Chemistry and Materials Research
ISSN 2224- 3224 (Print) ISSN 2225- 0956 (Online)
Vol.3 No.9, 2013

Application of support vector machines for prediction of anti-HIV activity of TIBO Derivatives.

Rachid Darnag [1]* Brahim Minaoui [2] Mohamed Fakir [3]

1. Département de Physique, Laboratoire de Traitement de l'Information et de Télécommunication, Faculté des
Sciences et Technique BP 523 Université Sultan Moulay Slimane, Béni-Mellal, Morocco,

2. Département de Physique, Laboratoire de Traitement de l'Information et de Télécommunication, Faculté des
Sciences et Technique BP 523 Université Sultan Moulay Slimane, Béni-Mellal, Morocco

3. Département de l'Informatique, Faculté des Sciences et Technique BP 523 Université Sultan Moulay Slimane, Béni-Mellal, Morocco

* E-mail of the corresponding author:r.darnag@gmail.com

Abstract

The performance and predictive power of support vector machines (SVM) for regression problems in quantitative structure-activity relationship were investigated. The SVM results are superior to those obtained by artificial neural network and multiple linear regression. These results indicate that the SVM model with the kernel radial basis function can be used as an alternative tool for regression problems in quantitative structure-activity relationship.

Keywords: support vector machine (SVM); ANN; QSAR

Introduction

The Quantitative Structure-Activity Relationship (QSAR) approach became very useful and largely widespread for the prediction of anti-HIV activity, particularly in drug design. This approach is based on the assumption that the variations in the properties of the compounds can be correlated with changes in their molecular features, characterized by the so-called "molecular descriptors". A certain number of computational techniques have been found useful for the establishment of the relationships between molecular structures and anti-HIV activity such as Multiple Linear Regression (MLR), Partial Least Square regression (PLS) and different types of Artificial Neural Networks (ANN) [1]. For these methods, linear model is limited for a complex biological system. The flexibility of ANN enables them to discover more complex nonlinear relationships in experimental data. However, these neural systems have some problems inherent to its architecture such as over training, over fitting and network optimization. Other problems with the use of ANN concern the reproducibility of results, due largely to random initialization of the networks and variation of stopping criteria. Owing to the reasons mentioned above, there is a growing interest in the application of SVM in the field of QSAR. The SVM is a relatively recent approach introduced by Vapnik [2] and Burges [3] in order to solve supervised classification and regression problems, or more colloquially learning from examples.

SVM have strong theoretical foundations and excellent empirical successes. They have been applied to tasks such as handwritten digit recognition, object recognition, text classification, cancer diagnosis, identification of

HIV protease cleavages sites. They have also been applied to the prediction of retention index of protein and the investigation of QSAR studies.

Methodology

Support vector machines

A SVM is a supervised learning technique from the field of machine learning applicable to both classification and regression. SVM developed by Cortes and Vapnik [4], as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance.

Originally it was worked out for linear two-class classification with margin, where margin means the minimal distance from the separating hyper plane to the closest data points. SVM learning machine seeks for an optimal separating hyper-plane, where the margin is maximal. An important and unique feature of this approach is that the solution is based only on those data points, which are at the margin. These points are called support vectors. The linear SVM can be extended to nonlinear one when first the problem is transformed into a feature space using a set of nonlinear basis functions. In the feature space which can be very high dimensional, the data points can be separated linearly. An important advantage of the SVM is that it is not necessary to implement this transformation and to determine the separating hyper-plane in the possibly very-high dimensional feature space, instead a kernel representation can be used, where the solution is written as a weighted sum of the values of certain kernel function evaluated at the support vectors.

All SVM model in our present study were implemented using the software Libsvm that is an efficient software for classification and regression developed by Chin-Chang and Chih-Jen Lin [5].

Artificial neural networks

ANN are artificial systems simulating the function of the human brain. Three components constitute a neural network: the processing elements or nodes, the topology of the connections between the nodes, and the learning rule by which new information is encoded in the network. While there are a number of different ANN models, the most frequently used type of ANN in QSAR is the three-layered feed-forward network [6]. In this type of networks, the neurons are arranged in layers (an input layer, one hidden layer and an output layer). Each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer.

According to the supervised learning adopted, the networks are taught by giving them examples of input patterns and the corresponding target outputs. Through an iterative process, the connection weights are modified until the network gives the desired results for the training set of data. A back-propagation algorithm is used to minimize the error function. This algorithm has been described previously with a simple example of application [7] and a detail of this algorithm is given elsewhere [8].

Data set

A series of 82 4,5,6,7-Tetrahydro-5-methylimidazo[4,5,1-jk][1,4]benzodiazepin-2(1H)-ones (TIBO) molecules [9] were taken under consideration in this study. All the molecules studied had the same parent skeleton. The structures and anti-HIV-1 activities of these compounds were described previously [9]. The anti-

HIV activity of the compounds has been expressed by the compound's ability to protect MT-4 cells against the cytopathic effect of the virus. The concentration of the compound leading to 50% effect has been measured and expressed as IC_{50}. The logarithm of the inverse of this parameter has been used as biological end points (log $1/IC_{50}$) in the QSAR studies.

In our study, each molecule was described by 4 descriptors, which are given by Garg et al. [9]. These descriptors characterize the hydrophobic, the steric and the electronic aspects, respectively:

logP: the calculated octanol/water partition coefficient of the molecule

B1(8-x): Verloop's sterimol parameter (width arameter of the X substituent at the position 8)

I_R = 1 if R = 3,3-dimethyallyl and I_R = 0 for others

I_Z = 1 if Z = Oxygen and I_Z = 0 if Z =Sulphur

82×5 matrix was obtained. 82 represents the number of the molecules and 5 represents the dependent variable

(log $1/IC_{50}$) and the four independent variables (the 4 mentioned descriptors).

Results and Discussion

Two different sessions have been achieved: computation and prediction. The first one was aimed at selecting the parameters of the SVM. The second one was aimed at determining the predictive ability of the SVM.

Computation

The performances of SVM depend on the combination of several parameters. They are capacity parameter C, ε of ε-insensitive loss function and the corresponding parameters of the kernel function. C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. However, Wang et al. [10] indicated that prediction error was scarcely influenced by C. In order to make the learning process stable, a large value should be set up for C.

The selection of the kernel function and corresponding parameters is very important because they implicitly define the distribution of the training set samples in the high dimensional feature space and also the linear model constructed in the feature space. There are four possible choices of kernel functions available in the LibSVM package i.e., linear, polynomial, radial basis function, and sigmoid function. For regression tasks, the radial basis function kernel is often used because of its effectiveness and speed in training process. In this work the form of the radial basis function used is:

$$\exp(-\gamma|\mu-\nu|^2)$$

where γ is a parameter of the kernel, μ and ν are the two independent variables.

The γ of the kernel function greatly affect the number of support vectors, which has a close relation with the performance of the SVM and training time. Many support vectors could produce over fitting and increase the training time. In addition, γ controls the amplitude of the RBF function, and therefore, controls the generalization ability of SVM.

The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ε, there is the practical consideration of the number of resulting support vectors. ε-insensitivity prevents the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value of ε is critical from theory.

To determine the optimal parameters, a grid search was performed based on leave-one-out cross validation on the original data set for all parameter combinations of C from 100 to 1000 with incremental steps of 50, γ ranging from 2 to 3.2 with incremental steps of 0.1 and ε from 0.04 to 0.16 with incremental steps of 0.01. The optimal values of C, γ and ε are 500, 2.8 and 0.09, respectively.

Prediction

After determining the optimum value of C, γ and ε, we turned to the most important predictive aspect of SVM: the prediction of the anti-HIV activity of new molecules. Cross-Validation is an approach particularly well adapted to the estimation of that ability. It consists in dividing a set of examples into N subsets. Each SVM model is trained on N-1 subsets and its performance tested on the remaining subset, which acts like a test set. This process is repeated for all the N subsets. When the subsets contain only one element, the process mentioned above is then called the LOO procedure. The drawback of such an approach is its greater computational demands. In this paper the LOO procedure was used to evaluate the predictive ability of the SVM.

In our previous study [11] ANN and MLR methods were applied to the same data set and the same four molecular descriptors. Nine ANN architectures of 4-x-1 (x = 5-13, x represents the number of hidden neurons) have been tested. The results of QSAR done by these ANN architectures, by the MLR analysis and by the SVM method are given in Table 1. The quality of the fitting is estimated by the root mean square error (RMSE) and by the statistical parameter q^2 [12].

As it can be seen in table 1, high correlation coefficient (q^2 = 0.96) and low RMSE (0.212) have been obtained by means of the SVM. According to this table, it is clear that the performance of SVM is better than those obtained by ANN and MLR techniques. Indeed, in every case, the SVM's correlation coefficient is greater and its standard deviation is lower than those of the ANN and MLR.

The plot of predicted versus experimental values for data set is shown in Fig. 1. This figure shows that the $\log(1/IC_{50})$ values predicted by the SVM are very close to the experimental ones.

Table 1: Predictive ability of SVM, ANN and MLR

Method	q^2	RMSE
SVM	0.960	0.212
4-5-1	0.910	0.432
4-6-1	0.924	0.395
4-7-1	0.925	0.394
4-8-1	0.924	0.395
4-9-1	0.923	0.399
4-10-1	0.922	0.401
4-11-1	0.927	0.388
4-12-1	0.923	0.399
4-13-1	0.928	0.387
MLR	0.861	0.550

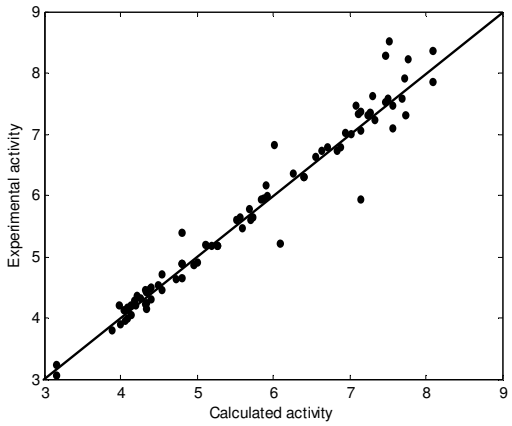

Figure 1. log ($1/IC_{50}$) observed experimentally versus log ($1/IC_{50}$) predicted by SVM.

Conclusion

The support vector machine was used to develop a QSAR model for the prediction of the anti-HIV-1 activity of TIBO derivatives. The results obtained show that the SVM technique was able to establish a satisfactory relationship between the molecular descriptors and the anti-HIV-1 activity. This technique is able to extract necessary information from examples, without explicitly incorporating rules into the SVM, in order to develop a reliable QSAR. The SVM approach would seem to have a great potential for determining quantitative structure-anti-HIV-1 activity relationships and as such be a valuable tool for the chemist.

REFERENCES

[1]L. Douali, D. Villemin, and D. Cherqaoui, "Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives," Curr. Pharm. Des, vol. 9, pp. 1817-1826, August 2003.

[2]V. N. Vapnik, "The Nature of Statistical Learning Theory, (Eds) Springer," Berlin, 1995.

[3] J. C. Burges, "A tutorial on support vector machines for pattern recognition" Data Min. Know. Discovery, vol. 2, pp. 121–167, 1998.

[4] C. Cortes, and V. Vapnik, "Support vector networks," Mach. Learn, vol. 20, pp. 273-297, September 1995.

[5] C. C. Chang, and C. J. Lin, LIBSVM-A Library for support vector machine.
http://www/csie.edu/tw/cjlin/libs/libsvm

[6] J. Zupan, and J. Gasteiger (Eds.), "Neural Networks for Chemists. An Introduction," VCH Publishers, Weinheim, 1993.

[7] D. Cherqaoui, and D. Villemin, "Use of neural network to determine boiling point of alkanes," J. Chem. Soc. Faraday. Trans, vol. 90, pp. 97-102, 1994.

[8] J. A. Freeman, and D. M. Skapura (Eds.), "Neural Networks Algorithms, Applications, and Programming Techniques," Addition Wesley Publishing Company, Reading, 1991.

[9] R. Garg, S. P. Gupta, H. Gao, M. S. Babu, and A. K. Debnath, "Comparative Quantitative Structure-Activity Relationship Studies on Anti-HIV Drugs," Chem. Rev, vol. 99, pp. 3525-3601, December 1999.

[10] W. J. Wang, Z. B. Xu, W. Z. Lu, and X.Y. Zhang, "Determination of the spread parameter in the Gaussian kernel for classification and regression," Neurocomputing, vol. 55, pp. 643–663, October 2003.

[11] L. Douali, D. Villemin, and D. Cherqaoui, "Exploring QSAR of Non-Nucleoside Reverse Transcriptase Inhibitors by Neural Networks: TIBO Derivatives," Int. J. Mol. Sci, vol. 5, pp. 48-55, January 2004.

[12] C.Y. Zhao, H.X. Zhang , X.Y. Zhang , M.C. Liu , Z.D. Hu, and B.T. Fan "Application of support vector machine (SVM) for prediction toxic activity of different data sets," Toxicology, vol. 217, pp. 105-119 August 2005