method called GA-SVM is proposed to predict the activity of COX-2 and PGE2 inhibitors and to select the essential factors which affect the activity at the same time. The comparison between the performance of GA-SVM and that of traditional BP net will be made and its advantage will be discussed.

The rest of this paper is organized as follows. In section 2 the main idea of GA-SVM is proposed, in section 3 two experiments are carried out and the results are discussed. Finally, some conclusions are drawn according to the results of experiments.

2.0 The main idea of GA-SVM

As is mentioned above, SVM has many advantages in dealing with problems in different areas and different sizes, and it is proved to be effective in small gene subsets [11]. But irrelevant features might destroy its performance, therefore feature selection will play an important role in such problems.

2.1 Theory of Support Vector Machine

Support Vector Machines map a vector $x \in R^n$ into a high (possibly infinite) dimensional space and construct an optimal hyper-plane in this space [12]. Different mappings construct different SVMs.

The mapping $\Phi(x)$ is performed by a kernel function $K(x_1, x_2)$ which defines an inner product in H. The decision function given by an SVM is thus:

$$f(x) = w \bullet \Phi(x) + b = \sum_i a_i^0 y_i K(x_i, x) + b \quad (1)$$

The optimal hyperplane is the one with the maximal distance (in H space) to the closest image $\Phi(x_i)$ from the training data. This reduces to maximizing the following optimization problem:

$$W^2(a) = \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i,j=1}^l a_i a_j y_i y_j K(x_i, x_j) \quad (2)$$

under constraints $\sum_{i=1}^l a_i y_i = 0$ and $a_i \geq 0, i = 1, L, l$. For the non-separable case one can quadratically penalize errors with the modified kernel $K \leftarrow K + \frac{1}{\lambda} I$ where I is the identity matrix and λ is a constant penalizing the training errors. In GA-SVM we choose RBF (Radial Basis Function) function as the kernel function.

2.2 Feature Selection Method – GA

Feature selection is an essential part for classification or regression to reduce computation time and to improve the performance. Several methods have been reported recently, which are divided into three groups: Filter [16], Wrapper [17] and Hybrid [18]. Among these methods PCA and MLR are classic Filter method which can be used independently while GA is a newly-proposed Wrapper method which is usually combined with classification or regression algorithms. The principle of GA is simple and can be described as follows:

- (1) Encoding of the problem in a binary string.
- (2) Random generation of a population. This one includes a genetic pool representing a group of possible solutions.
- (3) Reckoning of a fitness value for each subject which is called chromosome. It will directly depend on the distance to the optimum.
- (4) Selection of the chromosomes that will mate according to their share in the population global fitness.
- (5) Genomes crossover and mutations.
- (6) And then start again from (3).

In GA-SVM the problem is encoded into a 10-bit binary string, and each bit represents the selection of the corresponding feature, i.e. '1' indicates that the corresponding feature is selected while '0' indicates the opposite.

The fitness value of each chromosome is defined as formula (3) where MSE (Mean Square Error) is defined in formula (4). Here m is the number of subsets for cross validation (m-fold cross-validation) and n is the number of samples in each subset. In our experiment m is specified to be 3.

$$fitness(X) = \frac{1}{mse(X)} \quad (3)$$

$$mse(X) = \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{j=1}^n \left(\frac{y_{ij} - predict(x_{ij})}{\max y - \min y} \right)^2 \quad (4)$$

2.3 Some Modifications in GA-SVM

The main idea of GA-SVM is to combine GA with SVM so that feature selection and activity prediction can be achieved simultaneously. To reduce the computation time and to improve the performance of GA-SVM we made some modifications to the general model.

1. Speeding up convergence

When the fitness value of the best chromosome in a generation is only a little more than that of the second best one, the better one might not be easily selected and the convergence might become too slow. To speed up convergence we re-define the fitness function as formula (5) where MIN is the minimum of all the MSEs in a generation.

2. Reducing the effect of noise data

There might be some noise points in the test set so that the predicting errors at these points might be much larger than errors of other samples. When this happens, the MSE will be determined by these large errors and cannot correctly represent the general predicting accuracy. In order to reduce the influence of noise data, we define another mean square error MSE' as in formula (6), where j_{\max} is the index of the sample which has the maximal error.

$$fitness(X) = \frac{1}{mse(X) - \frac{3}{4} MIN} \quad (5)$$

$$MSE'(X) = \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{j=1, j \neq j_{\max}}^n \left(\frac{y_{ij} - predict(x_{ij})}{\max y - \min y} \right)^2 \quad (6)$$

3.0 Experiments and Discussion

Two experiments will be carried out as mentioned above, in each experiment we adopt GA-SVM and BP net, and then we use MLR to select the features so as to validate the results of GA.

3.1 Experiment on COX-2 inhibitors

This experiment is implemented on the 98-data set from [3], which is separated into three subsets, each time we choose two of the subsets as training subsets and one as testing subset, i.e., 3-fold cross validation. The statistical results of GA-SVM and BP net are given in Tab. 1 and the prediction results on one testing subset of the two methods are shown in Fig. 3. In GA-SVM the features are selected using GA and in BP net the features are selected according to the analysis result of MLR. The parameters in GA-SVM and BP net are optimized by the Steepest Descent Algorithm.

Tab. 1 Comparison of the result of GA-SVM and BP net in COX-2 experiment

	Minimal MSE	Features selected
GA-SVM	0.0490706	1 1 1 1 0 0 1 1 0 1
BP net	0.0653902	1 1 0 1 0 0 1 1 0 1

From Tab. 1, it's easy to find that GA-SVM has a higher accuracy than BP net to predict the activity of

COX-2 inhibitors, and the results of feature selection according to two methods are almost the same, the only different feature is the electric parameter on X substituent which needs to be validated in future work, the other 6 features (i.e. the hydrophobic parameter, the constant of gram-molecular refraction, the solid parameter of substituent Y and the electric parameters of substituent X1, Y, Z in Fig. 2(a)) are suggested to be important to the activity of COX-2 inhibitors.

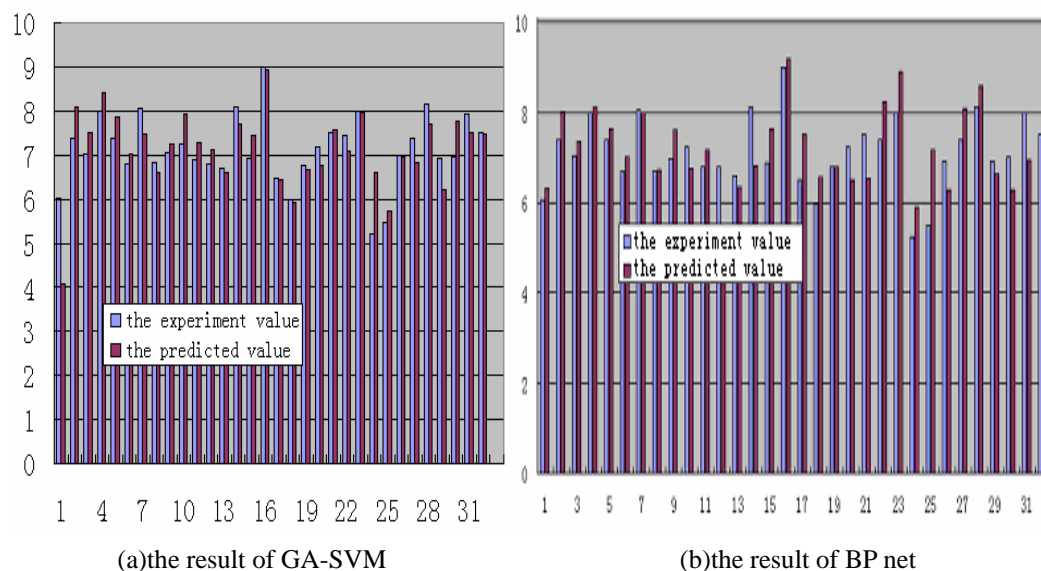


Fig. 3 results on a subset of GA-SVM and BP net in COX-2 experiment

3.2 Experiment on PGE2 inhibitors

The sample set of this experiment comes from practical biology experiment which has a very long experiment cycle and is hard to succeed, so the data is very limited—18 samples. This experiment is carried out with the purpose of providing more useful instruction to plan biology experiments and reducing the expense of experiments. We adopt GA-SVM while we don't use BP net in this experiment because the data is too limited to build BP model. The performance of prediction in GA-SVM is illustrated in Fig. 4 and the statistic value is given in Tab. 2.

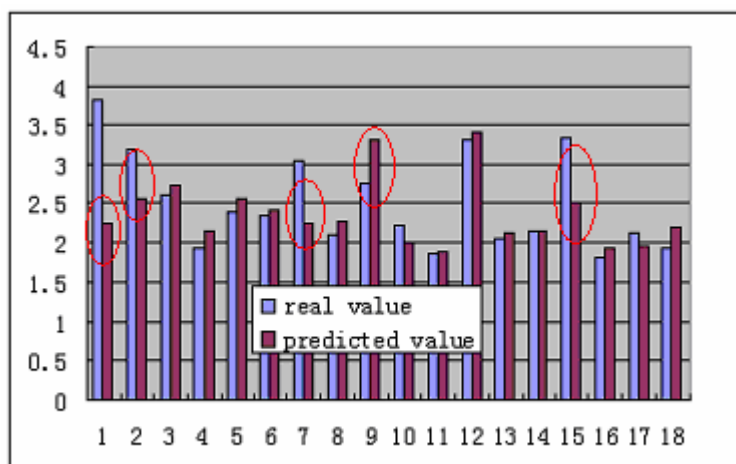


Fig. 4 result using GA-SVM in PGE2 experiment

Tab. 2 The result of GA-SVM in PGE2 experiment

	MSE'	Feature selected
GA-SVM	0.0322548	0 0 0 1 0 0 0 0 0 0
GA-SVM	0.037841	1 1 1 1 0 0 1 1 0 0

The result in Tab. 2 shows that the solid parameter on substituent X is the most important feature for the activity of PGE2 inhibitors. The result in Fig. 4 shows that GA-SVM doesn't perform well on the five samples marked above in Fig. 4, and we find that three of these five samples have a substituent different from all the other 17 samples after reviewing their structural formula, which might be a reason of bad performance.

Furthermore, we adopt MLR to analyze these data in order to validate the result of feature selection in GA-SVM and get the result in Tab.3. We find that the feature F4 has the maximal contribution value which is a criterion of feature importance to influence the activity. Comparing the result of MLR with that of GA-SVM we find that F4 (the solid parameter of substituent X in Fig. 2) is the most important feature in the 10 features, which indicates that different substituents should be tried in the position of X in Fig. 2 in the future biology experiments to find more active drugs.

Tab. 3 The analysis result of MLR

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Contribution	2.66	0.41	0.15	17.10	0	0.34	0	0.42	1.42	0

4.0 Conclusion and Future Work

In order to find the QSAR of COX-2 and PGE2 inhibitors, in this paper we propose a new hybrid model GA-SVM which combines GA and SVM to implement feature selection and activity prediction simultaneously, and make some improvement to reduce the computation time and improve the performance. Two relevant experiments are carried out and the predicting results of them are compared with the results of BP net and furthermore the feature selection results are validated by the results of MLR. GA-SVM is proved to be effective in selecting features and predicting the activity even if the sample set is very small.

In the first experiment 6 important features are selected and the prediction accuracy of GA-SVM is higher than that of BP net. The most important feature F4 (the solid parameter of the X substituent in Fig. 2(b)) is selected in the second experiment, and this result might provide useful instruction in pharmaceuticals.

The algorithm GA-SVM studied in this paper is not for Chinese medicine research's exclusive use, but feasible for many other similar areas in artificial science, such as Machine Learning, Natural Language Processing, with a little modification to this model.

Future work should be developed if more data of PGE2 inhibitors can be obtained and deeper research should be done to find out the other important features besides F4. With the help of modeling and predicting, biology experiments can be planned more purposively, and then the experiment cycle and the expense might be reduced remarkably, and moreover, some artificial compounds which have higher activity might be brought out. The research in this paper might have great significance in the application area of Chinese medicine if the quantitative relationship between structure and activity is made thoroughly clear which needs further study.

Acknowledge

*The work of authors is supported by National Natural Science Foundation of China (No. 90209006)

References

- [1] J.Y. Fu, J.L. Masferrer, K. Seibert, A. Raz, P. Needleman, The Induction and Suppression of Prostaglandin H2 Synthase (Cyclooxygenase) in Human Monocytes, *J. Biol. Chem.* 1990, 265, 16737-16740.
- [2] F. Julémont, L. Xavier, C. Michaux, J. Damas, C. Charlier, F. Durant, B. Pirotte, J.M. Dogné, Spectral and Crystallographic Study of Pyridinic Analogues of Nimesulide: Determination of the Active Form of Methanesulfonamides as COX-2 Selective Inhibitors, *J. Med. Chem.* 2002, 45, 5182-5185
- [3] Haifeng Chen, Qiang Li, Xiaojun Yao, BoTao Fan, Shengang Yuan, A. Panaye, J. P. Doucet, CoMFA/CoMSIA/HQSAR and Docking Study of the Binding Mode of Selective Cyclooxygenase (COX-2) Inhibitors, *QSAR & Combinatorial Science*, 2004, 23, 36-55
- [4] R.G. Kurumbail, A.M. Stevens, J.K. Gierse, J.J. McDonald, R.A. Stegeman, J.Y. Pak, D. Gildehaus, J.M. Miyashiro, T.D. Penning, K. Seibert, P.C. Isakson, W.C. Stallings, Structural Basis for Selective Inhibition of Cyclooxygenase-2 by Anti-inflammatory Agents, *Nature* 1996, 384, 644-648.
- [5] Y. Leblanc, W.C. Black, C.C. Chan, S. Charleson, D. Delorme, D. Denis, J.Y. Gauthier, E.L. Grimm, R. Gardon, D. Guay, P. Hamel, S. Kargman, C.K. Lau, J. Mancini, M. Ouellet, D. Percival, P. Roy, K. Skorey, P. Tagari, P. Vickers, E. Wong, L. Xu, P. Prasit, Synthesis and Biological Evaluation of Both Enantiomers of L-761,000 as Inhibitors of Cyclo-oxygenase-1 and 2, *Biorg. Med. Chem. Lett.* 1996, 6, 731-736.
- [6] T. Klein, R.M. Nusing, J. Pfeilschifter, V. Ullrich, Selective Inhibition of Cyclooxygenase-2, *Biochem. Pharmacol.* 1994, 48, 1605-1610.
- [7] Qiao Yanjiang, Wang Xi, et al. Application of Artificial Neural Networks to the Feature Extraction in Chemical pattern Recognition of the Traditional Chinese Medicine Venenum Bufonis, *Acta pharmaceutica Sinica* 1995, 30(9), 698-701 (in Chinese).
- [8] Wu Yuzhu, Luo Xu, et al. Quality Assessment of the Chinese Traditional Medicine Rhubarb by Chemical Pattern Recognition. *Acta pharmaceutica Sinica* 1991, 26(2), 132-138 (in Chinese).
- [9] Wang Xiukun, Li Jiashi, et al. Studies On radix Sophorae Flayescentis Quality by Chemical Pattern Recognition, *China Journal of Chinese Materia Medica* 1996, 21(4) (in Chinese).
- [10] Liu Qianguang, Chen Zhanguo, et al. Studies on the Quality of American Ginseng(Panax Quinque folius) by Chemical Pattern Recognition, *Chinese Traditional and Herbal Drugs* 1999, 30(11) (in Chinese).
- [11] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines, *J. Machine Learning*, 2000, 46(13), 389-422
- [12] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for support vector machines. *Advances in Neural Information Processing Systems*, 2000, 33
- [13] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio. Image representations for object detection using kernel classifiers. In *Asian Conference on Computer Vision*, 2000.
- [14] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio. Support vector machine classification of micro array data. *AI Memo 1677*, Massachusetts Institute of Technology, 1999.
- [15] Y. C. Martin, Quantitative drug design (Wang Erhua). Beijing: People's Hygiene Press, 1981, 109-128
- [16] M. Dash, K. Choi, P. Scheuermann, and H. Liu, Feature Selection for Clustering-a Filter Solution, *Proc. Second Int'l Conf. Data Mining*, 2002, 115-122.
- [17] R. Caruana and D. Freitag, Greedy Attribute Selection, *Proc. 11th Int'l Conf. Machine Learning*, 1994, 28-36.
- [18] S. Das, Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection, *Proc. 18th Int'l Conf. Machine Learning*, 2001, 74-81.