Research Article

Journal of CHEMOMETRICS

Received: 26 July 2010, Revised: 4 October 2010, Accepted: 9 October 2010, Published online in Wiley Online Library: 1 February 2011

(wileyonlinelibrary.com) DOI: 10.1002/cem.1364

Combination of kernel PCA and linear support vector machine for modeling a nonlinear relationship between bioactivity and molecular descriptors

Guang-Hui Fu[a], Dong-Sheng Cao[b], Qing-Song Xu[a]*, Hong-Dong Li[b] and Yi-Zeng Liang[b]

In this paper, a two-step nonlinear classification algorithm is proposed to model the structure–activity relationship (SAR) between bioactivities and molecular descriptors of compounds, which consists of kernel principal component analysis (KPCA) and linear support vector machines (KPCA + LSVM). KPCA is used to remove some uninformative gradients such as noises and then exactly capture the latent structure of the training dataset using some new variables called the principal components in the kernel-defined feature space. LSVM makes full use of the maximal margin hyperplane to give the best generalization performance in the KPCA-transformed space. The combination of KPCA and LSVM can effectively improve the prediction performance compared with the linear SVM as well as two nonlinear methods. Three datasets related to different categorical bioactivities of compounds are used to evaluate the performance of KPCA + LSVM. The results show that our algorithm is competitive. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: kernel methods; structure–activity relationship (SAR); kernel principal component analysis (KPCA); support vector machines (SVMs); classification; de-noising

1. INTRODUCTION

Structure–activity relationship (SAR), a very important area of chemometrics in the modern pharmaceutical industry, is urgently needed for predicting absorption, distribution, metabolism, excretion, toxicity (ADMET) properties to select lead compounds for optimization at the early stage of drug discovery and to screen drug candidates for clinical trials [1]. Much effort in recent SAR studies has been focused on predicting pharmacokinetic and toxicological properties that are collectively referred to as ADMET of compounds. The aim of SAR analysis is to investigate the relationship between chemical structure and biological activity. At present, many SAR modeling tools have been successfully employed to describe and build this relationship [2–7], for example, Artificial Neural Networks (ANN), Decision Tree (DT), Partial Least Squares (PLS), k-Nearest Neighbors (k-NN), Multiple Linear Regression (MLR), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM). Among all these modeling methods, SVM has been one of the most popular modeling tools in the SAR study due to its prediction performance in terms of accuracy [7–12]. However, many researchers have pointed out that SVM also suffered from the problem of feature subset selection [13–15]. Typically, redundant descriptors may destroy the pattern contained in the SAR and affect the prediction accuracy of the model. So how to avoid such a situation and improve the prediction ability for SVM is of practical importance in the SAR study.

In this paper, we proposed a two-step nonlinear classification algorithm to model the SAR between bioactivity and molecular descriptors, which consists of kernel principal component analysis (KPCA) and linear support vector machines (KPCA + LSVM). KPCA is used to remove some uninformative gradients such as noises and then capture the latent structure of the training dataset using some new variables called the principal components in the kernel-defined feature space. The use of LSVM is motivated by the construction of an optimal separating hyperplane in the sense of maximizing the distance to the closest point from either class. Three datasets related to different categorical bioactivities of compounds are employed to evaluate our method. These three datasets deal with human intestinal absorption (HIA), P-glycoprotein (P-gp) substrates and Torsade de Pointes (TdP), which are collectively related to the ADMET properties of compounds. The use of KPCA for dimensionality reduction or de-noising followed by LSVM computed on the reduced kernel feature space has shown good results in comparison with nonlinear SVM using the original data representation in three SAR datasets.

* *Correspondence to: Q.-S. Xu, School of Mathematical Science and Computing Technology, Central South University, Changsha 410083, P. R. China. E-mail: qsxu@mail.csu.edu.cn*

a *G.-H. Fu, Q.-S. Xu*
School of Mathematical Science and Computing Technology, Central South University, Changsha 410083, P. R. China

b *D.-S. Cao, H.-D. Li, Y.-Z. Liang*
Research Center of Modernization of Traditional Chinese Medicines, Central South University, Changsha 410083, P. R. China

The remainder of the paper is organized as follows: Sections 2 and 3 introduce the idea of KPCA method and linear support vector machines for classification, respectively. Section 4 describes our algorithm in detail. Section 5 introduces the dataset employed in this paper and the data pretreatment. Section 6 describes the experiments and the experimental results of the performance. Finally, in Section 7, we summarize and conclude the paper.

2. KERNEL PRINCIPAL COMPONENT ANALYSIS (KPCA)

Assuming that the dataset contains n observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i$ ($i = 1, 2, \ldots, n$) is a p-dimensional column vector. p is the number of predictors of the dataset. Let

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \tag{1}$$

be the predictor matrix and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ be response. For simplicity, we also assume that the data matrix has been centralized, namely $\sum_{i=1}^{n} \mathbf{x}_i = 0$.

Primal principal component analysis (PCA) is a simple method of extracting relevant information from complicated datasets. With minimal additional effort, PCA provides a roadmap for how to reduce a complex dataset to a lower dimension to reveal the hidden, simplified structure that often underlies it. PCA technique uses k ($k \leq p$) principal components to extract most information from the dataset. It is often the case that a small number of principal components is sufficient to account for most of the structure in the data.

Let the covariance matrix of the dataset be:

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \tag{2}$$

Principal component $\mathbf{v}_j (j = 1, 2, \cdots, p)$ can be computed by solving the following eigenvalue problem:

$$\lambda \mathbf{v} = \mathbf{C}\mathbf{v} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}\mathbf{v} \tag{3}$$

where $\lambda \geq 0$, $\mathbf{v} \neq 0$. We can employ singular value decomposition (SVD) technique to obtain p eigenvectors $\mathbf{v}_j$ ($j = 1, 2, \ldots, p$), as covariance matrix is positive semi-definite. Then one chooses first k eigenvectors corresponding first k largest eigenvalues as principal components, without loss of generality, denoted by $\mathbf{v}_d$ ($d = 1, 2, \ldots, k$). The projection of an observation $\mathbf{x} \in \mathbb{R}^p$ on them is

$$(\mathbf{v}_1^T\mathbf{x}, \mathbf{v}_2^T\mathbf{x}, \ldots, \mathbf{v}_d^T\mathbf{x})$$

which is the representation of $\mathbf{x}$ in the new coordinate system based on these k orthogonal principal components. So PCA actually is an orthogonal transformation of the coordinate system in which we describe our data.

Note that PCA is just applied to exploring linear pattern contained in the confusing dataset; KPCA is the natural generalization of PCA for finding nonlinear cases [16–19]. The basic idea of KPCA is to map the original dataset into some higher-dimensional feature space where we use the PCA method to establish linear model; however, this linear model established in the feature space is nonlinear in the original input space (see Figure 1).

KPCA method firstly maps original data points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ into a higher-dimensional feature space $\mathbf{F}$ by map ϕ:

$$\begin{aligned} \phi : \mathbf{R}^n &\rightarrow \mathbf{F} \\ \mathbf{x}_i &\rightarrow \phi(\mathbf{x}_i) \end{aligned} \tag{4}$$

Thus we get a new dataset $\{\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_n)\}$ in $\mathbf{F}$. The choice of the map ϕ aims to covert the nonlinear relations into linear ones. But it needs not to know what ϕ is.

We use the same denotations $\mathbf{X}$ and $\mathbf{C}$ to denote the predictor matrix and covariance matrix, respectively.

$$\mathbf{X} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_n)]^T \tag{5}$$

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \tag{6}$$

By the same argument as PCA, assume the dataset has centered.

Note that $\mathbf{C}\mathbf{v} = \frac{1}{n-1}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T v = \frac{1}{n-1}\sum_{i=1}^{n} (\phi(\mathbf{x}_i) \cdot \mathbf{v})\phi(\mathbf{x}_i)$, so all eigenvectors lie in the span of the data points. Thus they can be written as the linear combination of $\phi(\mathbf{x}_1)$, $\phi(\mathbf{x}_1)$,..., $\phi(\mathbf{x}_n)$. Namely,

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) \tag{7}$$

and by Equation (3):

$$\lambda(\phi(\mathbf{x}_i) \cdot \mathbf{v}) = (\phi(\mathbf{x}_i) \cdot \mathbf{C}\mathbf{v}) \quad \forall \quad i = 1, 2, \cdots, n \tag{8}$$

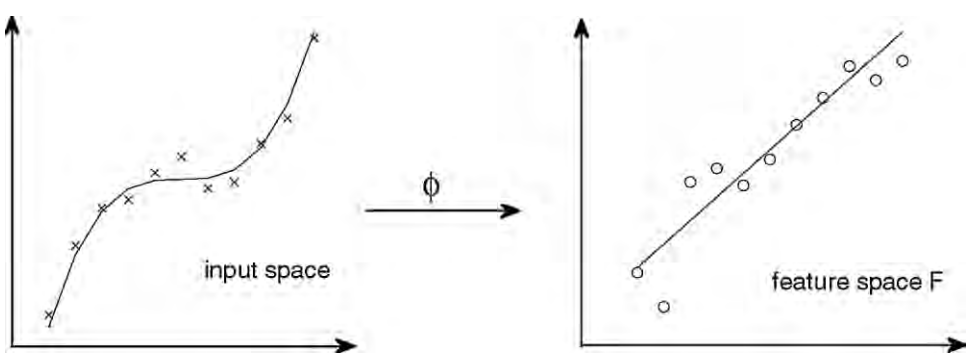

Figure 1. In the input space, the pattern (or relation) presented in the sample set is nonlinear. By mapping the sample set into feature space **F** later, the new dataset presents linear relation, and the inner products in the feature space can be calculated via some kernel function in the original input space.

Equation (7) tells that we should focus next on how to figure out the coefficient vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$. By combining Equations (7) and (8), the eigenvalue problem can be represented by the following simple form:

$$\lambda(n-1)\alpha = K\alpha \tag{9}$$

where

$$\mathbf{K} = \mathbf{X}\mathbf{X}^T \tag{10}$$

is kernel matrix or Gram matrix; its each entry can be represented as inner product form of two data points, namely,

$$\mathbf{K}_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \tag{11}$$

Kernel matrix plays a central role in the derivation of KPCA in that the inner product is equivalent to a so-called kernel function. Equation (3) also can be seen as an eigenvalue question of kernel matrix **K**, whose eigenvalue and corresponding eigenvector are $\lambda(n-1)$ and $\boldsymbol{\alpha}$, respectively. To normalize principal component **v**, we need to scale $\boldsymbol{\alpha}$ by factor $1/\sqrt{\lambda(n-1)}$. So Equation (7) reads:

$$\mathbf{v} = \sum_{i=1}^{n} \frac{1}{\sqrt{\lambda(n-1)}}\alpha_i\phi(\mathbf{x}_i) \tag{12}$$

We do not care what ϕ is, for the inner product can be computed by a so-called kernel function, such as Gaussian kernel function (radial basis function):

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left\{-\delta\|\mathbf{x}-\mathbf{y}\|^2\right\} \tag{13}$$

Let $\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$ be the matrix that consists of k principal components corresponding to first k largest eigenvalues. A new point **x** can be extracted by these k principal components, namely

$$\mathbf{v}_d^T\phi(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^d\phi(\mathbf{x}_i)^T\phi(\mathbf{x}) \quad (d = 1, 2, \cdots, k) \tag{14}$$

Remark: We cannot directly center the dataset in the feature space, but the kernel matrix $\hat{\mathbf{K}}$ of the centered dataset can be calculated by the kernel matrix **K** of noncentered case by the following formula:

$$\hat{\mathbf{K}} = \mathbf{K} - \mathbf{1}_n\mathbf{K} - \mathbf{K}\mathbf{1}_n + \mathbf{1}_n\mathbf{K}\mathbf{1}_n \tag{15}$$

where the matrix $(\mathbf{1}_n)_{ij} = 1/n$ for all $i, j = 1, 2, \ldots, n$

3. LINEAR SUPPORT VECTOR MACHINE (LSVM)

A detailed description of the theory of SVM can be easily found in several excellent books and literature [20–22]. SVM was originally developed by Vapnik *et al.* and has the capability to solve a number of biological classification problems. SVM is based on the structure risk minimization (SRM) principle from statistical learning theory. For linearly separable cases, SVM performs two classification tasks by constructing a hyperplane in the multidimensional space to differentiate two classes with a maximum margin (see Figure 2a). Given the supervised training dataset $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the decision function of SVM can be expressed in the following way:

$$f(\mathbf{x}_i) = \text{sign}\left(\mathbf{w}^T\mathbf{x}_i + b\right) \tag{16}$$

where **w** is a vector of weights and b is the constant coefficient. In the original feature space, the constraints for perfect classification can be described as:

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \quad (i = 1, 2, \ldots, n) \tag{17}$$

The vector **w** and parameter b can be estimated by solving the following quadratic optimization (QP) problem:

$$min \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{18}$$

subject to Equation (17).

In nonseparable case, slack variables, which are associated with the misclassified compounds (see Figure 2b), are added to the objective (18). Even though the erroneous classification cannot be avoided, the effect of the misclassified compounds can be reduced by means of these slack variables. Thus, Equation (18)

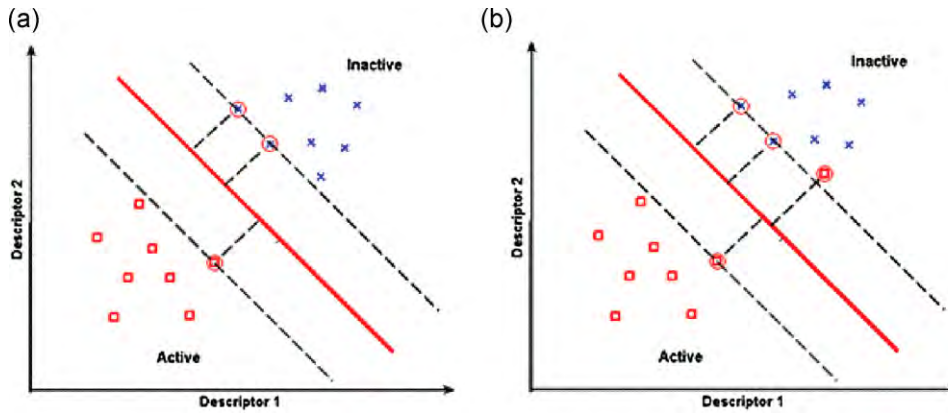

Figure 2. Support vector machines in linearly separable (a) and nonseparable (b) classification problems. The support vectors and margins are marked by red circles and dot lines, respectively. In nonseparable case, negative margins are associated with the misclassified compounds.

can further be re-expressed with a slack variable ξ:

$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{19}$$

$$s.t.\begin{cases} y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1-\xi_i \\ \xi_i \geq 0, \quad (i = 1, 2, \ldots, n) \end{cases} \tag{20}$$

By the Lagrange multiplier method, Equation (19) with constraints (20) has following the dual form

$$\min \quad \frac{1}{2}\alpha^T\mathbf{D}\mathbf{X}\mathbf{X}^T\mathbf{D}\alpha - e^T\alpha \tag{21}$$

$$s.t.\begin{cases} 0 \leq \alpha_i \leq C \quad (i = 1, 2, \ldots, n) \\ \sum_{i=1}^{n} y_i\alpha_i = 0 \end{cases} \tag{22}$$

where $\mathbf{D}$ is the $n \times n$ diagonal matrix with $\mathbf{D}_{ii} = y_i$ ($i = 1, 2, \ldots, n$) and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$ is the optimized Lagrange multiplier vector. $\mathbf{e}$ is the column vector of ones in n-dimensional real space. Equation (21) can be solved by means of QP methods. The above SVM algorithm is called linear SVM (LSVM) constructed in the original input space. A key property of LSVM is that it attempts to seek a 'safest' hyperplane maximizing the sum of squared distance between the hyperplane and all data points. The 'safest' hyperplane can give the correct prediction as far as possible when new samples arrive. That is, LSVM can define the 'safest' hyperplane to give the best generalization performance. Such obtained hyperplane is often referred to as the maximal margin hyperplane and is considered as the optimal hyperplane (see Figure 3). LSVM can use the optimal hyperplane for a better prediction performance compared with the other methods in most cases.

Lagrangian support vector machine (LagSVM) [23] is the generalization of the LSVM. It changes the 1-norm of slack variable ξ to a 2-norm squared, which makes the constraint $\xi \geq 0$ redundant. In addition, it appends the term b^2 to $||\mathbf{w}||^2$. Namely,

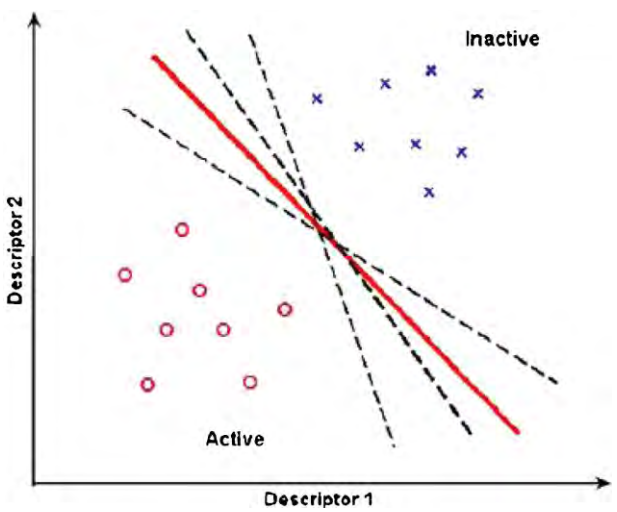

Figure 3. Even when the training set is linearly separable, there does not exist unique hyperplane to differentiate the two classes. However, the support vector machine can define the 'safest' hyperplane to give the best generalization performance (See the red line).

LagSVM is defined as the following objective:

$$\min \quad \frac{1}{2}\left(\|\mathbf{w}\|^2 + b^2\right) + C\sum_{i=1}^{n}\xi_i^2 \tag{23}$$

$$s.t. \quad y_i\left(\mathbf{w}^T\mathbf{x}_i - b\right) \geq 1-\xi_i \tag{24}$$

The dual of the above problem is

$$\min \quad \frac{1}{2}\alpha^T\left(\frac{\mathbf{I}}{C} + \mathbf{D}\left(\mathbf{X}\mathbf{X}^T + \mathbf{e}\mathbf{e}^T\right)\mathbf{D}\right)\alpha - \mathbf{e}^T\alpha \tag{25}$$

$$s.t. \quad \alpha_i \geq 0, \quad (i = 1, 2, \cdots, n) \tag{26}$$

where $\mathbf{I}$ is the identity matrix. The LagSVM is an algorithm similar to kernel ridge regression with constraints. The smaller the C is, the greater the amount of de-noising [24].

Note that LSVM and LagSVM are easy to generalize nonlinear form by replacing the term $\mathbf{X}\mathbf{X}^T$ in Equation (21) and $\mathbf{X}\mathbf{X}^T + \mathbf{e}\mathbf{e}^T$ in Equation (25) with a kernel matrix.

4. TWO-STEP NONLINEAR ALGORITHM

For nonlinear classification problems, the original SVMs firstly project the input feature vectors into a high-dimensional feature space using a kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ and then perform the LSVM algorithm in the kernel-defined feature space. However, as stated above, the variables in the kernel feature space may include some redundant information or noises, which may affect the prediction accuracy of the established model. It is necessary to remove such useless information before performing the LSVM algorithm. To deal with such situation, a two-step nonlinear algorithm based on the combination of KPCA and LSVM (KPCA + LSVM) is proposed. The two-step KPCA + LSVM algorithm is given below (see Figure 4).

4.1. Step 1: perform KPCA in input space

KPCA is carried out in the input space to extract the compact underlying structure of the dataset. We can calculate its orthonormal eigenvectors corresponding to first k largest nonzero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. So, the corresponding scores can be computed as $\mathbf{T}_k = \phi(\mathbf{X}) \cdot \mathbf{V}_k$ and the original kernel matrix can be re-constructed as $\mathbf{K} = \phi(\mathbf{X})\phi(\mathbf{X})^T \approx \mathbf{T}_k\mathbf{T}_k^T$. Here, k should be further optimized by means of model selection techniques such as cross-validation (CV).

4.2. Step 2: perform LSVM in KPCA-transformed space

The LSVM can be directly carried out by means of the reconstructed kernel matrix in the KPCA-transformed space.

Here are two remarks.

(1) The KPCA + LSVM method has a consistent framework with the existing nonlinear SVMs. KPCA + LSVM will be changed into the original nonlinear SVMs when all the scores in KPCA are used to perform the LSVM algorithm. However, the KPCA + LSVM algorithm is more flexible compared with the existing nonlinear SVMs, especially when the variables in the kernel-defined feature space include some redundant information or noises.

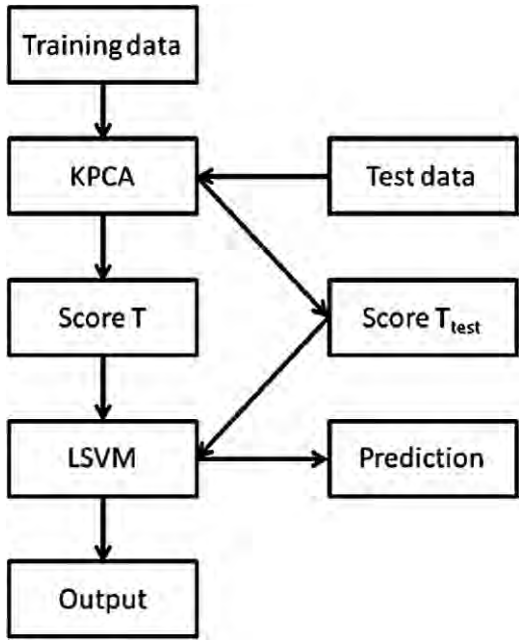

Figure 4. The flow chart of the KPCA + LSVM algorithm. T indicates the scores matrix of KPCA. In the KPCA + LSVM algorithm, there are some important parameters that need to be further optimized.

(2) Primal PCA algorithm may destroy the underlying nonlinear structure possessed by the training dataset. However, KPCA is performed in the kernel-defined feature space and is more likely to capture the underlying nonlinear structure of the training dataset. So, KPCA + LSVM is theoretically more reasonable compared with commonly used SVM coupled with PCA.

5. EXPERIMENTAL

5.1. Three datasets

As good pharmacokinetic properties are very important for the drug candidates, there have been increasing efforts in SAR research to address the prediction accuracy of the pharmacokinetic properties of compounds, including ADMET. For example, the studies of HIA [25,26], P-gp [27–29] and TdP [15] focus on prediction of the ADMET and adverse drug effects. We selected three datasets related to these pharmacokinetic and pharmacodynamic properties for evaluating the performance of our proposed method in the prediction of binary classes of SAR. A brief description of the three datasets including the number of compounds and their distribution into the active and inactive classes as well as the molecule descriptors used for each dataset is given in Table I. Here are more details for each dataset.

5.1.1. *Dataset 1: HIA*

The absorption of a drug compound through the human intestinal cell lining is an important property for potential drug candidates. There are 131 absorbable (HIA+) and 65 nonabsorbable (HIA−) compounds that are classified by the 'measured absorption rate' of 70% criterion. HIA comes from Xue *et al*. So do the datasets P-gp and TDP described below. We employ the original set of 159 descriptors provided by Xue *et al*. [15] for HIA, P-gp and TdP sets to facilitate the comparison among different studies. It includes 18 simple molecular properties, 28 molecular connectivity and shape descriptors, 84 electrotopological state descriptors, 13 quantum chemical properties as well as 16 geometrical properties.

5.1.2. *Dataset 2: P-gp*

P-gp is a transmembrane protein capable of transporting a wide variety of anticancer drugs out of the cell, hence hampering in chemotherapeutic treatment. An increased expression of P-gp is associated with multidrug resistance (MDR). Many studies have been undertaken to develop MDR-reversing compounds with potential clinical significance. P-gp substrates (P-gp+) are reported as being transported by P-gp or P-gp MDR reversals and nonsubstrates of P-gp (P-gp−) are those described as not transportable by P-gp. A total of 116 substrates and 85 nonsubstrates of P-gp were collected in the dataset of P-gp.

5.1.3. *Dataset 3: TdP*

TdP is a potentially fatal polymorphic ventricular tachycardia. It may also be induced by drugs that cause QT (Q wave and T wave) prolongation. This effect is present in different categories of therapeutic agents, for example, antihistamines, antidepressants or macrolide antibiotics. The TdP dataset included 85 TdP-inducing agents (TdP+) and 276 noninducing compounds (TdP−).

5.2. Data pretreatment and performance evaluation

Among 159 predictors of three datasets, constant variables exist and they are eliminated beforehand. Each predictor is also scaled to have zero mean and unit variance. One of the advantages of doing this is to bound the parameter δ of the Gaussian kernel function (see Equation (13)).

To evaluate the performance of a new algorithm, rigorous validation is necessary for the SAR model development. Evidence is presented that only models that have been validated by both external and internal validation can be considered reliable and

Table I. The three datasets used in the work

Dataset	Compound	Class+	Class−	Predictor set
HIA	196	131	65	159 descriptors include molecular properties, molecular connectivity, shape descriptors, etc.
P-gp	201	116	85	
TdP	361	85	276	

applicable for both external prediction and regulatory purpose [30,31]. So we use external and internal validation to evaluate the performance of the KPCA + LSVM algorithm in this paper. For external validation, the dataset is randomly split into training set used for establishing the model and test set for external validation. The training and test sets contain 80 and 20% observations of the dataset, respectively. For internal validation, 5-fold CV is employed to estimate the accuracy of our model. For 5-fold CV, the training set is randomly split into five roughly equal-sized parts firstly, and then we fit the model to four parts and calculate the prediction error of the fitted model with the remainder part. The process is repeated five times so that every part can be predicted as a validation set.

The parameters employed to evaluate the behavior in this investigation are some commonly used ones in classification problems: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). There are several criteria for assessing the prediction performance including sensitivity (SE) (the prediction accuracy of active compounds), specificity (SP) (the prediction accuracy of inactive compounds), the overprediction accuracy (R) and Matthews correlation coefficient (MCC), which are given by the following equations, respectively:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{27}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{28}$$

$$\mathrm{R} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{29}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{30}$$

5.3. Effects of model parameters

Gaussian kernel function (see Equation (13)), widely used in many works due to its good performance, is employed to construct the nonlinear mapping in our study. Thus, three important parameters, k (the number of principal components), C (the regularization parameter) as well as δ (the width of Gaussian kernel function), need to be further optimized in KPCA + LSVM algorithm. The parameter k, on the one hand, controls the ability to reconstruct the dataset in the kernel-defined feature space and hence measures the model complexity. On the other hand, k restricts the de-noising ability; the smaller the k, the greater the amount of de-noising. The choice of k depends on the contribution to the response values. The parameter C is the tradeoff between maximizing the margin and minimizing the training error, so it affects both trained and predicted results. Usually, k is an unknown parameter before modeling. If C is too small, an insufficient stress will be placed on fitting the training data. If C is too large, the algorithm will overfit the training data. The width δ of Gaussian kernel function is also crucial and should be tuned carefully. A very small δ can excessively model the local structure of the training data and so may overfit the training data, whereas a high δ does not capture the underlying structure of the training data and so may underfit the training data. Generally speaking, these three parameters of KPCA + LSVM are mutually interrelated and should be optimized jointly by means of model selection techniques such as CV, etc. In this paper, CV method is employed and a multiparameter grid search strategy is established to seek the optimal combination of model parameters simultaneously. It should be pointed out that the overprediction accuracy (R) (see Equation (29)) is employed to act as the optimal criterion in searching the parameter grids.

6. RESULTS AND DISCUSSION

The prediction accuracies of KPCA + LSVM for three SAR datasets were primarily evaluated by means of 5-fold CV.

On the one hand, all model parameters are further optimized by means of a grid search strategy. For the regularization parameter C, we set eight values ($C = 0.001$, 0.01, 0.1, 0.5, 1, 10, 100, 200). The number of principal component k is set to range from 1 to 50. For the width δ of Gaussian kernel function, we firstly estimate a suitable range and then set 10 values ($\delta = 0.0001$, 0.0002, 0.0003, 0.0005, 0.001, 0.005, 0.006, 0.01, 0.05, 0.1). Thus, we can make use of $8 \times 50 \times 10 = 4000$ grid points to search for the optimal combination of model parameters.

On the other hand, both internal and external validation are used to evaluate the performance of the KPCA + LSVM algorithm. The linear relationship is built by LSVM algorithm for the three SAR datasets. The prediction accuracy shows that it is wise to find a nonlinear model for these SAR data. LagSVM [23] and SVM [20,21] are also quoted for the purpose of comparisons. The results of internal and external validation for three SAR datasets are shown in Tables II and III, respectively.

6.1. Internal validation results for three SAR datasets

In internal validation, 5-fold CV method is used. The overprediction accuracy (R) acts as the optimal criterion in searching the parameter grids. As shown in Table II, the prediction accuracy of LSVM is low. The three SAR datasets given in section 5 do not exist obvious linear structure with their responses. That the results of three nonlinear methods (SVM, LagSVM and KPCA + LSVM) are better than that of LSVM further indicates that it is more suitable to establish nonlinear model for them. Among three nonlinear methods, KPCA + LSVM achieves the best prediction accuracy compared with SVM and LagSVM for each of the three SAR datasets. In view of the average prediction ability on four guidelines, KPCA + LSVM wins 81.74 and 55.71% accuracy on over prediction accuracy (R) and Matthews correlation coefficient (MCC), respectively. These results are the best among all the considered methods. For the specificity (SP), KPCA + LSVM obtains 73.00% accuracy, which is a bit low compared with the highest point 73.27% which is obtained by LagSVM. SVM achieves the best prediction accuracy of 78.76% on sensitivity (SE) and 77.94% on KPCA + LSVM.

The prediction ability of KPCA + LSVM is superior to that of SVM and LagSVM for all three SAR datasets. LagSVM is also better than SVM. SVM does not consider the noises of the data in training the classifier. However, both LagSVM and KPCA + LSVM have de-noising functions by controlling the regularization parameter C and the number of principal components k, respectively. The difference between them is as follows: The parameter C of LagSVM is not only to control the de-noising performance, but also to balance maximizing the margin of the

Table II. Internal validation results on three SAR datasets

	Datasets	SP (%)	SE (%)	R (%)	MCC (%)	Parameters
a	HIA	61.54	86.26	78.06	49.30	
	P-gp	61.18	72.41	67.66	33.64	
	TdP	85.87	62.35	80.33	46.95	
	Average	69.53	73.68	75.35	43.30	
b	HIA	61.54	87.79	79.08	51.40	$C=100$, $\delta=0.002$
	P-gp	57.65	89.66	76.12	50.83	$C=1$, $\delta=0.005$
	TdP	89.49	58.82	82.27	49.58	$C=100$, $\delta=0.0005$
	Average	69.56	78.76	79.16	50.60	
c	HIA	58.46	91.60	80.61	54.43	$C=1$, $\delta=0.01$
	P-gp	68.24	86.21	78.61	55.76	$C=1$, $\delta=0.01$
	TdP	93.12	54.12	83.93	52.16	$C=0.5$, $\delta=0.005$
	Average	73.27	77.31	81.05	54.12	
d	HIA	55.38	94.66	81.63	56.93	$C=10$, $\delta=0.01$, $k=27$
	P-gp	69.41	86.21	79.10	56.81	$C=1$, $\delta=0.006$, $k=38$
	TdP	94.20	52.94	84.49	53.38	$C=1$, $\delta=0.01$, $k=29$
	Average	73.00	77.94	81.74	55.71	

a: LSVM; b: SVM; c: LagSVM; d: KPCA + LSVM.

classifier and minimizing the training error. So it is possible that the de-noising function has to be sacrificed in order to achieve a better tradeoff between the maximal margin and minimal training error. It is more flexible for KPCA + LSVM to control the de-noising ability by using the number of principal components k. KPCA + LSVM can make full use of KPCA to carry out the dimensionality reduction or de-noising in the kernel-defined feature space and thus remarkably improve the prediction performance.

6.2. External validation results for three SAR datasets

In external validation, the dataset is randomly split into training set and test set which contain 80 and 20% observations of the dataset, respectively. Methods similar to those used in internal validation are employed to evaluate our algorithm. Table III shows the comparison results of the prediction accuracies on three SAR datasets from linear SVM and three nonlinear classification methods. The external validation results on LSVM again exhibit the worst prediction performance.

Among three nonlinear methods of SVM, LagSVM and KPCA + LSVM, it seems that LagSVM outperforms. Particularly, LagSVM wins the best prediction accuracy on the datasets HIA and P-gp on the overprediction accuracy (R). However, on the dataset TdP, KPCA + LSVM achieves the over prediction accuracy of 87.67%, which is the highest in comparison with SVM and LagSVM. Table I shows that datasets HIA, P-gp and TdP contain 196, 201 and 361 observations, respectively. It seems that LagSVM is more suitable for dealing with small sample problems.

Table III. External validation results on three SAR datasets

	Datasets	SP (%)	SE (%)	R (%)	MCC (%)	Parameters
a	HIA	84.62	85.19	85.00	67.53	
	P-gp	70.59	70.83	70.73	40.92	
	TdP	82.14	58.82	76.71	38.82	
	Average	79.12	71.61	77.48	49.09	
b	HIA	69.23	96.30	87.50	70.88	$C=100$, $\delta=0.0003$
	P-gp	58.82	91.67	78.05	54.67	$C=0.5$, $\delta=0.002$
	TdP	92.86	58.82	84.93	55.48	$C=10$, $\delta=0.002$
	Average	73.64	82.26	83.49	60.34	
c	HIA	84.62	100.00	94.87	88.64	$C=10$, $\delta=0.1$
	P-gp	88.24	95.65	92.50	84.65	$C=100$, $\delta=0.0005$
	TdP	98.18	41.18	84.72	53.19	$C=0.01$, $\delta=0.05$
	Average	90.35	78.94	90.70	75.49	
d	HIA	76.92	96.30	90.00	76.80	$C=1$, $\delta=0.01$, $k=25$
	P-gp	82.35	79.17	80.49	60.78	$C=0.5$, $\delta=0.005$, $k=33$
	TdP	98.21	52.94	87.67	62.88	$C=10$, $\delta=0.01$, $k=27$
	Average	85.83	76.14	86.05	66.82	

a: LSVM; b: SVM; c: LagSVM; d: KPCA + LSVM.

In fact, LagSVM for nonlinear kernel, which does not make use of the Sherman–Morrison–Woodbury identity, does not scale up to very large problems [23]. KPCA + LSVM is more competitive while investigating data with many observations. That the results of KPCA + LSVM on internal validation (Table II) are the best compared with SVM and LagSVM can indirectly support this point. , as the samples are used repeatedly in 5-fold CV.

7. CONCLUSIONS

In this paper, we proposed a new strategy, KPCA + LSVM, to carry out a nonlinear classification problem for the SAR datasets. This strategy is exactly consistent with the existing nonlinear SVM algorithms when all the scores are used in the LSVM algorithm. However, KPCA + LSVM is more flexible and can obtain a better prediction performance compared with the original SVMs, especially when the variables in the kernel feature space include some redundant information or noises. The results from internal and external validations on three SAR datasets have demonstrated that the use of KPCA for dimensionality reduction or de-noising followed by LSVM computed on the reduced kernel feature space can obtain a better prediction performance in comparison with nonlinear SVMs using the original data representation. LagSVM, another similar kernel algorithm with de-noising function, is also quoted to further investigate the de-noising performance. In view of both internal and external validation, KPCA + LSVM is better as least competitive compared with LagSVM, especially when it comes to the large sample problem. However, we also notice that the parameter of principal components k is large to some extent at the optimal point. Reducing the value of k requires further effort.

Acknowledgements

This work is financially supported by the National Nature Foundation Committee of P.R. China (grants no. 10771217 and 20875104), the International Cooperation Project on Traditional Chinese Medicines of Ministry of Science and Technology of China (grant no. 2007DFA40680). The studies are approved by the university's review board.

REFERENCES

1. Li H, Sun J, Fan X, Sui X, Zhang L, Wang Y, He Z. Considerations and recent advances in qsar models for cytochrome p450-mediated drug metabolism prediction. *J. Comput. Aided Mol. Des.* 2008; **22**(11): 843–855. DOI: 10.1007/s10822-008-9225-4
2. Agatonovic-Kustrin S, Davies P, Turner JV. Structure-activity relationships for serotonin transporter and dopamine receptor selectivity. *Med. Chem.* 2009; **5**(3): 271–278.
3. Asikainen A, Kolehmainen M, Ruuskanen J, Tuppurainen K. Structure-based classification of active and inactive estrogenic compounds by decision tree, lvq and knn methods. *Chemosphere* 2006; **62**(4): 658–673.
4. Askjaer S, Langgard M. Combining pharmacophore fingerprints and pls-discriminant analysis for virtual screening and sar elucidation. *J. Chem. Inf. Model.* 2008; **48**(3): 476–488.
5. Li JZ, Liu HX, Yao XJ, Liu MC, Hu ZD, Fan BT. Structure-activity relationship study of oxindole-based inhibitors of cyclin-dependent kinases based on least-squares support vector machines. *Anal. Chim. Acta* 2007; **581**(2): 333–342.
6. Waske B, Schiefer S, Braun M. and Ieee. Random feature selection for decision tree classification of multi-temporal sar data. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Denver, CO, 31 July–04 August 2006; 168–171.
7. Yuan YN, Zhang RS, Hu RJ, Ruan XF. Prediction of ccr5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression. *Eur. J. Med. Chem.* 2009; **44**(1): 25–34.
8. Darnag R, Schmitzer A, Belmiloud Y, Villemin D, Jarid A, Chait A, Seyagh M, Cherqaoui D. Qsar studies of hept derivatives using support vector machines. *QSAR Comb. Sci.* 2009; **28**(6–7): 709–718.
9. Dong XW, Jiang CY, Hu HY, Yan JY, Chen J, Hu YZ. Qsar study of akt/protein kinase b (pkb) inhibitors using support vector machine. *Eur. J. Med. Chem.* 2009; **44**(10): 4090–4097.
10. Leong MK, Chen YM, Chen TH. Prediction of human cytochrome p450 2b6-substrate interactions using hierarchical support vector regression approach. *J. Comput. Chem.* 2009; **30**(12): 1899–1909.
11. Vasanthanathan P, Taboureau O, Oostenbrink C, Vermeulen NPE, Olsen L, Jorgensen FS. Classification of cytochrome p450 1a2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* 2009; **37**(3): 658–664.
12. Yuan H, Huang JP, Cao CZ. Prediction of skin sensitization with a particle swarm optimized support vector machine. *Int. J. Mol. Sci.* 2009; **10**(7): 3237–3254.
13. Bierman S, Steel S. Variable selection for support vector machines. *Commun. Stat. Simul. Comput.* 2009; **38**(8): 1640–1658.
14. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* 2002; **46**(1–3): 389–422.
15. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* 2004; **44**(5): 1630–1638. DOI: 10.1021/ci049869h
16. Liu XQ, Kruger U, Littler T, Xie L, Wang SQ. Moving window kernel pca for adaptive monitoring of nonlinear processes. *Chemom. Intell. Lab. Syst.* 2009; **96**(2): 132–143.
17. Rosipal R, Girolami M, Trejo LJ, Cichocki A. Kernel pca for feature extraction and de-noising in nonlinear regression. *Neural Comput. Appl.* 2001; **10**(3): 231–243.
18. Scholkopf B, Mika S, Burges CJC, Knirsch P, Muller KR, Ratsch G, Smola AJ. Input space versus feature space in kernel-based methods. *IEEE Trans. Neural Netw.* 1999; **10**(5): 1000–1017.
19. Wu W, Massart DL, de Jong S. The kernel pca algorithms for wide data. part i: theory and algorithms. *Chemom. Intell. Lab. Syst.* 1997; **36**(2): 165–172. DOI: 10.1016/S0169-7439(97)00010-5
20. Cristianini N, Shawe-Taylor J (eds). *An Introduction to Support Vector Machines*. Cambridge University Press: Cambridge, 2000.
21. Schokopf B, Smola A (eds). *Learning with Kernels*. MIT Press: Cambridge, 2002.
22. Vapnik V. *The Nature of Statistical Learning Theory*. Springer: New York, USA, 1995.
23. Mangasarian O, Musicant D. Lagrangian support vector machines. *J. Mach. Learn. Res.* 2001; **1**: 161–177.
24. Hansen P. *Rank-deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion*. Society for Industrial Mathematics: Philadelphia, USA, 1998.
25. Hou T, Wang J, Zhang W, Xu X. Adme evaluation in drug discovery. 7. prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* 2006; **47**(1): 208–218. DOI: 10.1021/ci600343x
26. Willmann S, Schmitt W, Keldenich J, Lippert J, Dressman JB. A physiological model for the estimation of the fraction dose absorbed in humans. *J. Med. Chem.* 2004; **47**(16): 4022–4031. DOI: 10.1021/jm030999b
27. Kaiser D, Terfloth L, Kopp S, Schulz J, de Laet R, Chiba P, Ecker GF, Gasteiger J. Self-organizing maps for identification of new inhibitors of p-glycoprotein. *J. Med. Chem.* 2007; **50**(7): 1698–1702. DOI: 10.1021/jm060604z
28. Wang Y-H, Li Y, Yang S-L, Yang L. Classification of substrates and inhibitors of p-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* 2005; **45**(3): 750–757. DOI: 10.1021/ci050041k
29. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ. Prediction of p-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* 2004; **44**(4): 1497–1505.
30. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* 2007; **26**(5): 694–701.
31. Tropsha A, Gramatica P, Gombar V. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 2003; **22**(1): 69–77.