Available online at www.sciencedirect.com
SCIENCE DIRECT
Solid State Sciences 8 (2006) 129–136
Solid State Sciences
www.elsevier.com/locate/ssscie

Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors

Tianhong Gu, Wencong Lu *, Xinhua Bao, Nianyi Chen

Department of Chemistry, College of Sciences, Shanghai University, Shanghai, 200444, China

Received 23 May 2005; received in revised form 13 September 2005; accepted 12 October 2005

Available online 28 December 2005

Abstract

In this work, atomic parameters support vector regression (APSVR) was proposed to predict the band gap and melting point of III–V, II–VI binary and I–III–VI_2, II–IV–V_2 ternary compound semiconductors. The predicted results of APSVR were in good agreement with the experimental ones. The prediction accuracies of different models were discussed on the basis of their mean error functions (*MEF*) in the leave-one-out cross-validation. It was found that the performance of APSVR model outperformed those of back propagation-artificial neural network (BP-ANN), multiple linear regression (MLR) and partial least squares regression (PLSR) methods.

Keywords: Semiconductor; Band gap; Melting point; Support vector regression; Atomic parameters

1. Introduction

III–V and II–VI binary compounds are important semiconductors for microwave, optoelectron and infrared devices, while I–III–VI_2 and II–IV–V_2 ternary compounds are largely developed as nonlinear optical devices and solar cell materials. The band gaps (E_g) and melting points (T_m) are essential properties of these compounds. It would be helpful for materials scientists to estimate the band gap (E_g) and melting point (T_m) of a compound before synthesizing it. Up to now, no satisfactory results are reported about the computation of these properties only using the principle of theoretical physics and quantum chemistry. However, on the basis of known data set available, it is reasonable to predict the properties of unseen samples by using chemometric methods. For example, the band gaps of these compounds were predicted by using back propagation-artificial neural network (BP-ANN) [1]. Since there are a lot of chemometric methods, one has to deal with the troublesome problem about model selection for a particular data set with finite number of samples and multiple features. It is very important to select a proper model with good generalization ability, i.e., low mean relative error for the properties of new compounds (unseen samples).

Generally speaking, the modeling problem is actually ill-posed in the sense of Hadamard [2]. So, how to choose the right balance between model flexibility and overfitting to a limited training set is one of the most difficult obstacles for obtaining a model with good generalization ability. As an effective way to overcome the problem of overfitting, support vector machine (SVM) based on statistical learning theory (SLT) has been proposed by V.N. Vapnik [3]. Now SVM has gained successful application in such research fields as drug design [4], combinatorial chemistry [5], proteins [6,7], etc. We also reported the applications of SVM in chemistry [8]. The goal of this work is to build a quantitative Support Vector Regression (SVR) model for predicting the band gaps (E_g) and melting points (T_m) of some semiconductors, with special consideration of its generalization abilities in the leave-one-out cross-validation (LOOCV) test, which was widely used as a reliable way for assessing model validity.

2. Method

2.1. Support vector regression [3]

SVM can be applied to regression by the introduction of an alternative loss function and the results appear to be very en-

* Corresponding author. Tel.: +86-21-6613-3513; fax: +86-21-6613-4080.
E-mail address: wclu@staff.shu.edu.cn (W. Lu).

1293-2558/$ – see front matter
doi:10.1016/j.solidstatesciences.2005.10.011

couraging. In SVR, the basic idea is to map the data X into a higher-dimensional feature space F via a nonlinear mapping Φ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(\mathbf{x}_i; d_i)\}_{i=1}^{l}$ ($\mathbf{x}_i$ is input vector, d_i is the desired value). SVM approximates the function in the following form:

$$y = \sum_{i=1}^{l} w_i \Phi_i(\mathbf{x}) + b. \tag{1}$$

Where $\{\Phi_i(\mathbf{x})\}_{i=1}^{l}$ is the set of mappings of input features, and $\{w_i\}_{i=1}^{l}$ is a vector of weights in the features space, and b are coefficients. They are estimated by minimizing the regularized risk function $R(C)$:

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|\mathbf{w}\|^2, \tag{2}$$

where

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon & \text{for } |d - y| \geqslant \varepsilon, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and ε is a prescribed parameter in the insensitive loss function.

In Eq. (2), $C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i)$ is the so-called empirical error (risk), which is measured by ε-insensitive loss function $L_\varepsilon(d, y)$, which indicates that it does not penalize errors below ε. The second term, $(1/2)\|\mathbf{w}\|^2$, is used as a measurement of function flatness. C is a regularized constant determining the tradeoff between the training error and the model flatness. Introduction of slack variables "ξ" leads Eq. (2) to the following constrained function:

$$\text{Max } R(\mathbf{w}, \xi^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C^*\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{4}$$

$$\text{s.t. } w\Phi(\mathbf{x}_i) + b - d_i \leqslant \varepsilon + \xi_i,$$
$$d_i - w\Phi(\mathbf{x}_i) - b \leqslant \varepsilon + \xi_i^*,$$
$$\xi_i, \xi_i^* \geqslant 0. \tag{5}$$

Thus, decision function Eq. (1) becomes the following form:

$$f(\mathbf{x}, \alpha_i, \alpha_i^*) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}_j) + b. \tag{6}$$

In Eq. (6), α_i, α_i^* are the introduced Lagrange multipliers. They satisfy the equality $\alpha_i \cdot \alpha_i^* = 0$, $\alpha_i \geqslant 0$, $\alpha_i^* \geqslant 0$; $i = 1, \ldots, l$, and are obtained by maximizing the dual form of Eq. (4), which has the following form:

$$\Phi(\alpha_i, \alpha_i^*) = \sum_{i=1}^{l} d_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

with the following constrains:

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, l,$$
$$0 \leqslant \alpha_i^* \leqslant C, \quad i = 1, \ldots, l,$$
$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0. \tag{8}$$

Solving Eq. (7) with constraints Eq. (8) determines the Lagrange multipliers, α_i, α_i^*. Based on the Karush–Kuhn–Tucker (KKT) conditions of quadratic programming, only a number of coefficients $(\alpha_i - \alpha_i^*)$ will assume nonzero values, and the data points associated with them could be referred to as support vectors.

In Eq. (6), $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function. The value is equal to the inner product of two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in the feature space $\Phi(\mathbf{x})$, i.e. $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$. The elegance of using kernel function lied in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(\mathbf{x})$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function.

The SVR algorithm is implemented as follow:

Step 1: Normalize all the data.

Step 2: Set variables C and ε.

Step 3: Structure a quadratic programming (QP) problem Eq. (4).

Step 4: Transfer QP problem into the formula of Lagrange function.

Step 5: Solve QP problem by SMO (Sequential Minimal Optimization) [9].

Step 6: Obtain the parameter w and b.

2.2. Support vector regression using kernels

Kernel function is originally a kind of functions used in integral operator research. But Vapnik implemented this function into his newly invented SVM method. The use of kernel function makes SVM able to treat nonlinear data processing problems by using linear algorithms.

There are four commonly used kernel functions:

- Linear kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \big(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + \theta\big), \tag{9}$$

- Gaussian (RBF) kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right), \tag{10}$$

- Polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \big(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + \theta\big)^d. \tag{11}$$

2.3. Implementation of SVM

According to the literature [3] the SVM software package including SVR was programmed in our lab. The validation of the software has been tested in some applications in chemistry and chemical engineering [8]. All the computations were carried out on a Pentium IV computer with a 2.0G Hz processor.

3. Results and discussion

3.1. $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds

3.1.1. Data set

The data set consists of 25 compounds, including AlP, AlAs, AlSb, GaP, GaAs, GaSb, InP, InAs, InSb, ZnS, ZnSe, ZnTe, CdS, CdSe, CdTe, HgS, HgSe, HgTe, AlN, GaN, InN, PbO, PbS, PbSe, and PbTe [10–13].

3.1.2. Atomic parameters

Atomic parameters are successful used in the prediction of some property of some compounds [14]. The physic-chemical behavior of an alloy system is mainly related to geometrical, charge transfer and energy band factors. The effective atomic radius or its function can roughly describe the geometrical factor. The charge transfer factor should be explained by the difference of electronegativities of constituent elements. The energy band factor is usually considered as the valence or its function [15]. The crystal structure of the compound semiconductors is based on the principle of closest packing of anions. The radius affects the physic-chemical property of semiconductor compounds. According to Mooser and Pearson's formula [16], electrovalent is an important factor in the criterion of the formation of semiconductor. In this work, the quantitative model was built on the basis of atomic parameters including electronegativity, valence, radius and atomic mass. However, some functions of atomic parameters can also be chosen as features in modeling. So, SVM-RFE (Recursive feature elimination) [8,17] method is used to select features. SVM-RFE is good for selecting relevant features from data set.

3.1.3. Model building for E_g of $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds

The experimental E_g values of 25 binary compounds were used to construct the regression model. In this work, the sum of proportion of atomic electrovalent and covalent radius $\sum(z/r_{\mathrm{cov}})$ [18], mean atomic number $\overline{N}$, atomic electrovalent Z_A and Z_B were selected as parameters in the model of band gap [10,11], where

$$\overline{N} = \frac{N_A + N_B}{2}, \tag{12}$$

$$\sum(z/r_{\mathrm{cov}}) = (z/r_{\mathrm{cov}})_A + (z/r_{\mathrm{cov}})_B. \tag{13}$$

All data were scaled in the interval between 0 and 1. In the present work, the leaving-one-out cross-validation (LOOCV) test was undertaken to find the suitable capacity parameter C, ε-insensitive loss function and kernel function for SVR model. In order to measure the generalization ability of SVR model, we defined the mean error function (*MEF*) as Eq. (14)

$$MEF = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - e_i|}{e_{\max} - e_{\min}} \times 100\%, \tag{14}$$

where e_i is the experimental value of sample i, p_i is the predicted value of sample i, n is the number of the whole samples, $e_{\max}, e_{\min}$ is the maximum and minimum experimental value of whole samples, respectively. In general, the smaller the value of *MEF* obtained, the better generalization ability expected. Fig. 1 (a, b, c) illustrated *MEF* versus ε and C under different kernel functions (linear, polynomial, radial basis ($\sigma = 1.00$)), respectively.

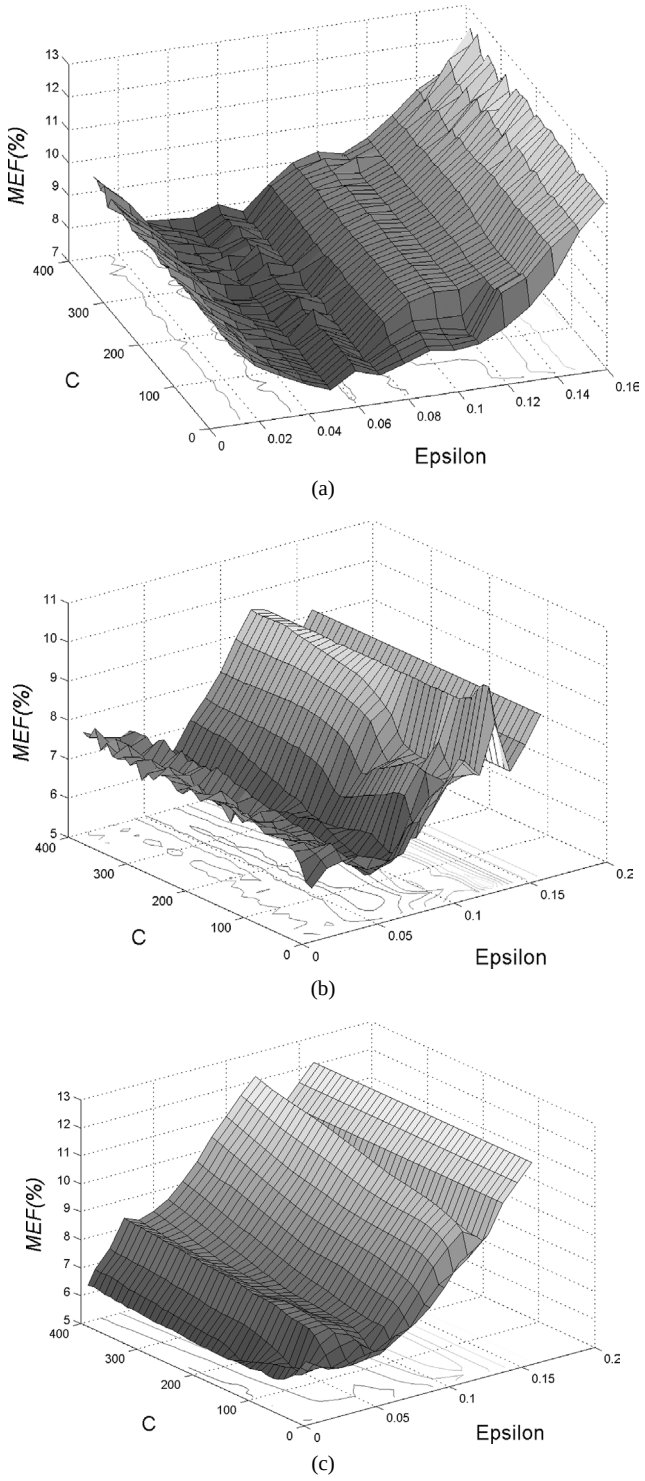

Fig. 1. (a) *MEF* versus C and ε using LOOCV with linear kernel function. (b) *MEF* versus C and ε using LOOCV with polynomial kernel function. (c) *MEF* versus C and ε using LOOCV with RBF kernel function ($\sigma = 1.00$).

It is found that the optimal SVR model with the least *MEF* is available when the kernel function is polynomial with the form:

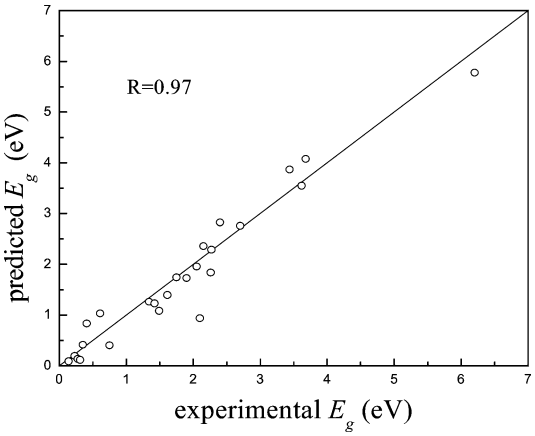

Fig. 2. Experimental E_g vs predicted E_g of binary compound semiconductors with trained SVR model.

$K(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1)^2$ with $\varepsilon = 0.07$ and the regularized constant $C = 70$. By using above kernel function and parameters optimized, the trained SVR model for E_g of $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds with original data is available as follows:

$$E_g = 3.479 \times \left[\sum_{i,j=1}^{25} (\alpha_i - \alpha_i^*) \big(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1 \big)^2 + 0.7473 \right] + 0.1410, \tag{15}$$

where $(\alpha_i - \alpha_i^*)$ is Lagrange coefficient corresponding to support vector. Fig. 2 illustrates the relationship of predicted E_g and experimental E_g of $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds, with related coefficient (R) 0.97.

3.1.4. Model building for T_m of $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds

In the present work, the regression model for T_m was built by using experimental data from a set of 25 binary compounds. Electronegativity in the Pauling scale of values $\Delta\chi$, mean Born exponent $\bar{n}$ (Born exponent is an important parameter determining physical properties of ionic crystals [19]), the sum of proportion of atomic electrovalent and covalent radius $\sum(z/r_{\mathrm{cov}})$, mean atomic mass $\bar{m}$ were selected to model melting point [10–13], where

$$\Delta\chi = |\chi_A - \chi_B|, \tag{16}$$

$$\bar{n} = \frac{n_A + n_B}{2}, \tag{17}$$

$$\bar{m} = \frac{m_A + m_B}{2}. \tag{18}$$

Fig. 3 (a, b, c) illustrates *MEF* versus ε and C under different kernel functions (linear, polynomial, radial basis ($\sigma = 1.00$)) respectively.

According to Fig. 3 (a, b, c), it is suitable to adopt the polynomial functional $K(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1)^2$ with $C = 2$ and $\varepsilon = 0.02$ in the model. The trained SVR model for T_m of $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds with original data is available as follows:

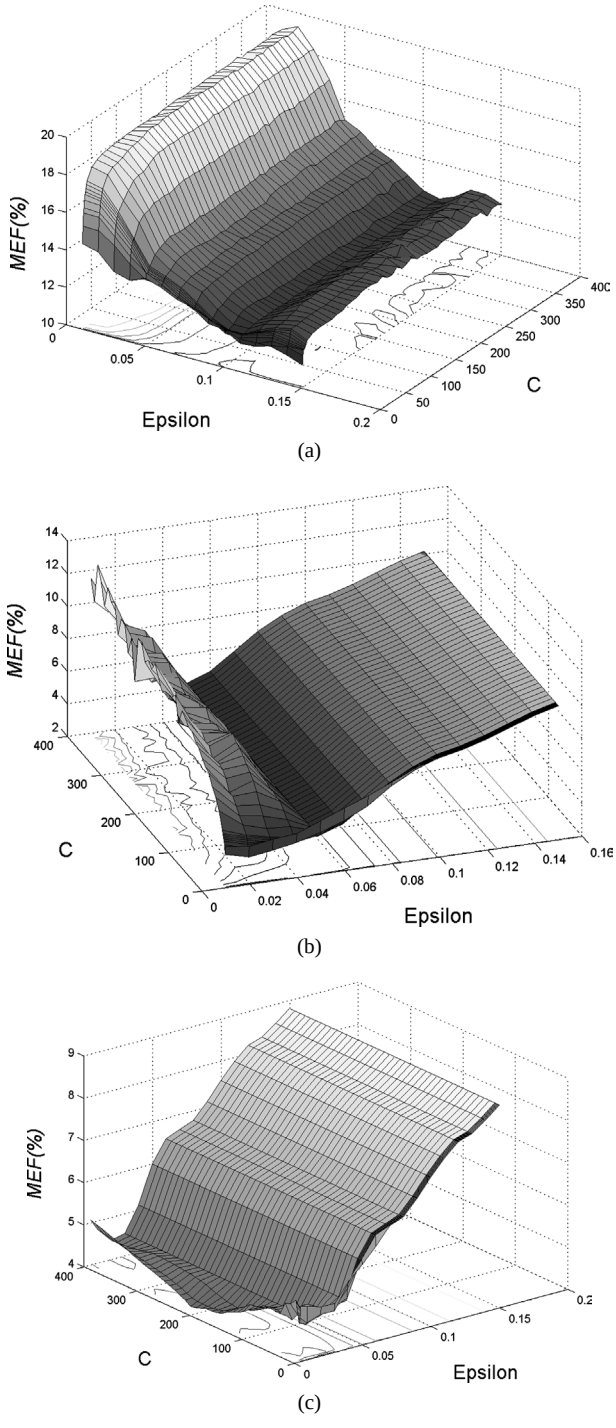

Fig. 3. (a) *MEF* versus C and ε using LOOCV with linear kernel function. (b) *MEF* versus C and ε using LOOCV with polynomial kernel function. (c) *MEF* versus C and ε using LOOCV with RBF kernel function ($\sigma = 1.00$).

$$T_m = 2300 \times \left[\sum_{i,j=1}^{25} (\alpha_i - \alpha_i^*) \big(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1 \big)^2 + 0.5711 \right] + 800,$$

where $(\alpha_i - \alpha_i^*)$ is Lagrange coefficient corresponding to support vector. Fig. 4 illustrates the relationship of predicted T_m and experimental T_m of $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds, with related coefficient (R) 0.99.

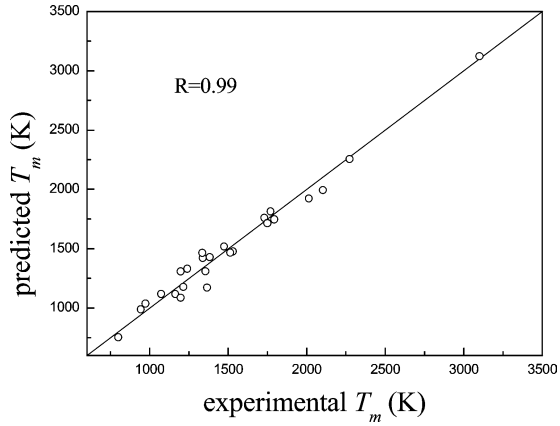

Fig. 4. Experimental T_m vs predicted T_m of binary compound semiconductors with trained SVR model.

3.2. $A^IB^{III}C^{VI}_2$ and $A^{II}B^{IV}C^{VI}_2$ ternary compounds

3.2.1. Data set

For ternary compounds, the data set consists of 31 compounds including $CuAlS_2$, $CuGaS_2$, $CuInS_2$, $CuAlSe_2$, $CuGaSe_2$, $CuInSe_2$, $CuAlTe_2$, $CuGaTe_2$, $CuInTe_2$, $AgAlS_2$, $AgGaS_2$, $AgAlSe_2$, $AgGaSe_2$, $AgInS_2$, $AgInSe_2$, $AgAlTe_2$, $AgGaTe_2$, $AgInTe_2$, $ZnSiP_2$, $ZnSiAs_2$, $ZnGeP_2$, $ZnGeAs_2$, $ZnSnP_2$, $ZnSnAs_2$, $ZnSnSb_2$, $CdSiP_2$, $CdGeP_2$, $CdSiAs_2$, $CdGeAs_2$, $CdSnP_2$ and $CdSnAs_2$. However, the number of samples decreases to 28 in the modeling of T_m due to the lack of T_m data of $AgAlS_2$, $AgAlSe_2$ and $ZnSnSb_2$.

3.2.2. Atomic parameters

Similarly, the features of modeling are some atomic parameters including lattice number u, Born exponent, atomic mass, atomic radius, electrovalent and electronegativity.

3.2.3. Model building for E_g of $A^IB^{III}C^{VI}_2$ and $A^{II}B^{IV}C^{VI}_2$ ternary compounds

In this work, the experimental E_g values of 31 ternary compounds were used to build the SVR model. Value of u [20,21], the average of proportion of electrovalent and radius $\overline{z/r_{\text{cov}}}$ [18], the average of atomic mass $\overline{m}$ and the average proportion of final ionization potential and atomic electrovalent $\overline{I_z/z}$ [15] were taken as features to model E_g [10,11], where

$$\overline{m} = \frac{m_A + m_B + 2m_C}{4}, \tag{19}$$

$$\overline{z/r_{\text{cov}}} = \frac{(z/r_{\text{cov}})_A + (z/r_{\text{cov}})_B + 2(z/r_{\text{cov}})_C}{4}, \tag{20}$$

$$\overline{I_z/z} = \frac{(I_z/z)_A + (I_z/z)_B + 2(I_z/z)_C}{4}. \tag{21}$$

Fig. 5 (a, b, c) illustrates *MEF* versus ε and C under different kernel functions (linear, polynomial, radial basis ($\sigma = 1.00$)), respectively.

From Fig. 5 (a, b, c), it is feasible to employ polynomial function with the form: $K(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1)^2$ with $\varepsilon = 0.07$ and $C = 50$ in the model. The trained model for E_g of $A^IB^{III}C^{VI}_2$ and $A^{II}B^{IV}C^{VI}_2$ ternary compounds with original

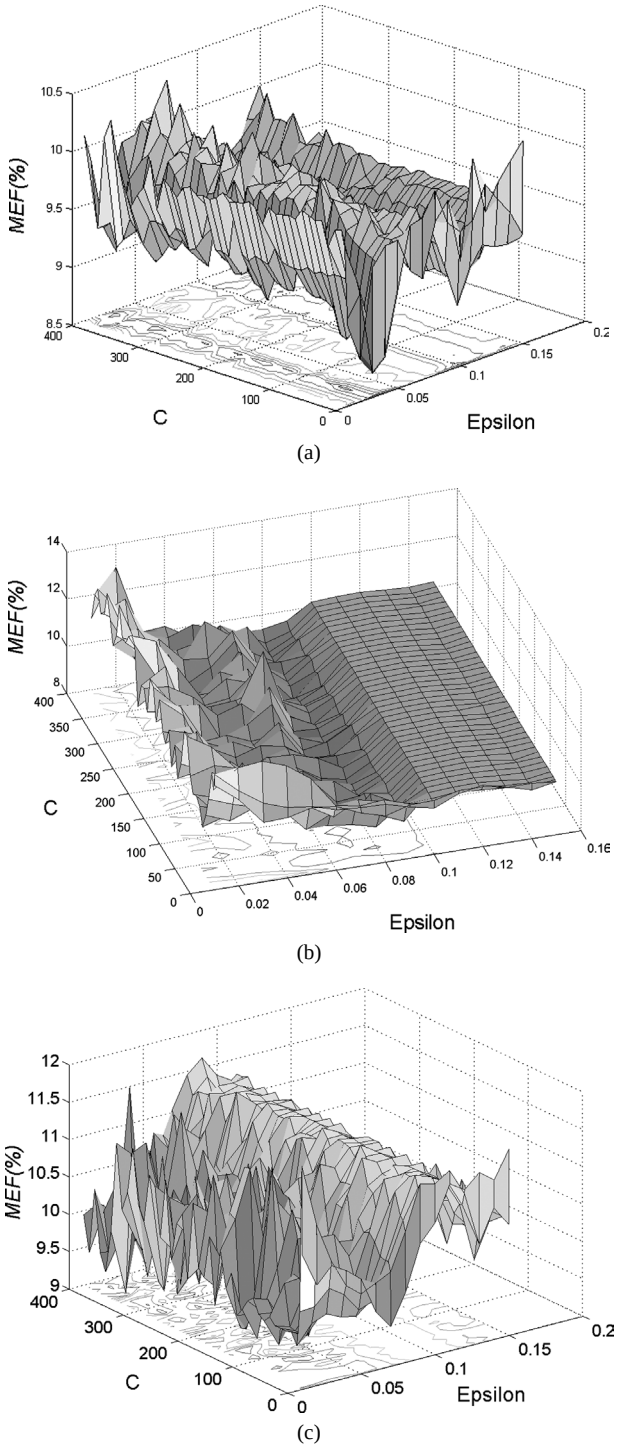

Fig. 5. (a) *MEF* versus C and ε using LOOCV with linear kernel function. (b) *MEF* versus C and ε using LOOCV with polynomial kernel function. (c) *MEF* versus C and ε using LOOCV with RBF kernel function ($\sigma = 1.00$).

data is available as follows:

$$E_g = 3.23 \times \left[\sum_{i,j=1}^{31} (\alpha_i - \alpha_i^*) \left(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1 \right)^2 + 0.000141 \right] + 0.26,$$

where $(\alpha_i - \alpha_i^*)$ is Lagrange coefficient corresponding to support vector. Fig. 6 illustrates the relationship of predicted E_g

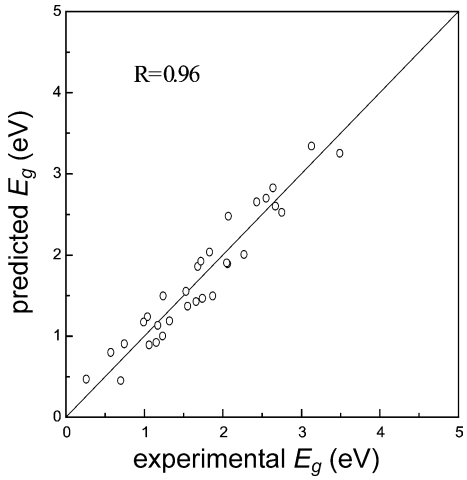

Fig. 6. Experimental E_g vs predicted E_g of ternary compound semiconductors with trained SVR model.

and experimental E_g of $A^{I}B^{III}C^{VI}_2$ and $A^{I}B^{III}C^{VI}_2$ ternary compounds. Fig. 6 illustrates the relationship of predicted E_g and experimental E_g of $A^{I}B^{III}C^{VI}_2$ and $A^{II}B^{IV}C^{VI}_2$ ternary compounds, with related coefficient (R) 0.96.

3.2.4. Model building for T_m of $A^{I}B^{III}C^{VI}_2$ and $A^{II}B^{IV}C^{VI}_2$ ternary compounds

In this work, the regression model for T_m is constructed using experimental data from a group of 28. The average of proportion of electrovalent and radius $\overline{z/r_{\mathrm{cov}}}$ [18], average of atomic mass $\overline{m}$, the average Born exponent $\bar{n}$ and difference of electronegativity in the Allred–Rochow scale of values $\Delta\chi$ are selected to model T_m [10,11,20,21], where

$$\bar{n} = \frac{n_A + n_B + 2n_C}{4}, \tag{22}$$

$$\Delta\chi = |2\chi_C - \chi_B - \chi_A|. \tag{23}$$

Fig. 7(a,b,c) illustrates *MEF* versus ε and C under different kernel functions (linear, polynomial, radial basis ($\sigma = 1.00$)) respectively. It is shown that using the RBF $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$ with $C = 1$ and $\varepsilon = 0.07$ is suitable in the model. The trained SVR model for T_m of $A^{I}B^{III}C^{VI}_2$ and $A^{II}B^{IV}C^{VI}_2$ ternary compounds with original data is as follows:

$$T_m = 730 \times \left[\sum_{i,j=1}^{28} (\alpha_i - \alpha_i^*) \times \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) + 0.5560 \right] + 840,$$

where $(\alpha_i - \alpha_i^*)$ is Lagrange coefficient corresponding to support vector. Fig. 8 illustrates the relationship of predicted T_m and experimental T_m of $A^{I}B^{III}C^{VI}_2$ and $A^{I}B^{III}C^{VI}_2$ ternary compounds.

3.3. Results of leaving-one-out cross-validation (LOOCV) of SVR model

In this work, the leaving-one-out cross validation (LOOCV) method was undertaken to evaluate the performances of the models obtained. As such, the data set of n samples was divided into two disjoint subsets including a training data set ($n - 1$ samples) and a test data set (only 1 sample). After developing each model based on the training set, the omitted data was predicted and the difference between experimental value and predicted value was calculated. Figs. 9–12 are plot of the predicted values employing LOOCV of SVR versus experimental values for E_g and T_m of binary and ternary compound semiconductor, respectively.

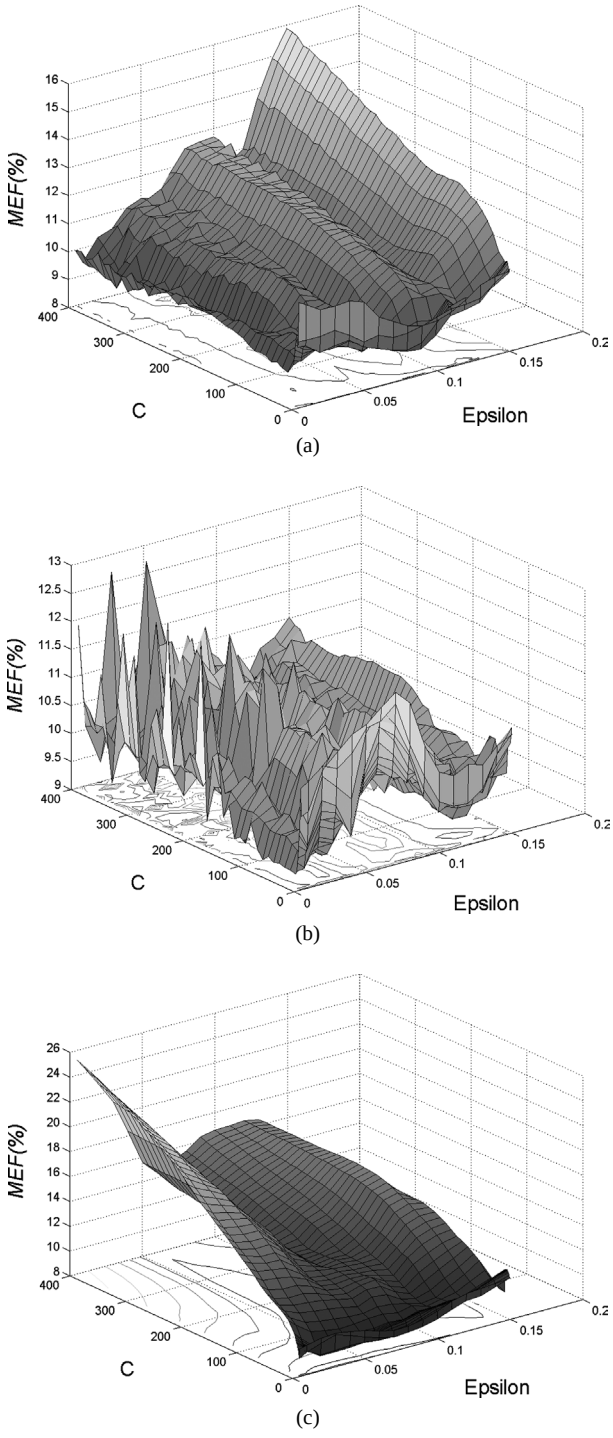

Fig. 7. (a) *MEF* versus C and ε using LOOCV with linear kernel function. (b) *MEF* versus C and ε using LOOCV with polynomial kernel function. (c) *MEF* versus C and ε using LOOCV with RBF kernel function ($\sigma = 1.00$).

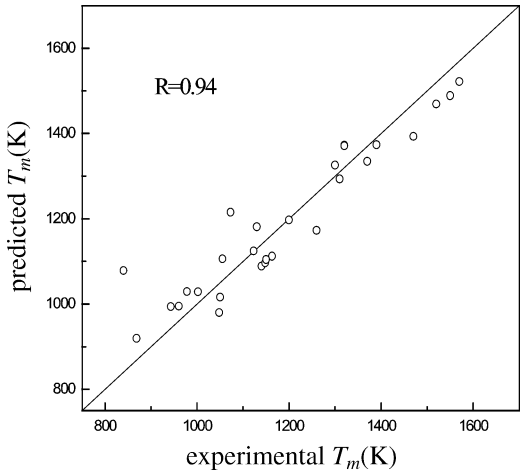

Fig. 8. Experimental T_m vs predicted T_m of ternary compound semiconductors with trained SVR model.

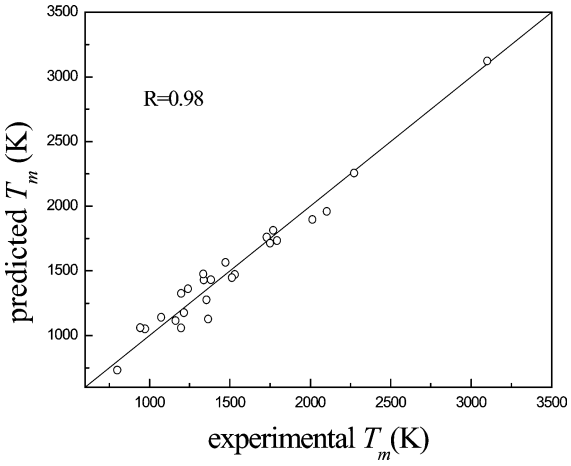

Fig. 10. Experimental T_m vs predicted T_m of binary compound semiconductors by using LOOCV of SVR.

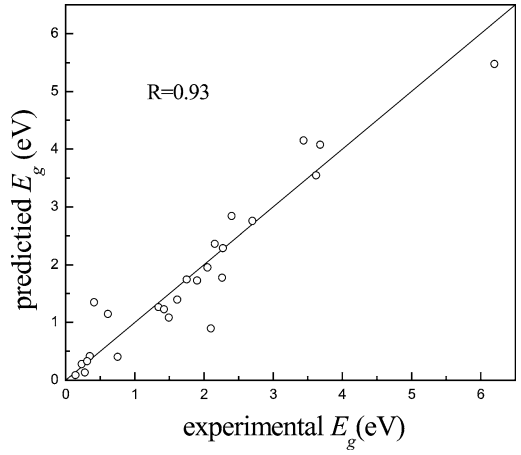

Fig. 9. Experimental E_g vs predicted E_g of binary compound semiconductors by using LOOCV of SVR.

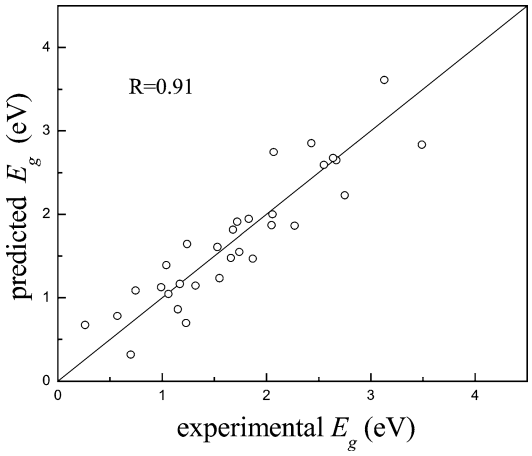

Fig. 11. Experimental E_g vs predicted E_g of ternary compound semiconductors by using LOOCV of SVR.

From Figs. 9–12, it can be concluded that the predicted results are in good agreement with experimental ones.

4. Discussion and conclusion

In a benchmark test, the support vector regression (SVR) was compared with several techniques of data mining including back propagation-artificial neural network (BP-ANN), multiple linear regression (MLR), and partial least squares regression (PLSR), with special consideration of their prediction abilities (generalization abilities) in LOOCV test. BP-ANN with three layers was used to build the model. The parameters of BP-ANN model used were as follows: the number of hidden nodes was two. The transformation function was sigmoid. In this work, the predictive errors were measured by *MEF* (Eq. (14)) and *MRE* (Eq. (24)).

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{p_i - e_i}{e_i}\right| \times 100\%, \tag{24}$$

where e_i is the experimental value of sample i, p_i is the predicted value of sample i, n is the number of the whole samples.

Table 1
The predicted error of E_g, T_m of binary and ternary compound by using LOOCV of different methods

Method	Binary compound semiconductors				Ternary compound semiconductors			
	E_g		T_m		E_g		T_m	
	MEF (%)	*MRE* (%)	*MEF* (%)	*MRE* (%)	*MEF* (%)	*MRE* (%)	*MEF* (%)	*MRE* (%)
SVR	5.87	39.31	3.63	6.26	8.36	22.31	9.07	5.90
BP-ANN	7.83	57.04	6.54	12.36	9.60	29.98	10.78	6.80
PLSR	8.71	56.59	11.65	18.60	9.73	30.82	10.72	6.90
MLR	7.87	55.63	13.26	20.91	9.59	30.21	10.73	6.86

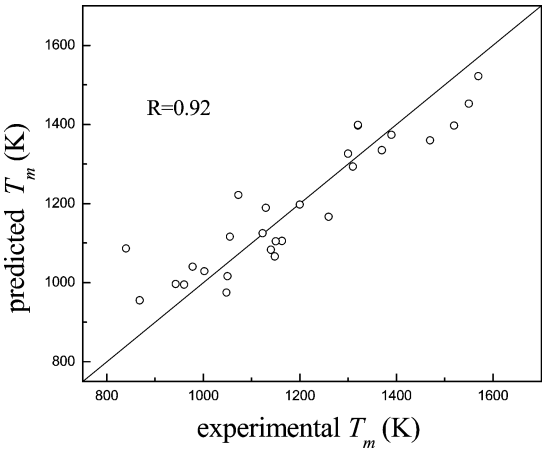

Fig. 12. Experimental T_m vs predicted T_m of ternary compound semiconductors by using LOOCV of SVR.

Table 1 lists the predictive errors for E_g and T_m of binary and ternary compound semiconductors by using LOOCV of SVR, BP-ANN, PLSR and MLR methods.

From Table 1, it can be concluded that the performance of SVR model outperforms those of BP-ANN, MLR, and PLS models for the data set available, which indicates that the SVR model has had better generalization ability. The SVR has been introduced as a robust and highly accurate intelligent regression technique, likely to be a useful chemometric tool. The SVM exhibits the better whole performance due to embodying the structural risk minimization principle, which has been shown to be superior to traditional empirical risk minimization principle, employed by conventional techniques of machine learning. It has the advantage over the other techniques of converging to the global optimum, and not to a local optimum that depends on the initialization and parameters affecting rate of convergence.

From the above results, it can be concluded that SVR is a good method for prediction of the band gaps (E_g) and melting points (T_m) of III–V, II–VI, I–III–VI_2 and II–IV–V_2 compound semiconductors. The computation of SVR model is faster compared with other machine learning techniques, because there are fewer free parameters and only support vectors (only a fraction of all data) are used in the generalization process. Since there are many tasks for generalization of useful information from small size of data sets in materials science. It can be expected that the SVM will be further applied in materials science.

Acknowledgement

We thank National Science Fund of China for the financial support of this work (No. 20373040).

References

[1] Z.C. Zhang, R.W. Peng, N.Y. Chen, Mater. Sci. Eng. B 54 (1998) 149.
[2] J. Hadamard, Lectures on the Cauchy's Problem in the Linear Differential Equations, Yale University Press, New Haven, 1923.
[3] V.N. Vapnik, Statistical Learning Theory, A Wiley-Interscience Publication, John Wiley and Sons, Inc., USA, 1998.
[4] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Comput. Chem. 26 (2001) 5.
[5] M. Trotter, B. Buxton, S. Holden, Meas. Control UK 34 (2001) 235.
[6] Y.D. Cai, X.J. Liu, Y.X. Li, X.B. Xu, K.C. Chou, Peptides 24 (2003) 665.
[7] Y.D. Cai, K.Y. Feng, Y.X. Li, K.C. Chou, Peptides 24 (2003) 629.
[8] N.Y. Chen, W.C. Lu, G.Z. Li, J. Yang, Support Vector Machine in Chemistry, World Scientific Publishing Co. Pte. Ltd., 2004.
[9] J. Platt, Sequential minimal optimization: A fast algorithm for training support vector machines, available at http://www.research.microsoft.com/users/jplatt/smo.html.
[10] M. Otfried, Semiconductors—Basic Data, Springer-Verlag, Berlin, 1996.
[11] R. Boca, Semiconductors Materials, CRC Press, New York, 1997.
[12] Y. Shizhong, Y. Shuren, K. Changhe, The Application of Semiconductor Materials, Press of Engineering Industry, Beijing, 1986.
[13] N.Kh. Abrikosov, Semiconducting II–VI, IV–VI and V–VI Compounds, Plenum Press, 1969.
[14] N.Y. Chen, et al., Pattern Recognition Applied to Chemistry and Chemical Technology, Science Press, Beijing, 2000.
[15] N.Y. Chen, W.C. Lu, R.L. Chen, P. Qin, P. Villars, J. Alloy. Compd. 289 (1999) 120.
[16] E. Mooser, W.B. Pearson, Phys. Rev. 101 (1956) 1608.
[17] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Machine Learning 46 (2002) 389.
[18] N.Y. Chen, Application of Bond Parameter Function, Press of Science, Beijing, 1976.
[19] L.C. Pauling, The Nature of the Chemical Bond, Cornell, Ithaca, 1960.
[20] H. Matsushita, S. Endo, T. Irie, Jpn. J. Appl. Phys. 30 (1991) 1181.
[21] A. Zunger, Appl. Phys. Lett. 50 (1987) 164.