Journal of Hazardous Materials 168 (2009) 962–969

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

A novel QSPR model for prediction of lower flammability limits of organic compounds based on support vector machine

Yong Pan, Juncheng Jiang*, Rui Wang, Hongyin Cao, Yi Cui

Jiangsu Key Laboratory of Urban and Industrial Safety, Institute of Safety Engineering, Nanjing University of Technology, Nanjing 210009, China

ARTICLE INFO

Article history:
Received 27 October 2008
Received in revised form 21 February 2009
Accepted 23 February 2009
Available online 6 March 2009

Keywords:
Quantitative structure–property relationship
Lower flammability limit
Genetic algorithm
Support vector machine

ABSTRACT

A quantitative structure–property relationship (QSPR) study is suggested for the prediction of lower flammability limits (LFLs) of organic compounds. Various kinds of molecular descriptors were calculated to represent the molecular structures of compounds, such as topological, charge, and geometric descriptors. Genetic algorithm was employed to select optimal subset of descriptors that have significant contribution to the overall LFL property. The novel chemometrics method of support vector machine was employed to model the possible quantitative relationship between these selected descriptors and LFL. The resulted model showed high prediction ability that the obtained root mean square error and average absolute error for the whole dataset were 0.069 and 0.051 vol.%, respectively. The results were also compared with those of previously published models. The comparison results indicate the superiority of the presented model and reveal that it can be effectively used to predict the LFL of organic compounds from the molecular structures alone.

1. Introduction

In chemical industry and engineering, there is a wide application of various physicochemical data. For example, risk assessment calculations often require a wide range of physicochemical parameter inputs. Similarly, in process design, material and energy balances must be based on accurate data to properly size equipment and determine utility consumption and cost. As a result, reliable and accurate data of physicochemical properties are always required and also considered to be absolutely necessary. However, in the practical industry process, the required data are often absent when needed, especially for the properties that related to the combustion such as the flash point, the auto-ignition temperature and the lower and upper flammability limits.

The lower flammability limit (LFL), which is usually in percentage volume (vol.%) at 298 K, is defined as the minimum concentration of a combustible substance that is capable of propagating a flame through a homogeneous mixture of the combustible and the air under the specified conditions of test. LFL is one of the most important indices used to rate the flammability and combustibility of chemical substances in the chemical industries. Knowledge of LFL values is essential to maximize safety in process design and operational procedures, such as starting up a reactor without passing through a flammable range, and storing or shipping the flammable product safely. So, reliable and accurate LFL data are always required and also considered to be absolutely necessary in the practical industry process.

Experimental LFL values are the most accurate and main source of the LFL data used in production. However, as we known, the experimental LFL is not absolute, but depend on several factors, such as the geometry of the apparatus, the type and strength of the ignition source, the test pressure and temperature, the degree of mixing, and so on [1]. Accordingly, the measured LFL values reported by different literatures are often inconsistent, sometimes quite different. Besides, the measurement of LFL is expensive and time consuming, sometimes even impossible. Consequently, in order to support and expand the LFL dataset used for industry, the development of theoretical prediction methods which are desirably convenient and reliable for predicting the LFL is required.

There have already been several methods reported in the literatures for predicting the LFL of organic compounds, which can be classified into several categories containing group contribution models [2], empirical correlations [3–7], and the quantitative structure–properties relationship (QSPR) models [8]. These methods have been extensively reviewed by Vidal et al. [1] and Albahri [9].

The most important disadvantage of group contribution models is their limitations in use. For example, the applicability range of these models are too related to the studied dataset, and the new chemicals with functional groups not included in those used for the model development will be out of the model applicability range and thus will not be predicted. Moreover, group contribution

* Corresponding author at: Mail Box 186, No.5 Xinmofan Road, Nanjing University of Technology, Nanjing 210009, China. Tel.: +86 25 83587305; fax: +86 25 83587411.
E-mail addresses: yongpannjut@163.com (Y. Pan), ypnjut@126.com, jcjiang@njut.edu.cn (J. Jiang), hades44@126.com (R. Wang), daqiao517@sohu.com (H. Cao), flyan0208@163.com (Y. Cui).

0304-3894/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2009.02.122

models also provide a weak ability in distinguishing the isomeric compounds.

The empirical correlations also suffer from some important disadvantages. Firstly, the use or application of these models requires unconventional physicochemical properties, and the availability or the lack of which may result in some limitations on their applicability range. Moreover, the prediction accuracy of these models is quite dependent on the accuracy of the needed physicochemical properties.

A current trend in predicting the physicochemical properties is the use of quantitative structure–properties relationship (QSPR) method. QSPR is a mathematical method that relates the properties of interest to the molecular structures of compounds which are represented by a variety of molecular descriptors. Molecular descriptors are various molecular-based theoretical parameters which can be calculated using known mathematical algorithms solely from molecular structures. Several molecular descriptors are always selected as the QSPR input to correlate the desired property of compounds with special principles. The QSPR method possesses some obvious advantages. Firstly, the number of descriptors used in the QSPR method is almost always lower than that used in the group contribution method for the same studied dataset. This fact may bring on more robust models. Secondly, the descriptors used in the QSPR models have definite physical meanings, which would be useful to probe the physicochemical information that has significant contribution to the targeted properties. Thirdly, because only theoretical descriptors derived solely from the molecular structure would be involved and have continuous values, the QSPR models developed should theoretically be applicable to any organic compound. Consequently, the QSPR methods have been widely used in predicting various physicochemical properties, such as boiling point, melting point, flash point, vapor pressure, critical properties, water solubility, auto-ignition temperatures, octanol/water coefficients, and so on, which have been extensively reviewed elsewhere [10–14].

In QSPR studies, the selection of appropriate modeling techniques which can be applied for construction of model is one of the key problems involved. At present, many different technologies, such as multiple linear regression (MLR), partial least squares (PLS), and different types of artificial neural networks (ANNs) have been widely used in the QSPR modeling, which can be used for inspection of linear and nonlinear relation between interested property and molecular descriptors, respectively. However, as we known, the linear method is much limited for a complex nonlinear system. Meanwhile, the neural networks also suffered some disadvantages inherent to its architecture, such as overtraining, overfitting, network optimization, and reproducibility of results. Due to these reasons above, a more accurate and informative modeling technique which can be effectively used in QSPR analysis is desirably needed.

The support vector machine (SVM) is recently developed from the machine learning community by Vapnik and co-workers [15,16]. As a novel type of machine learning method, SVM is gaining popularity due to many attractive features and promising empirical performance. Originally, SVM was developed for classification problems, and has demonstrated a good performance in solving these problems by numerous successful applications [17–22]. In recent years, with the introduction of ε-insensitive loss function, SVM has also been extended to solve regression problems, and has shown great performance in QSPR studies due to its remarkable ability to interpret the nonlinear relationships between molecular structure and properties [23–28]. In the most of these cases, the performance of SVM modeling either matches or is significantly better than that of traditional machine learning approaches [25].

In the present work, the main aim was to establish a new QSPR model for predicting the LFL of organic compounds from their molecular structures by using SVM techniques. The performance of this model was compared with those obtained by MLR, PLS and ANN methods as well as those of previous works.

2. Support vector machine

Support vector machine is a new type of machine learning method developed for solving both classification and regression problems. We focus here on SVM for regression problems, the task studied in this work. Theories of SVM for regression can be found in the tutorials for SVM [16], and here we will take only a brief description of the SVM in the following.

For a given regression problem, the basic idea of SVM is to map the input vectors X onto a very high-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 $G = \{(x_i, d_i)\}_{i=1}^{l}$ (x_i is input vector, d_i is the desired value). SVM approximates the function in the following form

$$f(x) = \sum_{i=1}^{l} w_i \cdot \phi_i(x) + b \tag{1}$$

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

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

where

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

In Eq. (2), $C(1/N)\sum_{i=1}^{N} L_\varepsilon(d_i, y_i)$ is the so-called empirical error, which is measured by ε-insensitive loss function $L_\varepsilon(d, y)$. $(1/2)\|w\|^2$ is used as a measurement of function flatness. C is a regularized constant determining the trade-off between the training error and the model flatness. When introducing slack variables ξ and ξ^*, Eq. (2) can be written as below:

$$\begin{aligned} Max\ R(w, \xi^*) = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (y_i - wx - b \leq \varepsilon \\ & + \xi_i, wx + b - y_i \leq \varepsilon + \xi_i^*, \xi_i, \xi_i^* \geq 0) \end{aligned} \tag{4}$$

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

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x, x_i) + b \tag{5}$$

where $K(x, x_i)$ is the kernel function. A given kernel function corresponds to the inner product in transformed space, that is, the inner product in that transformed space is equivalent to a kernel function of the input space ($K(x, x_i) = (\phi(x) \cdot \phi(x_i))$). The kernel function can effectively solve the contradiction between high dimension and computing complexity, and is thus a great progress in the development of SVM.

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.

The performance of SVM for regression depends on the combination of several parameters. They are kernel function type and

its corresponding parameters, capacity parameter C, and ε of ε-insensitive loss function. There are four possible choices of kernel functions, such as linear, polynomial, sigmoid, and radial basis function. For the regression problems, the radial basis function kernel is commonly used because of its effectiveness and speed in training. For the radial basis function kernel, the most important parameter is the width γ of the radial basis function, which controls the amplitude of the kernel function and, therefore, controls the generalization ability of SVM. C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data [29]. The optimal value for ε depends on the type of noise present in the data and the number of resulting support vectors. ε-insensitivity prevents the entire training set meeting boundary conditions and so the value of ε can affect the number of support vectors used to construct the regression function. The bigger ε, the fewer support vectors are selected.

In this study, we used the grid search method to find the optimum values for C, γ and ε. The leave-five-out cross-validation was employed to determine the optimal parameters, and the set of values with the best leave-five-out cross-validation performance, which is scaled by the mean square error (MSE), was selected as the optimal and final parameters for further analysis.

The overall performance of SVM model was evaluated in terms of the average absolute error (AAE) and root mean square error (RMSE), which were calculated with the following equation:

$$\mathrm{AAE} = \frac{\sum_{i-1}^{n} \left| y_i - y_o \right|}{n} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_o)^2}{n}} \tag{7}$$

where y_i is the observed value, y_o is the predicted value, and n is the number of compounds in the dataset.

The internal predictive capability of SVM model was evaluated by leave-one-out cross-validation (Q^2_{LOO}) on the training set, which was calculated with the following equation [30]:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{training} (y_i - y_o)^2}{\sum_{i=1}^{training} (y_i - \bar{y})^2} \tag{8}$$

where y_i, y_o, and $\bar{y}$ were respectively the observed, predicted, and mean observed LFL values of the compounds in the training set.

The external predictive capability of SVM model on the external test set was evaluated by Q^2_{ext}, which was calculated with the following equation [30]:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test} (y_i - y_o)^2}{\sum_{i=1}^{test} (y_i - \bar{y}_{tr})^2} \tag{9}$$

where y_i and y_o were the observed and predicted LFL values of the compounds in the test set, and $\bar{y}_{tr}$ was the mean observed LFL values of the compounds in the training set.

3. Materials and methods

3.1. Dataset

The dataset used in this study was taken from the work of Gharagheizi [8]. This set consists of a diverse set of 1038 organic compounds, which includes hydrocarbons, halogenated compounds, alcohols, ethers, esters, aldehydes, ketones, carboxylic acid, amines, amides, nitriles, nitro compounds, heterocyclic compounds and compounds with multiple functional groups. The LFL values of these compounds range from 0.185 to 3.6 vol.%. A complete list of the compounds and its corresponding observed LFL values are presented as supplementary materials.

The dataset is randomly divided into a training set and an external test set. The training set is used for model development, while the external test set is used for model validation. A QSPR model cannot be verified for its predictivity by checking only a few compounds, as in such cases the results could be obtained by chance and it is impossible to obtain general conclusions [30]. Consequently, the model must be tested on a sufficiently large number of compounds not used in the model development (at least 20% of the complete dataset is recommended [30]. So in this work, the whole dataset is randomly divided into a training set with 830 compounds (80% of the dataset) and a test set with 208 compounds (20% of the dataset).

3.2. Descriptor calculation and reduction

In QSPR studies, compounds must be represented using molecular descriptors. A wide variety of descriptors have been reported for QSPR analysis [31,32], such as topological descriptors, geometrical descriptors, electrostatic descriptors and quantum chemical descriptors. In the present work, the molecular descriptors used to search for the best model of the LFL prediction are calculated by the Dragon program (web version 2.1) [33], which is a sophisticated program for the calculation of molecular descriptors. All the calculation are on the basis of the minimum energy molecular geometries optimized by the HyperChem 7.5 Package based on MM+ molecular mechanics force field and AM1 semi-empirical method. The detailed description on the types of the molecular descriptors that Dragon can calculate and the procedure of calculation of the descriptors can refer to Dragon software user's guide [33]. In all, a total of 1481 descriptors were calculated for each compound in the dataset.

After the calculation of molecular descriptors, those stayed constant and near constant for all molecules were removed from the descriptor pool, since those descriptors were not encoding the structural differences between compounds that accounts for their different LFL values. Further reduction of the descriptor pool was attained by examining pairwise correlations between descriptors so that only one descriptor was retained from a pair contributing similar information (correlation coefficient >0.95 in this study). Finally, a total set of 578 remaining descriptors are achieved and used to select optimal subset of descriptors that have significant contribution to the LFL property.

3.3. Genetic algorithm based descriptor selection

The basic strategy of QSPR analysis is to find optimum quantitative relationships between the molecular descriptors and desired property, which can be then used for the prediction of the property from only molecular structures. One of the most important problems involved in QSPR studies is to select optimal subset of descriptors that have significant contribution to the desired property. The well-known genetic algorithm is just a well-accepted method for solving this kind of problems.

Genetic algorithm (GA) is a powerful optimization method to search for the global optima of solutions. This algorithm is developed to mimic some of the processes observed in natural evolution. The detailed description of which can be found in Ref. [34]. In recent years, GA has been successfully applied to feature selection in QSPR studies [8,10,24,35–40]. In this study, the GA, along with partial least squares (PLS) method (GA-PLS), was employed to find the optimal subset of descriptors that accurately represented the relationships between molecular structure and LFL. GA-PLS is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variable

Table 1
Descriptors selected for the SVM model for prediction of LFL.

Descriptor	Type	Definition
AAC	Topological descriptors	Mean information index on atomic composition
PW5	Topological descriptors	Path/walk 5 - Randic shape index
SIC0	Topological descriptors	Structural information content (neighborhood symmetry of 0-order)
GATS1v	2D autocorrelations	Geary autocorrelation - lag 1/weighted by atomic van der Waals volumes

selection. This algorithm was presented by Leardi and Lupiáñez [41] for the first time. In this study, the GA-PLS programs are implemented using the software package PLS_Genetic Algorithm Toolbox written by Leardi and Lupiáñez [41]. In this procedure, the chromosome and its fitness in the species correspond to a set of variables and internal prediction of the derived PLS model, respectively. Selection of useful variables is based on their frequency of occurrence in the best models obtained for each program. The used parameters of GA-PLS and detailed description of how to use GA-PLS can be found in the work of Leardi and Lupiáñez [41].

Before modeling, it must be indicated that the used GA-PLS software cannot be applied when the number of starting variables are greater than 200. This is due to the fact that a higher variables/compounds ratio may increase the risk of overfitting. In the presented study, this limitation is overcome by performing the work of "heats and finals", which can be described as following:

(1) Firstly, splitting the remaining 578 variables into three random groups of about 200 variables each. Then, a total of three "heats" is achieved.
(2) Running GA-PLS on each "heats", and for each "heats" a set of selected variables is retained.
(3) Combining the selected variables for each "heats" together, and a much smaller set of variables are achieved. If they are less than 200, the GA-PLS can be run on them (the "final"). If not, the variables can be split again and GA-PLS can be run on the different groups (the "semifinals"), and then the "final" can be achieved.

3.4. Software

In the present study, all calculation programs implementing GA-PLS are written in M-file by using Matlab V. 4.0, and the SVM model is implemented based on the shareware program Libsvm V.2.84 [42]. All the calculations involved in this study are performed on a 2.4 GHz Intel Pentium IV with 1 GB RAM under windows XP.

4. Results and discussion

4.1. Results of descriptor selection

GA-PLS procedure was performed on the training set to select the optimal set of descriptors. Since the GA is mainly a stochastic algorithm, the results of different GA applications would therefore be slightly different. In order to get more consistent results, the GA process needs to repeat many times to give a more reliable model. In this work, all the GA process were repeated 100 times and the selection of useful variables was based on their frequency of occurrence in the models with the maximal C.V. % (Cross-validated explained variance) obtained for each operation. The frequency was calculated by the following equation:

$$\text{Frequency}(i) = \frac{\text{the total number of descriptor}(i)\text{ selected by GA-PLS}}{\text{the times of operation using GA-PLS}} \quad (10)$$

where i was the ith descriptor.

The descriptors with higher frequency were considered as more important in identifying the molecular structures that have significant contribution to the overall LFL property. In this particular work, the descriptors with a frequency above 90% in each 100 operations were considered to be important. With this criterion, a set of four descriptors were finally selected and used to build the following model of SVM. The types and definitions of these descriptors were presented in Table 1, while their corresponding values for the whole 1038 compounds were presented as supplementary materials.

As can be seen from Table 1, all the four descriptors selected in the model are 2D descriptors, which could also be calculated from the Simplified Molecular Input Line Entry System (SMILES). The physical meanings of these descriptors are interpreted as following.

AAC is a topological descriptor to describe each atom by its own atom type and the bond types and atom types of its first neighbors. This descriptor is related to molecular complexity in terms of atom types.

PW5 is a topological descriptor related to molecular shape, such as the molecular geometry and molecular branching.

SIC0 is a topological descriptor to measure the degree of the diversity of the elements in the molecule and also describe the molecular shape.

GATS1v is a 2D autocorrelation descriptor, which is obtained from molecular graphs, by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag). This descriptor is mainly related to the atomic van der Waals volumes of a molecule, which is one of the primary dimensional features of the chemicals.

All the four descriptors selected are mainly related to the dimensional features of the molecules like molecular shape and

Table 2
Performance comparison between models obtained by SVM, MLR, PLS, and ANN.

Model	Training set				Test set			
	R^2	Q^2_{LOO}	RMSE	AAE	R^2	Q^2_{ext}	RMSE	AAE
SVM	0.979	0.979	0.068	0.050	0.979	0.977	0.076	0.054
MLR	0.971	0.971	0.079	0.061	0.976	0.975	0.079	0.062
PLS	0.971	0.971	0.079	0.061	0.976	0.975	0.079	0.062
ANN	0.977[a] 0.976[b]	–	0.075[a] 0.075[b]	0.057[a] 0.057[b]	0.977	0.973	0.082	0.061

[a] Derived from 554 training samples.
[b] Derived from 554 training samples plus 276 validation samples.

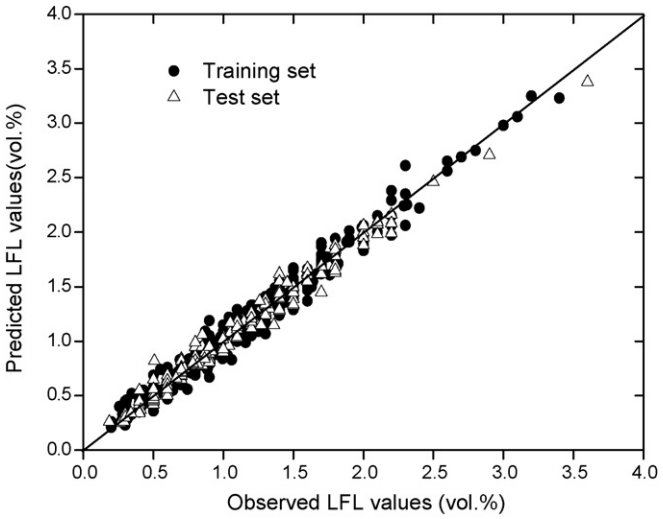

Fig. 1. Correlation between the predicted and observed LFL values for both the training and test sets.

complexity. Therefore, the overall LFL property of organic compounds can be reasonably explained by their steric effects.

Also, it was noteworthy that the descriptors selected in this study were quite agreed with those selected in the work of Gharagheizi [8], with three topological descriptors (AAC, PW5, SIC0) in common. This further validates the conclusion made previously that the molecular shape and complexity have positive relationships with the overall LFL property of compounds.

4.2. Results of SVM model

The powerful modeling method of SVM is then employed to investigate the possible nonlinear relation between the selected descriptors and the LFL values. The data are scaled in range [−1 to 1] before modeling, and the leave-five-out cross-validation is employed on the training set to determine the optimal parameters (ε, γ, and C). The resulting parameters of SVM model are fixed as follows: $C = 4$, $\varepsilon = 0.03125$, $\gamma = 0.5$. The corresponding number of support vectors is 303. With the optimal SVM model, the LFL values of the compounds in the test set were predicted for external validation. Finally, the predicted LFL values for all 1038 organic compounds are obtained and presented in the supplementary materials. The main statistical parameters of the obtained model are shown in Table 2. A plot of the predicted LFL values versus the observed ones for both the training and test sets is shown in Fig. 1.

4.3. Results analysis and interpretation

As can be seen from Table 2, for the SVM model, the resulting AAE values for both training and test sets are within the experimental error of LFL determination, which is around ±0.1 vol.% [1]. Meanwhile, it is noteworthy that the RMSE values are not only low but also as similar as possible for the training and test sets, which suggests that the proposed model has both predictive ability (low values) and generalization performance (similar values) [30].

Moreover, the predicted percentage error of all the 1038 organic compounds was calculated. The obtained average percentage error (APE) for these compounds was 5.60%, while the maximum percentage error was 61.86%. The results were shown in detail in Fig. 2. As can be seen from Fig. 2, the percent error of 622 organic compounds is less than 5%, which is more than half of the 1038 organic compounds used in the presented work.

Thus, it can be reasonably concluded that: (1) the SVM method can effectively investigate the nonlinear relationship existed between molecular structure and LFL property, which showed that SVM is a very promising tool for the QSPR studies. (2) The four descriptors selected by the GA-PLS approach can account for the structural features of the compounds related to the LFL property, which indicates that the GA-PLS approach is a very effective method for variable selection. (3) A new QSPR model would have been developed, which could be successfully used to predict the LFL of compounds with an accuracy that can approach the accuracy of experimental LFL determination.

4.4. Model validation

In order to further analysis the model stability, the obtained model was tested for chance correlation. A y-scrambling experiment was performed in which the dependent variables were scrambled. This y-scrambling was repeated 100 times. As expected, the models generated would produce high RMSE values with the minimum RMSE of 1.21 and 1.34 for the training and test sets, respectively. These errors were much higher than the errors calculated when the dependent variables were not scrambled. It can be thus concluded that only the correct dependent variable can be used to generate reasonable models, and the chance correlation had little or no effect in the presented model.

Also, the residuals of the predicted values of the LFL against the observed values for the model were listed in Fig. 3. As most of the calculated residuals are distributed on both sides of the zero line, one may conclude that there is no systematic error in the development of the present model.

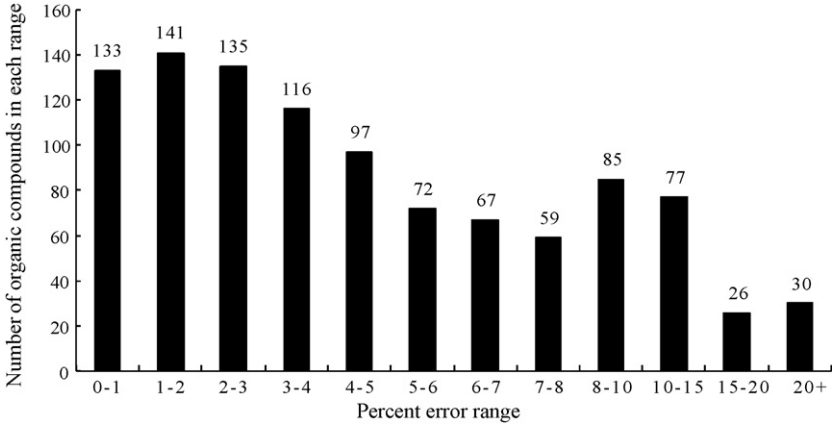

Fig. 2. The percent errors obtained by the presented model and the number of organic compounds in each range.

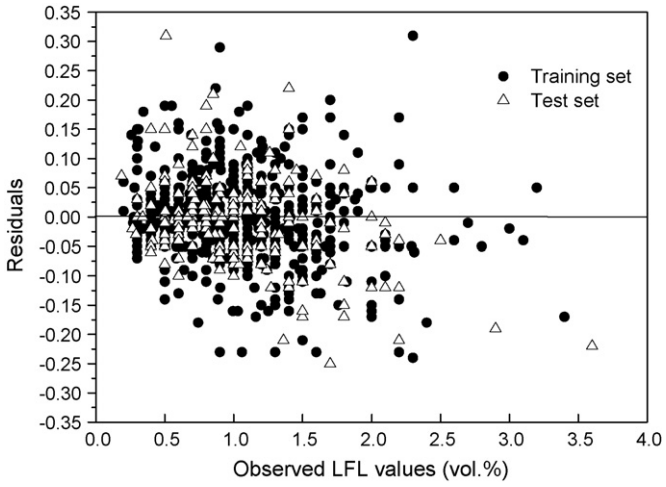

Fig. 3. Plot of the residuals versus the observed LFL values for the SVM model.

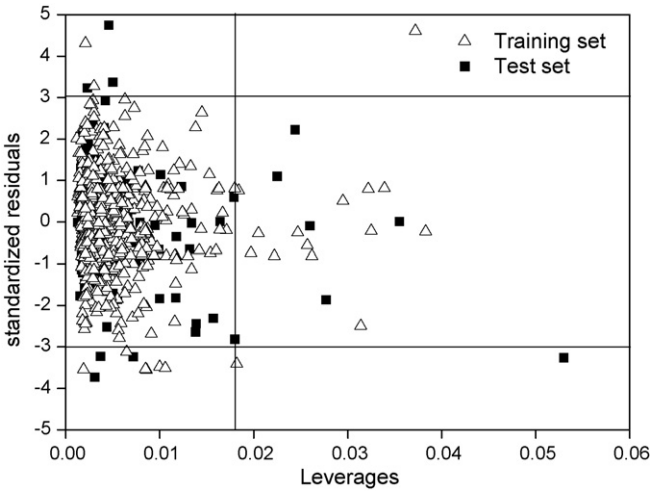

Fig. 4. The Williams plot describing the applicability domain of the SVM model (h^* = 0.018).

All the results discussed above showed that the presented SVM model is a valid model and can be effectively used to predict the LFL of organic compounds.

4.5. Definition of the applicability domain of the model

Once a QSPR model is obtained, another crucial problem is the definition of its applicability domain (AD). For any QSPR model, only the predictions for chemicals falling within its AD can be considered reliable and not model extrapolations.

There are several methods for defining the AD of QSPR models [43], but the most common one is determining the leverage values for each compound [30]. To visualize the AD of a QSPR model, the plot of standardized residuals versus leverage values (h) (the Williams plot) was exploited in this study, which played a double role. Firstly, it described the impacts of the objects on models by the values of their leverages. Leverage indicates a compound's distance from the centroid of X. The leverage of a compound in the original variable space is defined as [44]:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (11)$$

where x_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) is defined as [43]:

$$h^* = \frac{3p}{n} \quad (12)$$

where n is the number of training compounds, p is the number of model variables plus one. The leverage (h) greater than the warning leverage (h^*) suggested that the compound was very influential on the model. Secondly, it presented the Euclidean distances of the compounds to the model measured by the cross-validated standardized residuals. The cross-validated standardized residuals greater than three standard deviation (s) units classified the compound as a response outlier.

The Williams plot for the presented SVM model was showed in Fig. 4. From this plot, the applicability domain is established inside a squared area within ±3 standard deviations and a leverage threshold h^* of 0.018. For making predictions, predicted LFL data must be considered reliable only for those compounds that fall within this AD on which the model was constructed. It can be seen from Fig. 4 that the majority of compounds in the dataset are inside of this area. However, two compounds (compounds ethene and decamethyltetrasiloxane) in the training set with $h > h^*$ and the standardized residuals >3s, as do one of the test set (compound chloroethene). They are both response outliers and high leverage chemicals. Meanwhile, 13 compounds in the training set with $h > h^*$ and the standardized residuals <3s, as do five of the test set. However, the 13 compounds in the training set fit the model well, thus they can stabilize the model and make it more precise, which implies that they should not be considered outliers but influential compounds. Also, it can be concluded from the five compounds in the test set that the developed SVM model has good generalizability and predictivity for the compounds with descriptor values significantly far from the centroid of the descriptor space. Moreover, eight compounds in the training set and six compounds in the test set are wrongly predicted (>3s), but with lower leverage values ($h < h^*$). These erroneous predictions could probably be attributed to wrong experimental data rather than to molecular structures [30].

4.6. Comparison with other models

4.6.1. Comparison with MLR, PLS and ANN models

The MLR, PLS and Back-Propagation NN (BPNN) methods were also employed to describe the relation between LFL and the selected descriptors. By using the same four descriptors and the same 830 training samples used for SVM modeling, the optimal MLR, PLS and ANN models were obtained and the corresponding results were shown in Table 2. Then all the three developed models were used to predict the LFL values of the compounds in the test set, which have not been used for the model development. The obtained predicted LFL values for all 1038 organic compounds were presented as supplementary materials. The main statistical parameters of the three models were also shown in Table 2. As can be seen from the table, the results of SVM are a little better than those of MLR, PLS and ANN for both training and test sets, but for all the latter three models, the obtained AAE values for both training and test sets are also within the experimental error of LFL determination. This further validates the conclusion made previously that the four descriptors selected in this study can properly account for the structural features of the compounds related to the LFL property no matter what particular modeling methods are employed.

4.6.2. Comparison with previous models

General comparisons have also been made between the SVM model and previous models. As we known, the various models had been developed based on different dataset and different methods, and each model possesses its own advantages and disadvantages.

So it is suggested that not only the prediction results but also more other important characteristics of models should be taken into account and analyzed, such as the applicability efficiency and applicability range of models. Consequently, detailed comparisons between the SVM model and some previous works are presented as follows.

Suzuki [4] developed a nonlinear empirical model between the LFL and the standard enthalpies of combustion of a diverse set of organic compounds. The reported standard error for 112 compounds was 0.23, which is definitely worse than the presented SVM model. Moreover, it must be noted that a big set of 11 compounds that show large deviation have been regarded as outliers and already removed from the final model. When the 11 reported outliers are included, the corresponding standard error will increase to 0.35, which is more than 400% higher than that of the presented model. In addition, compared with the work of Suzuki [4] our model: (1) includes more compounds (830 versus 112); (2) has been externally validated with compounds not used in model development; (3) can be used to predict the LFL of unknown compounds solely from the molecular structure without requiring any extra information on physicochemical properties.

As the work of Suzuki and Ishida [5], a general comparison is also made with the presented SVM model. Regarding the input parameters used in the models, the models of work [5] employed four physicochemical parameters, while the presented model employs four molecular descriptors which can be directly calculated from the molecular structure. Moreover, these theoretical descriptors have definite physical meanings, which are useful to probe the physicochemical information that has significant contribution to the LFL property. Regarding the statistical parameters of the models, both the MLR and NN models of work [7] were worse in terms of *AAE* and *RMSE* than the presented model. Moreover, the presented model is developed based on larger number of compounds in the dataset (830 versus 144), and also more compounds are employed in the test set for model external validation (208 versus 50). Finally, regarding the applicability range of the models, the use of the models of work [7] requires extra data of needed physicochemical properties, and if only one of the needed properties is missing, calculation cannot be performed to predict the LFL. Oppositely, because only theoretical descriptors derived solely from the molecular structure is involved, the presented model would theoretically be used to reliably predict the LFL for any organic compound belonging to its applicability domain.

As the work of Gharagheizi [8], in which a QSPR model had also been developed for prediction of LFL, a general comparison can also be made with the presented SVM model. Regarding the input parameters used in the models, both models employed four molecular descriptors which can be directly calculated from the molecular structure. Regarding the statistical parameters of the models, the presented SVM model is obviously better in terms of *AAE* and *RMSE* than the model of work [8]. However, regarding the applicability efficiency of the models, the model of work [8] is a simple MLR model, which can be easily understood and expediently applied, while the presented SVM model is more complicated and professional knowledge needed. In addition, the model of work [8] is developed based on a little larger number of compounds in the dataset (845 versus 830), and also external validated by employing more compounds in the test set (211 versus 208). Finally, regarding the applicability range of the models, both models were QSPR models, which would theoretically be used to reliably predict the LFL for any organic compound belonging to their applicability domains. However, the AD of the presented SVM model has been verified, while the one of the model of work [8] not.

5. Conclusion

In the present work, based on the novel modeling technique of support vector machine, a new QSPR model has been developed for predicting the LFL of a diverse set of organic compounds from the molecular structure alone. By performing the model validation, it can be concluded that the presented SVM model is a valid model and can be effectively used to predict the LFL of organic compounds with an accuracy that can approach the accuracy of experimental LFL determination. Moreover, the mechanism of the model was interpreted, and the applicability domain of the model was defined. When comparing the results of the model to those of previously published models, it showed that the presented model possesses some obvious superiority. Thus it can be reasonably concluded that the proposed model would be expected to predict LFL for new organic compounds or for other organic compounds for which experimental values are unknown. Additionally, the presented method could also identify and provide some insight into what structural features are related to the LFL property of organic compounds.

Acknowledgements

This research is supported by National Natural Science Fund of China (No. 50774048), Research Fund for the Doctoral Program of Higher Education of China (No. 200802910007), and Program for New Century Excellent Talents in University (No. NCET-05-0505). Y. Pan acknowledges the support of Jiangsu Graduate Scientific Innovation Projects (No. CX07B_150z). The authors also would like to thank professor Riccardo Leardi of University of Genoa for providing genpls matlab code program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhazmat.2009.02.122.

References

[1] M. Vidal, W.J. Rogers, J.C. Holste, M.S. Mannan, A review of estimation methods for flash points and flammability limits, Process Saf. Prog. 23 (2004) 47–55.
[2] W.H. Seaton, Group contribution method for predicting the lower and the upper flammable limits of vapors in air, J. Hazard. Mater. 27 (1991) 169–185.
[3] D.A. Crowl, J.F. Louvar, Chemical Process Safety, Fundamentals with Applications, Prentice Hall, Englewood Cliffs, NJ, 1990.
[4] T. Suzuki, Note: Empirical relationship between lower flammability limits and standard enthalpies of combustion of organic compounds, Fire Mater. 18 (1994) 333–336.
[5] T. Suzuki, M. Ishida, Neural network techniques applied to predict flammability limits of organic compounds, Fire Mater. 19 (1995) 179–189.
[6] F. Hshieh, Predicting heats of combustion and lower flammability limits of organosilicon compounds, Fire Mater. 23 (1999) 79–89.
[7] L.G. Britton, Using heats of oxidation to evaluate flammability hazards, Process Saf. Prog. 21 (2002) 31–54.
[8] F. Gharagheizi, Quantitative structure–property relationship for prediction of the lower flammability limit of pure compounds, Energy Fuels 22 (2008) 3037–3039.
[9] T.A. Albahri, Flammability characteristics of pure hydrocarbons, Chem. Eng. Sci. 58 (2003) 3629–3641.
[10] A.R. Katritzky, D.C. Fara, How chemical structure determines physical, chemical, and technological properties: an overview illustrating the potential of quantitative structure–property relationships for fuels science, Energy Fuels 19 (2005) 922–935.
[11] A.R. Katritzky, V.S. Lobanov, M. Karelson, QSPR: the correlation and quantitative prediction of chemical and physical properties from structure, Chem. Soc. Rev. 24 (1995) 279–287.
[12] A.R. Katritzky, U. Maran, V. Lobanov, M. Karelson, Structurally diverse quantitative-structure–property relationship correlations of technologically relevant physical properties, J. Chem. Inf. Comput. Sci. 40 (2000) 1–18.
[13] D.L. Yaffe, A Neural Network Approach for Estimating Physicochemical Properties using Quantitative Structure–Property Relationships (QSPRs), University of California, Los Angeles, 2001.
[14] J. Taskinen, J. Yliruusi, Prediction of physicochemical properties based on neural network modeling, Adv. Drug Deliv. Rev. 55 (2003) 1163–1183.

[15] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[16] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[17] A.I. Belousov, S.A. Verzakov, J.V. Frese, A flexible classification approach with optimal generalisation performance: support vector machines, Chemomet. Intell. Lab. Syst. 64 (2002) 15–25.
[18] L.H. Chiang, M.E. Kotanchek, A.K. Kordon, Fault diagnosis based on fisher discriminant analysis and support vector machines, Comput. Chem. Eng. 28 (2004) 1389–1401.
[19] A. Kulkarni, V.K. Jayaraman, B.D. Kulkarni, Support vector classification with parameter tuning assisted by agent-based technique, Comput. Chem. Eng. 28 (2004) 311–318.
[20] A. Kulkarni, V.K. Jayaraman, B.D. Kulkarni, Knowledge incorporated support vector machines to detect faults in Tennessee Eastman Process, Comput. Chem. Eng. 29 (2005) 2128–2133.
[21] A. Niazi, J. Ghasemi, M. Zendehdel, Simultaneous voltammetric determination of morphine and noscapine by adsorptive differential pulse stripping method and least-squares support vector machines, Talanta 74 (2007) 247–254.
[22] J. Wang, H.Y. Du, X.J. Yao, Z.D. Hu, Using classification structure pharmacokinetic relationship (SCPR) method to predict drug bioavailability based on grid-search support vector machine, Anal. Chim. Acta 601 (2007) 156–163.
[23] P.C. Lima, A. Golbraikh, S. Oloff, Y. Xiao, A. Tropsha, Combinatorial QSAR modeling of P-glycoprotein substrates, J. Chem. Inf. Model. 46 (2006) 1245–1254.
[24] M.H. Fatemi, S. Gharaghani, A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine, Bioorg. Med. Chem. 15 (2007) 7746–7754.
[25] M.H. Fatemi, S. Gharaghani, S. Mohammadkhani, Z. Rezaie, Prediction of selectivity coefficients of univalent anions for anion-selective electrode using support vector machine, Electrochim. Acta 53 (2008) 4276–4282.
[26] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines, J. Hazard. Mater. 151 (2008) 603–609.
[27] Y. Pan, J.C. Jiang, R. Wang, H.Y. Cao, Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds, Chemomet. Intell. Lab. Syst. 92 (2008) 169–178.
[28] Y. Pan, J.C. Jiang, R. Wang, H.Y. Cao, J.B. Zhao, Quantitative structure–property relationship studies for predicting flash points of organic compounds using support vector machines, QSAR Comb. Sci. 27 (2008) 1013–1019.
[29] J. Wang, H.Y. Du, H.X. Liu, X.J. Yao, Z.D. Hu, B.T. Fan, Prediction of surface tension for common compounds based on novel methods using heuristic method and support vector machine, Talanta 73 (2007) 147–156.
[30] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.
[31] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (1996) 1027–1043.
[32] R. Todeschini, V. Consonni, Handbook of molecular descriptors, in: R. Mannhold, H. Kubinyi, H. Timmerman (Eds.), Methods and Principles in Medicinal Chemistry, Wiley-VCH, Weinheim, 2000.
[33] R. Todeschini, V. Consonni, M. Pavan, DRAGON. Software for the calculation of molecular descriptors, web version 2.1, 2002. http://www.disat.unimib.it/chm/.
[34] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, 1975.
[35] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemomet. 6 (1992) 267–281.
[36] R. Leardi, Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection, J. Chemomet. 8 (1994) 65–79.
[37] M.H. Fatemi, M. Jalali-Heravi, E. Konuze, Prediction of bioconcentration factor using genetic algorithm and artificial neural network, Anal. Chim. Acta 486 (2003) 101–108.
[38] M.H. Fatemi, Prediction of ozone tropospheric degradation rate constant of organic compounds by using artificial neural networks, Anal. Chim. Acta 556 (2006) 355–363.
[39] P. Ghosh, B. Chawla, P.V. Joshi, S.B. Jaffe, Prediction of chromatographic retention times for aromatic hydrocarbons, Energy Fuels 20 (2006) 609–619.
[40] S.S. Godavarthy, R.L. Robinson, K.A.M. Gasem, An improved structure–property model for predicting melting-point temperatures, Ind. Eng. Chem. Res. 45 (2006) 5117–5126.
[41] R. Leardi, A. Lupiáñez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemomet. Intell. Lab. Syst. 41 (1998) 195–207.
[42] C.C. Chang, C.J. Lin, LIBSVM: A library for support vector machines, 2001. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
[43] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.
[44] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang, Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships, the report and recommendations of ECVAM Workshop 52, Altern. Lab. Anim. 33 (2005) 155–173.