Adv. Manuf. (2013) 1:151–159
DOI 10.1007/s40436-013-0025-2

Using support vector machine for materials design

Wen-Cong Lu · Xiao-Bo Ji · Min-Jie Li · Liang Liu · Bao-Hua Yue · Liang-Miao Zhang

Received: 20 April 2013 / Accepted: 7 May 2013 / Published online: 5 June 2013

Abstract Materials design is the most important and fundamental work on the background of materials genome initiative for global competitiveness proposed by the National Science and Technology Council of America. As far as the methodologies of materials design, besides the thermodynamic and kinetic methods combing databases, both deductive approaches so-called the first principle methods and inductive approaches based on data mining methods are gaining great progress because of their successful applications in materials design. In this paper, support vector machine (SVM), including support vector classification (SVC) and support vector regression (SVR) based on the statistical learning theory (SLT) proposed by Vapnik, is introduced as a relatively new data mining method to meet the different tasks of materials design in our lab. The advantage of using SVM for materials design is discussed based on the applications in the formability of perovskite or $BaNiO_3$ structure, the prediction of energy gaps of binary compounds, the prediction of sintered cold modulus of sialon-corundum castable, the optimization of electric resistances of VPTC semiconductors and the thickness control of In_2O_3 semiconductor film preparation. The results presented indicate that SVM is an effective modeling tool for the small sizes of sample sets with great potential applications in materials design.

Keywords Support vector machine · Materials genome initiative · Materials design · Data mining · Quantitative structure–property relationship · Materials exploration and optimization

W.-C. Lu (✉) · X.-B. Ji · M.-J. Li · L. Liu · B.-H. Yue · L.-M. Zhang
Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, People's Republic of China
e-mail: wclu@shu.edu.cn

1 Introduction

Over the last several decades, it is a great challenge for scientists to develop, manufacture, and deploy advanced materials as fast as possible. In June of 2011, the materials genome initiative (MGI) for global competitiveness was proposed by the National Science and Technology Council of America for the development of an infrastructure to shorten the materials development cycle. The most important and fundamental goal of MGI is to accelerate materials design through the use of computational capabilities, data management, and an integrated approach to materials science and engineering [1].

In principle, there are two strategies for materials design. One strategy is to start from the first principle, i.e., from quantum mechanics and statistical mechanics, to predict the properties of unknown materials. Although the first principle method has been widely used in materials design [2, 3], up to now, it is still impossible to solve most of complicated problems in materials exploration work by using this strategy solely. The other strategy is to start from the semi-empirical way, i.e., from the known data of some materials to find semi-empirical rules, which can be used to predict the properties of unknown materials. In general, the second strategy is more practicable than the first one in materials design or new materials exploration area, since a variety of data mining methods can be utilized to construct statistical models for a lot of data sets available from scientific experiments [4–9].

In the research of materials design by using data mining methods, principal component analysis (PCA), partial least

squares (PLS) and artificial neural networks (ANNs) are very helpful because of their relative good performance, speed, simplicity to construct statistical models [10, 11]. However, ANN may give rise to over-fitting problems [12] (i.e., may lead to good performance in fitting but poor performance in prediction) in treating finite, multivariate data set. At the meanwhile, nonlinear relations can only be modeled in limited way by using PCA or PLS algorithm [13].

In the semi-empirical method of materials design, the data of known materials is usually used as the training set. In most cases, the numbers of available known data in the training sets are rather limited, which means that the data processing tasks usually deals with the problem of small sample size, and hence cause serious over-fitting problem. As an effective way to overcome the problem of over-fitting, support vector machine (SVM) based on statistical learning theory (SLT) has been proposed by Vapnik [14]. SVM has been shown to perform well in various applications including drug design [15, 16], materials design [17–19] and chemistry researches [20].

In new materials exploration work, there are two questions with general significance in need of answers: the first question is "what is the chemical composition of the substance having desirable properties?", and the second one is "what are the optimal conditions of preparation or production for this material at low cost?". Since both of these two questions involve with very complicated systems or processes, we have to solve these problems by using some semi-empirical methods. To answer the first question, the relationships between the microscopic structure of materials and their properties need to be addressed. This relationship is usually known as quantitative structure–property relationship (QSPR). To deal with the second question, mathematical models are usually set up for the optimization of the processes.

Based on the problems concerning materials design, the tasks of materials design can be classified into four different categories. The first type of task is to solve the "formability problems", i.e., to find some mathematical model or criterion for the stability of some unknown substances. The second type of task is the "property prediction", i.e., to make mathematical models for the structure–property relationships and use these models to predict the properties of new materials (or the inverse problem: to search the unknown new materials with some pre-assigned properties). The third type of task is to solve the "optimization problems", i.e., to find the conditions for optimizing some properties of certain materials. The last but not the least type of task is to solve the "problem of control", i.e., to find the mathematical model to control some index of materials within a desired range. Different data mining techniques should be adopted for these different purposes. In this paper, we demonstrate some examples of applying SVM methods including support vector classification (SVC) and support vector regression (SVR) as a relatively new tool to meet the different tasks of materials design in our lab. The advantage of using SVM for materials design is discussed based on the applications presented.

2 Methods of SVM

The foundations of SVM have been developed by Vapnik [14] and are gaining popularity due to many attractive features, and promising empirical performance. In this paper the term SVM will refer to both SVC and SVR methods, which can be used for solving qualitative and quantitative problems respectively [21–25].

2.1 SVC

SVC has been recently proposed as a very effective method for solving classification problems, which can be restricted to consideration of the two-class problem without loss of generality [14, 20]. In this problem the goal is to separate the two classes by a classifier induced from available examples. It is expected that the classifier constructed has good performance on unseen examples, i.e., it generalizes well.

The geometrical interpretation of SVC is that it determines the optimal separating surface, i.e., a hyperplane, which is equidistant from two sets of data points. This hyperplane has many interesting statistical properties as discussed by Vapnik [14]. Consider the problem of separating the set of training vectors belonging to two separate classes, $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \cdots, (y_n, \boldsymbol{x}_n), \boldsymbol{x} \in \boldsymbol{R}^m, y \in -1, +1$, with a hyperplane

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0, \tag{1}$$

where $\boldsymbol{w}$ and b are the weight vector and bias, respectively.

If the training data are linearly separable, then there exists a pair of parameter set $(\boldsymbol{w}, b)$, for which we can write

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 \geq 0, \quad i = 1, 2, \cdots, l, \tag{2}$$

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \geq +1, \quad \text{for all } \boldsymbol{x} \in \boldsymbol{P}, \tag{3}$$

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \leq -1, \quad \text{for all } \boldsymbol{x} \in \boldsymbol{N} \tag{4}$$

where $\boldsymbol{P}$ is the set of positive sample, and $\boldsymbol{N}$ is the set of negative sample.

The decision rule is

$$f_{w,b}(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b). \tag{5}$$

The pair (w, b) can be rescaled without loss of generality

$$\min_{i=1,2,\cdots,l} \left| \boldsymbol{w}^{T}\boldsymbol{x}_i + b \right| = 1. \tag{6}$$

The learning problem is hence reformulated as follows. Let us minimize $\|\boldsymbol{w}\|^2$ subject to the constraints of linear separability. This is equivalent to maximizing the distance, normal to the hyperplane, between the convex hulls of two classes and the optimisation becomes a quadratic programming (QP) problem

$$\underset{\boldsymbol{w},b}{\text{Min}}\ \phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{7}$$

subject to $y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1, \quad i = 1, 2, \cdots, l$. This problem has global optimum, and the Lagrangian is written as

$$L(\boldsymbol{w}, b, \Lambda) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{t} \lambda_i [y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1], \tag{8}$$

where $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_l\}$ are the Lagrange multipliers, one for each data point. Hence we can write

$$F(\Lambda) = \sum_{i=1}^{t} \lambda_i - \frac{1}{2}\|\boldsymbol{w}\|^2 = \sum_{i=1}^{l} \lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j. \tag{9}$$

The Lagrange multipliers are only non-zero when $y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) = 1$. Vectors fulfilling this requirement are called support vectors since they lie closest to the separating hyperplane. Then, the optimal separating hyperplane is given as follows

$$\boldsymbol{w}^* = \sum_{i+1}^{l} \lambda_i^* \boldsymbol{x}_i y_i, \tag{10}$$

and the bias is given by

$$b^* = -\frac{1}{2}(\boldsymbol{w}^*)^{\mathrm{T}}(\boldsymbol{x}_{\mathrm{s}} + \boldsymbol{x}_{\mathrm{r}}), \tag{11}$$

where $\boldsymbol{x}_{\mathrm{r}}$ and $\boldsymbol{x}_{\mathrm{s}}$ are any support vectors from each class satisfying the following equation

$$y_{\mathrm{r}} = 1, \quad y_{\mathrm{s}} = -1. \tag{12}$$

The hard classifier is then

$$f(\boldsymbol{x}) = \mathrm{sgn}((\boldsymbol{w}^*)^{\mathrm{T}}\boldsymbol{x} + b^*). \tag{13}$$

In the case where a linear boundary is inappropriate, the SVC can map the input vector, $\boldsymbol{x}$, into a high dimensional feature space, $\boldsymbol{F}$. By choosing a non-linear mapping Φ, the SVC constructs an optimal separating hyperplane in this higher dimensional space. Among acceptable mappings are polynomials, radial basis functions and certain sigmoid functions. Then the optimisation problem becomes

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j) \rangle. \tag{14}$$

In this case, the decision function in SVC is as follows

$$\begin{aligned} g(\boldsymbol{x}) &= \mathrm{sgn}(f(\boldsymbol{x})) = \mathrm{sgn}\left(\sum_{i\in SV} \alpha_i y_i \langle \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}_i) \rangle + b\right) \\ &= \mathrm{sgn}\left(\sum_{i\in SV} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right), \end{aligned} \tag{15}$$

where the $\boldsymbol{x}_i$ is the set of support vectors and $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is called the kernel function.

2.2 SVR [14, 20]

In SVR, the basic idea is to map the data x into a higher-dimensional feature space $\boldsymbol{F}$ 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 $G = \{(\boldsymbol{x}_i, d_i)\}_{i=1}^{l}$ ($\boldsymbol{x}_i$ is input vector, and d_i is the desired value). SVR approximates the function in the following form

$$y = \sum_{i=1}^{l} w_i \Phi(\boldsymbol{x}_i) + b, \tag{16}$$

where $\{\Phi(\boldsymbol{x}_i)\}_{i=1}^{l}$ is the set of mappings of input features, and $\{w_i\}_{i=l}^{I}$ and b are coefficients. They are estimated by minimizing the regularized risk function $R(C)$:

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{17}$$

where

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon, & \text{for } |d - y| \geq \varepsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

and ε is a prescribed parameter.

In Eq. (17), $C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i)$ 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, $\frac{1}{2}\|\boldsymbol{w}\|^2$, 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. (17) to the following constrained function:

$$\mathrm{Min}R(\boldsymbol{w}, \xi, \xi^*) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C^*\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{19}$$

s.t.

$$\begin{cases} \boldsymbol{w}\Phi(x_i) + b - d_i \leq \varepsilon + \xi_i, \\ d_i - \boldsymbol{w}\Phi(x_i) - b \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0. \end{cases} \tag{20}$$

Thus, decision function Eq. (16) becomes the following form:

$$f(\boldsymbol{x},\alpha,\alpha^*)=\sum_{i=1}^{l}\left(\alpha_i-\alpha_i^*\right)K(\boldsymbol{x},\boldsymbol{x}_i)=b. \tag{21}$$

In Eq. (21), α_i,α_i^* are the introduced Lagrange multipliers. They satisfy the equality: $\alpha_i\cdot\alpha_i^*=0,\quad \alpha_i\geq 0,\quad i=1,2,\cdots,l$, and are obtained by maximizing the dual form of Eq. (19), which has the following form:

$$\begin{aligned} w(\alpha,\alpha^*)=&\sum_{i=1}^{l}d_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\\ &-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\boldsymbol{x}_i,\boldsymbol{x}_j), \end{aligned} \tag{22}$$

with the following constrains:

$$\begin{cases} 0\leq\alpha_i\leq C \quad i=1,2,\cdots,l,\\ 0\leq\alpha_i^*\leq C \quad i=1,2,\cdots,l,\\ \sum_{i+1}^{l}(\alpha_i-\alpha_i^*)=0. \end{cases} \tag{23}$$

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. (21), $K(\boldsymbol{x},\boldsymbol{x}_i)$is the kernel function. The value is equal to the inner product of two vectors $\boldsymbol{x}$ and $\boldsymbol{x}_i$ in the feature space $\Phi(\boldsymbol{x})$. That is, $K(\boldsymbol{x},\boldsymbol{x}_i)=\Phi(\boldsymbol{x})\Phi(\boldsymbol{x}_i)$. 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 $\Phi(\boldsymbol{x})$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function.

2.3 Implementation of SVM

According to the Ref. [14], the SVM software package *ChemSVM* including SVC and SVR has been programmed in our lab. The free version of *ChemSVM* can be downloaded on the website of Laboratory of Computational Chemistry in Shanghai University (http://chemdata.shu.edu.cn:8080/MyLab/Lab/download.jsp). The validation of the software has also been performed in the applications of chemistry [20].

3 Applications

3.1 SVC applied to the formability of perovskite or $BaNiO_3$ structure

The most exciting achievement of materials research is to find some new compound (or new phases) with specified structure and outstanding properties. In this work, the materials design problems of compounds with perovskite-type structures or $BaNiO_3$ structure will be discussed based on SVC model.

There are numerous complex oxides or halides with general formula ABX_3 (X = oxygen or halogen) having perovskite-type crystal structure and outstanding functional properties [26]. Since 1945, when the ferroelectric properties of barium titanate were discovered, a series of complex oxides and complex halides with perovskite-type structure have been found to be valuable functional materials. In recent years, searching new complex oxides and complex halides with perovskie-type structure has become an active research field of new materials exploration.

The crystal structure of compounds with ideal perovskite structure is illustrated in Fig. 1. It is the structure of a unit cell of $SrTiO_3$ crystal. In this structure, tetravalent Ti^{4+} cation is surrounded by 6 oxygen anions to form octahedral structure, and bivalent Sr^{2+} cation is surrounded by 12 oxygen anions to form cubo-octahedral structure. Based on the understanding of such type of crystal structure, Goldschmidt proposed a famous crystal-chemical criterion of the formability or the stability of perovskite structure for ABX_3-type compounds:

$$t=\frac{R_a+R_x}{\sqrt{2}(R_b+R_x)}, \tag{24}$$

where t is called tolerance factor. R_a, R_b are cationic radii of A, B ions respectively, and R_x is the ionic radius of anion of X. According to Goldschmidt, the cubic perovskite structure is stable only if the value of tolerance factor has an approximate range of $0.8<t<0.9$, and the distorted perovskite structure can be stable in a somewhat larger range of tolerance factor. This criterion is widely used in the exploration work of new compounds with perovskite-type or perovskite-like type structure. Owing to the accumulation of the crystallographic data of compounds with

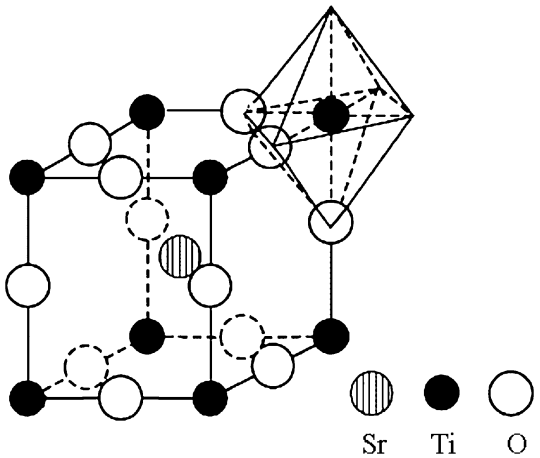

Fig. 1 Crystal structure of $SrTiO_3$, a typical compound with ideal perovskite structure

perovskite structure, it is now widely recognized that the range of tolerance factor for the stability of perovskite structure should be $0.75 < t < 1.00$.

Although Goldschmidt's tolerance factor t is indeed very useful for the exploration of new compounds with perovskite structure, it is only a necessary condition but not the sufficient condition for the formation or the stability of perovskite structure [27]. Many systems having t in the range of 0.75–1.0 do not form perovskite-type compound.

For example, $MnSiO_3$ has $t = 0.856$, but it has $CdGeO_3$ structure; $RbMnCl_3$ has $t = 0.88$, but it has hexagonal $BaTiO_3$ structure. NaI-MgI_2 system has $t = 0.826$, but it has no intermediate compound at all. Therefore it is desirable to investigate the complementary conditions for the formability of perovskite structure, in order to help the computerized materials design for new materials with perovskite structure. Atomic parameters and SVM technique can be used for this purpose.

Since BX_6 octahedra and AX_{12} cubo-octahedra are the basic sub-structures of perovskite lattice, as shown in Fig. 1, it is easy to see that the stability of BX_6 octahedra and AX_{12} cubo-octahedra are also necessary conditions for the stability of perovskite lattice. It is obvious that the condition $0.75 < t < 1.00$ is not enough to assure the stability of the BX_6 octahedral and AX_{12} cubo-octahedral structure. It is necessary to find the suitable criteria for the above-mentioned stability requirements.

Some ABX_3 type compounds, such as $RbNiCl_3$, although having $0.75 < t < 1.00$, do not form perovskite-type lattice but the crystal lattices with $BaNiO_3$ structure. The chief difference of $BaNiO_3$ structure (or hexagonal $BaTiO_3$ structure) from perovskite structure is that in $BaNiO_3$ structure (or hexagonal $BaTiO_3$ structure) the BX_6 octahdra are shared by their face with each other (see Fig. 2), while in perovskite structure they are shared with each other by their corners.

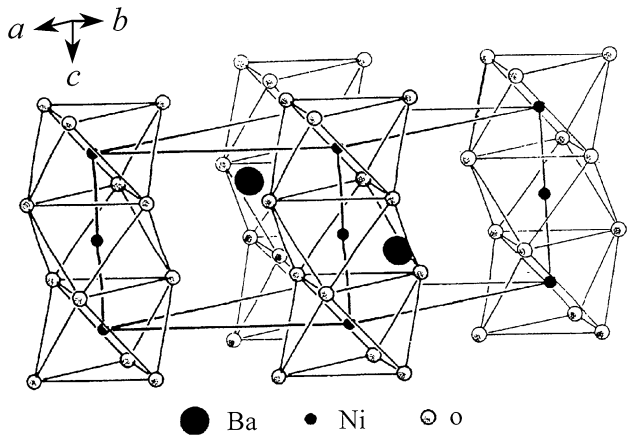

Fig. 2 Face-shared BX_6 structure in $BaNiO_3$ lattice

In order to find the criterion of relative stability between $BaNiO_3$ structure and perovskite structure, the data set containing 23 samples is used for data mining [28]. By using SVC combined with atomic parameters of compounds, the mathematical model of SVC was found to differentiate between perovskite structure and the hexagonal ABX_3 structures involving face-shared octahedral. The SVM model with linear kernel function manifests that 100 % of separation of the chlorides with perovskite structure and the chlorides with face-shared structures can be achieved. 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 by the model developed. In LOOCV test, the rate of correctness of prediction is 91 %. The criterion for the formation of face-shared structure found by SVC can be expressed as follows

$$4.52R_b - 1.83R_a + 2.23X_a - 0.142X_b - 4.10N_d + 0.589 < 0, \quad (25)$$

where R_a and R_b denote the Shanon-Preweit ionic radii of A and B in $ABCl_3$ respectively. X_a and X_b denote the Basanov electronegativity of A and B respectively. N_d denotes the number of d electrons in the unfilled shell of d electrons of B ions. It implies that large A^+ cation, small B^{2+} cation and large electronegativity of B favor the formation of face-sharing of BX_6 octahedra. This fact can be explained as follows: In perovskite-type lattice the network of corner shared BX_6 octahedra form cages of A^+ ions. The repulsive force due to the large A^+ and small cage formed by small B^{2+} and X^- will make perovskite-type lattice unstable, while the BX_6 octahedra in face-shared structure form parallel chains, and A^+ ions are located between these chains without strict confinement. Thus large A^+ and small B^{2+} favor the face-shared structure.

3.2 SVR applied to the prediction of energy gaps of binary compounds

III-V and II-VI binary compounds are important semiconductors for microwave, optoelectron and infrared devices. The band gaps (E_g) are essential properties of these compounds. It would be helpful for materials scientists to estimate the E_g of a compound before synthesizing it. On the basis of known data set available, it is reasonable to predict the properties of unseen samples by using data mining methods. Since there are a lot of data mining methods available, 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).

In this work, 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 [29, 30]. Based on the data set available, the SVR model for predicting E_g of $A^{III}B^{V}$ and $A^{II}B^{VI}$ binary compounds was constructed by using atomic parameters as features including electronegativity, valence, radius, atomic mass and their functions. The data mining results indicated that the sum of proportion of atomic electrovalent and covalent radius $\sum(z/r_{cov})$ [31], mean atomic number, $\bar{N}$ atomic electrovalent Z_A and Z_B should be selected as parameters in the model of band gap,

$$\text{where} \quad \bar{N} = \frac{N_A + N_B}{2}, \tag{26}$$

$$\sum (z/r_{cov}) = (z/r_{cov})_A + (z/r_{cov})_B. \tag{27}$$

In the present work, the 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) U_m as Eq. (28)

$$U_m = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - e_i|}{e_{max} - e_{min}} \times 100\,\%, \tag{28}$$

where e_i is the experimental value of sample i, p_i the predicted value of sample i, n the number of the whole samples. e_{max}, e_{min} are the maximum and minimum experimental value of whole samples respectively. In general, the smaller the value of U_m obtained, the better generalization ability expected. It is found that the optimal SVR model with the least U_m is available when the kernel function is polynomial form of $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\langle \boldsymbol{x}_i \cdot \boldsymbol{x}_j \rangle + 1\right)^2$, while $\varepsilon = 0.07$ 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

$$E_g = 3.479 \times \left(\sum_{i,j=1}^{25}\left(\alpha_i - \alpha_i^*\right)\left(\langle \boldsymbol{x}_i \cdot \boldsymbol{x}_j \rangle + 1\right)^2 + 0.7473\right) + 0.1410, \tag{29}$$

where $(\alpha_i - \alpha_i^*)$is the Lagrange coefficient corresponding to support vector. Figure 3 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) of 0.97.

In this work, the LOOCV method was also undertaken to evaluate the performances of the models obtained.

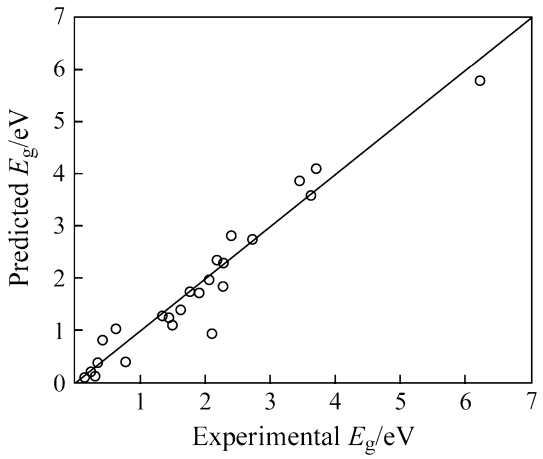

Fig. 3 Experimental E_g versus predicted E_g of binary compound semiconductors with trained SVR model

Figure 4 is the plot of the predicted values employing LOOCV of SVR versus experimental values for E_g of binary compounds.

From Fig. 4, it can be concluded that the predicted results are in good agreement to experimental ones [25].

3.3 SVR applied to the prediction of sintered cold modulus of sialon-corundum castable

Although the discovery of new materials is very exciting in materials research, the most part of tasks of materials research everyday is to try to improve the preparation technology of known materials. The economic effect of such kind of improvement is very significant because these efforts eventually determine the cost and the quality of products or the competitive ability in international market. Here the example of using SVR in materials optimization of preparing Sialon Ceramic will be described.

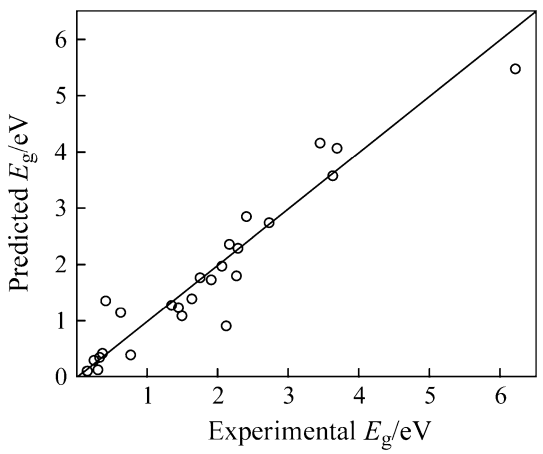

Fig. 4 Experimental E_g versus predicted E_g of binary compound semiconductors by using LOOCV of SVR ($R = 0.93$)

Sialons are silicon aluminium oxynitride ceramic materials with a range of technically important application, from cutting tools to specialized refractories. Furthermore, they can have a wide range of compositions and occur in several different families of crystal structures, the properties of sialons can be tailored for specific purposes [32]. β-sialon corundum find applications as high temperature, corrosion resistant, thermal shock resistant, high strength and toughness structural material [33]. The sintered cold modulus of sialon-corundum castable is an important property of sialon material, but the relationship between the property and process parameters is very complicated [34]. Hence it is necessary to find some computational methods to correlate the properties of sialon-corundum with their process parameters.

In this work, the data set consists of 20 samples from our experiments. The root mean square error (RMSE) V_{m} of LOOCV was adopted to estimate the quality of the model for predicting the sintered cold modulus of sialon-corundum castable. The V_{m} is defined as follows

$$V_{\mathrm{m}} = \sqrt{\frac{\sum_{i=1}^{n} (p_i - e_i)^2}{n}}, \tag{30}$$

where e_i is the experimental value of i sample, p_i the predicted value of i sample, n the number of the whole samples in LOOCV. Based on the data mining work, it was found that the linear kernel function with $C = 7.0$ and $\varepsilon = 0.16$ can be used to construct SVR model for the quantitative relationship of sintered cold modulus with their process parameters. Finally, the SVR model obtained can be presented as follows [19]

$$\begin{aligned} S_{\mathrm{pred}} = {} & -2.78C_{\mathrm{Water}} + 0.40C_{\mathrm{SiO_2}} - 1.35C_{\mathrm{Al_2O_3}} \\ & + 2.25C_{\mathrm{Disperser}} + 36.77, \end{aligned} \tag{31}$$

where S_{pred} means the experimental values of sintered cold modulus of sialon-corundum castable (strength, unit: MPa), while $C_{\mathrm{Water}}, C_{\mathrm{SiO_2}}, C_{\mathrm{Al_2O_3}}$ and $C_{\mathrm{Disperser}}$ are the content (mass %) of water, SiO_2 powder and ρ-Al_2O_3, dispersant and water respectively.

Figure 5 illustrates the experimental values versus predicted values of sialon-corundum cold modulus using LOOCV of SVR model with linear kernel ($C = 7.0$ and $\varepsilon = 0.16$).

According to the Eq. (31), in order to increase the S_{pred}, the content of SiO_2 and dispersant should be increased, at the mean while the content of ρ-Al_2O_3 should be decreased, which is consistent with the mechanism as follows. Addition of SiO_2 improves the flowability, but it reduces the added water content of castable. Simultaneously, SiO_2 reacts with water and then forms net structure of siloxene, which accelerates the sintering of castable

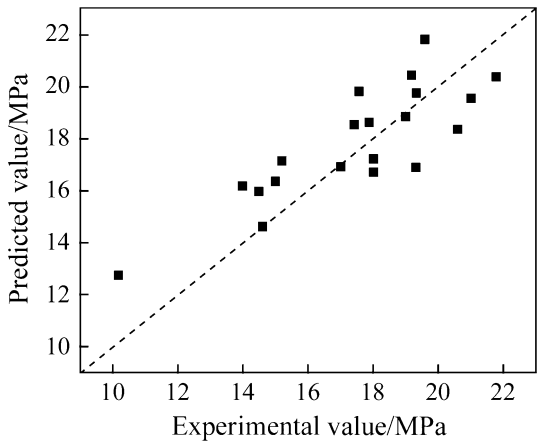

Fig. 5 Experimental values versus predicted values of sialon-corundum cold modulus using LOOCV of SVR model with linear kernel ($C = 7.0$ and $\varepsilon = 0.16$)

and improves its cold strength after sintering. However, ρ-Al_2O_3 which usually hydrates into $Al(OH)_3$ and AlOOH as bonder, cannot accelerate the sintering with water increasing, because mass agglomeration γ-Al_2O_3 exists in the commercial ρ-Al_2O_3, which holds lots of water resulting bad flowability. Therefore, the increase of ρ-Al_2O_3 restrains the sintering and reduces the sintered cold strength of castable. As for dispersant, it improves the flowability of castable by avoiding the flocculation structure of micelle and making water difficultly into this structure.

3.4 SVM applied to the optimization of electric resistances of VPTC semiconductors

VPTC materials are a kind of ceramic semiconductors for electronic uses. The task of the research work of VPTC materials is to search the optimal composition and the optimal preparation conditions for high value of $\rho_0/\rho_{\min}$ (the ratio of the electric resistance at zero degree centigrade to the minimum electric resistance) of these materials. There are five influencing factors including Yb_2O_3 content (W_1), excess TiO_2 (W_2), sintering temperature (T_c), sintering time (T_k), and relative cooling rate (V). By using linear kernel function ($C = 10$ and $\varepsilon = 0.15$), the trained SVR model for predicting $\rho_0/\rho_{\min}$ is available as follows

$$\begin{aligned} \rho_0/\rho_{\min} = {} & 29.52[W_1] + 1.315[W_2] + 0.03098[T_c] \\ & + 1.075[T_k] - 3.487[V] - 40.56. \end{aligned} \tag{32}$$

It is found that the relationship between the property of $\rho_0/\rho_{\min}$ and the five influencing factors is nearly linear one. Figure 6 shows the comparison between the experimental values and the predicted values of $\rho_0/\rho_{\min}$ by SVR in LOOCV test. By using the SVR model combined with pattern recognition techniques, some unseen samples can be designed with new compositions and technological

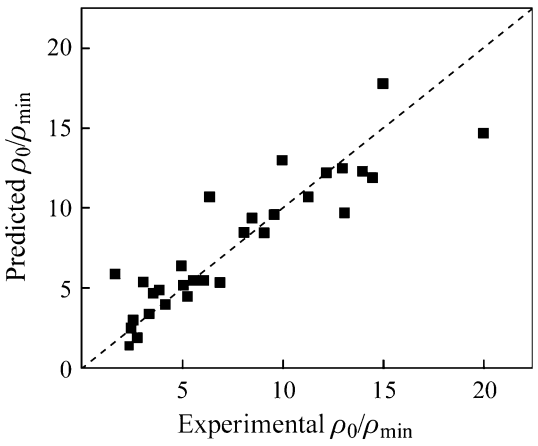

Fig. 6 Comparison between the experimental values and predicted values of ρ_0/ρ_{min} by SVR in LOOCV test

conditions for the optimization of VPTC semiconductors. The experimental results prove that the property ρ_0/ρ_{min} of new sample designed by using data mining increases to 27, which is much higher than that of the best sample ($\rho_0/\rho_{min} = 21$) obtained before the optimization.

3.5 SVM applied to the thickness control of In_2O_3 semiconductor film preparation

In_2O_3 semiconductor nanometer film is a new material for combustible gas detector uses. It can be prepared by sol-gel method. How to control the thickness of the semiconductor film is one of the crucial problems in the preparation work. There are several factors influencing the thickness of film: the mass percentage of In_2O_3 and PVA in the bath, the viscosity of coating liquids, the drawing rate and the drawing number in preparation. So it is desirable to have a mathematical model for the automatic control in the film production. In our lab, SVM methods have been used for data mining of this purpose.

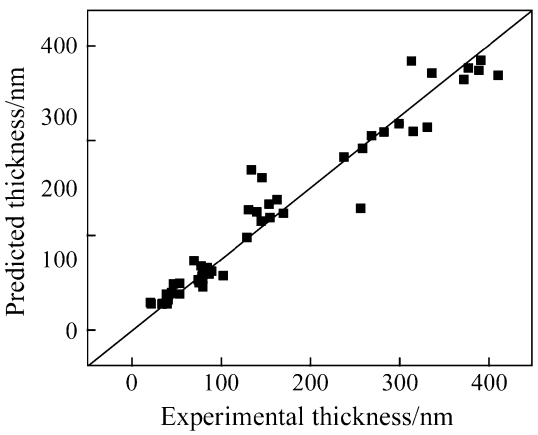

Fig. 7 Result of prediction of thickness of In_2O_3 film by SVR in LOOCV test

It has been found that the SVR with polynomial kernel of second degree can make the mathematical model for the thickness control of the semiconductor films. Figure 7 shows the comparison between the experimental thickness data and the predicted thickness in LOOCV test [35].

4 Discussion and conclusions

Generally speaking, how to choose the right balance between model flexibility and over-fitting to a limited training set is one of the most difficult obstacles for obtaining a model with good generalization ability to predict properties of materials. In the computation of SVM model, it should be noted that the selection of appropriate value for the regularization parameter C is very important because of its possible effects on both trained and predicted results, since it controls the tradeoff between maximizing the margin and minimizing the training error. Usually, C should be optimized for fear of neither under-fitting nor over-fitting. It is also noticed that the predicted results are largely affected by the kernel functions and its parameters adopted.

It should be emphasized that the advantage of SVM is workable with a small size of sample set. In many cases, obtaining a sufficient number of experimental samples is still time-consuming and costly in the development of novel materials. Therefore, efficient learning from a limited number of samples becomes increasingly important for shortening the materials development cycle.

Although our research results indicate that the performance of SVM outperforms those of traditional data mining methods, it should be realized that different data mining methods would have their own advantages and disadvantages in different applications. Sometimes the best approach is a combination of different methods since the complementary approaches can provide helpful information from different point of views.

From the examples introduced in this paper, it can be concluded that the SVM is an effective modeling tool with great potential in materials design. Therefore, it can be expected that the SVM method will be further applied in various fields of materials science.

Acknowledgments Financial supports to this work from the National Natural Science Foundation of China (Grant No. 21273145) and the 085 Project of Materials Genome Engineering of Shanghai University are gratefully acknowledged.

References

1. National Science and technology Coucil (2011) Materials genome initiative for global competitiveness, Washington DC, America, June 24, 2011

2. Choi YM, Lin MC, Liu ML (2010) Rational design of novel cathode materials in solid oxide fuel cells using first-principles simulations. J Power Sources 195(5):1441–1445
3. Ceder G (2010) Opportunities and challenges for first-principles materials design and applications to Li battery materials. MRS Bull 35(9):693–701
4. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O (2013) The high-throughput highway to computational materials design. Nat Mater 12(3):191–201
5. Kong CS, Rajan K (2012) Rational design of binary halide scintillators via data mining. Nucl Instrum Methods Phys Res A 680(1):145–154
6. Suh C, Kim K, Berry JJ, Lee J, Jones WB (2010) Data mining-aided crystal engineering for the design of transparent conducting oxides materials research society fall meeting. Cambridge University Press, Cambridge
7. Liu X, Lu WC, Peng CR, Su Q, Guo J (2009) Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO_3 perovskites. Comp Mater Sci 46(4):860–868
8. Gu TH, Lv W, Shao X, Lu WC (2012) Detection of high energy materials using support vector classification, Adv Mater Res 554–556:1628–1631
9. Wu ML, Zhang LM, Gu TH, Qian N, Ma WJ, Lu WC (2013) Shape-controlled synthesis and pattern recognition of dendritic Co_3O_4 superstructures. Adv Mater Res 652–654:352–355
10. Liu HL, Guo J, Chen NY (1996) A PLS-BPN pattern recognition method applied to computer-aided materials design. Anal Lett 29(2):341–350
11. Chen NY, Li CH, Qin P (1998) KDPAG expert system applied to materials design and manufacture. Eng Appl Artif Intell 11(5): 669–674
12. Patterson DW (1996) Artificial neural networks: theory and applications. Prentice Hall, New Jersey
13. Wold S, Sjostroma M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab 58(2):109–130
14. Vapnik VN (1998) Statistical learning theory. Wiley, New York
15. Burbidge R, Trotter M, Buxton B, Holden S (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. J Comput Chem 26(1):5–14
16. Lu WC, Dong N, Nάray-Szabó G (2005) Predicting anti-HIV-1 activities of HEPT-analog compounds by using support vector classification. QSAR Comb Sci 24(9):1021–1025
17. Li J, Qi M, Kong J, Wang J, Yan Y, Huo W, Yu J, Xu R, Xu Y (2010) Computational prediction of the formation of microporous aluminophosphates with desired structural features. Micropor Mesopor Mat 129(1–2):251–255
18. Yan Q (2012) Prediction of porosity of porous NiTi alloy from processing parameters based on SVR. Adv Mater Res 393–395:231–235
19. Liu X, Lu WC, Jin SL, Li YW, Chen NY (2006) Support vector regression applied to materials optimization of sialon ceramics. Chemometr Intell Lab 82(1–2):8–14
20. Chen NY, Lu WC, Yang J, Li GZ (2004) Support vector machine in chemistry. World Scientific Publishing Company, Singapore
21. Niu B, Lu WC, Yang SS, Cai YD, Li GZ (2007) Support vector machine for SAR/QSAR of phenethyl-amines. Acta Pharmacol Sin 28(7):1075–1086
22. Zhu JX, Lu WC, Liu L, Gu TH, Niu B (2009) Classification of Src kinase inhibitors based on support vector machine. QSAR Comb Sci 28(6–7):719–727
23. Yang SS, Lu WC, Gu TH, Yan LM, Li GZ (2009) QSPR study of *n*-octanol/water partition coefficient of some aromatic compounds using support vector regression. QSAR Comb Sci 28(2): 175–182
24. Liu X, Chen HC, Liu TA, Li YL, Lu ZR, Lu WC (2007) Application of PCA-SVR to NIR prediction model for tobacco chemical composition. Spectrosc Spect Anal 27(12):2460–2463
25. Gu TH, Lu WC, Bao XH, Chen NY (2006) Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. Solid State Sci 8(2):129–136
26. Galasso FS (1990) Perovskites and high Tc superconductors. Wiley, New York
27. Liu L, Lu WC, Chen NY (2004) On the criteria of formation and lattice distortion of perovskite-type complex halides. J Phys Chem Solids 65(5):855–860
28. Müller O, Roy R (1974) The major ternary structural families. Springer, Berlin
29. Madelung O (1996) Semiconductors—basic data. Springer, Berlin
30. Boca R (1997) Semiconductors Materials. CRC Press, New York
31. Chen NY (1976) Application of bond parameter function. Press of Science, Beijing
32. MacKenzie KJD, Temuujin J, Smith ME, Okada K, Kameshima Y (2003) Mechanochemical processing of sialon compositions. J Eur Ceram Soc 23(7):1069–1082
33. Kudyba-Jansen AA, Hintzen HT, Metselaar R (2001) The influence of green processing on the sintering and mechanical properties of β-sialon. J Eur Ceram Soc 21(12):2153–2160
34. Li YW, Zhang X, Jin SL (2001) Corundum castables containing nitrogen for purging plug in Ladle. In: Proceedings of 44th international colloquium on refractories, pp 26–27, Aachen, Germany
35. Bao XH, Pan QY, Chen NY (2002) Support vector regression model for controlling the thickness of semiconductor In_2O_3 film. Comput Appl Chem 19(6):733–736