Korean J. Chem. Eng., **29**(2), 145-153 (2012)
DOI: 10.1007/s11814-011-0164-8

Flash point prediction of organic compounds using a group contribution and support vector machine

Chang Jun Lee*, Jae Wook Ko, and Gibaek Lee***,†**

*Department of Chemical and Biological Engineering, Seoul National University, Shilim-dong, Gwanak-gu, Seoul 151-742, Korea
**Department of Chemical Engineering, Kwangwoon University, Wolgye-dong, Nowon-gu, Seoul 139-701, Korea
***Department of Chemical and Biological Engineering, Chungju National University, Chungju, Chungbuk 380-702, Korea
(*Received 4 April 2011 • accepted 25 June 2011*)

Abstract–The flash point is one of the most important properties of flammable liquids. This study proposes a support vector regression (SVR) model to predict the flash points of 792 organic compounds from the DIPPR 801 database. The input variables of the model consist of 65 different functional groups, logarithm of molecular weight and their boiling points in this study. Cross-validation and particle swarm optimization were adopted to find three optimal parameters for the SVR model. Since the prediction largely relies on the selection of training data, 100 training data sets were randomly generated and tested. Moreover, all of the organic compounds used in this model were divided into three major classes, which are non-ring, aliphatic ring, and aromatic ring, and a prediction model was built accordingly for each class. The prediction results from the three-class model were much improved than those obtained from the previous works, with the average absolute error being 5.11-7.15 K for the whole data set. The errors in calculation were comparable with the ones from experimental measurements. Therefore, the proposed model can be implemented to determine the initial flash point for any new organic compounds.

Key words: Flash Point, Property Estimation, Group Contribution Methods, Support Vector Regression, Particle Swarm Optimization

INTRODUCTION

The flash point (FP) is one of the most important properties for evaluating the hazards of liquids, and it has received increased concern for process safety in recent years [1]. The FP of a liquid is the lowest temperature at which the mixture of air and vapor near its surface can be ignited (by spark or flame). There are two basic experimental methods to measure the FP: open cup and closed cup. The errors in the measurements are approximately 5-8 K for both methods [1].

Experimental methods are the most accurate way to collect FP data, which are of vital importance in the design of chemical processes. However, FP measurements are very expensive and time consuming. In addition, to obtain the FP for toxic, explosive, or radioactive compounds is extremely difficult. Among the millions of chemical compounds, FPs of only a few thousands are reported. Even if the FP is listed in the literature, the references are often not accessible. Therefore, a reliable and accurate FP prediction method is absolutely essential to chemical industries.

In the literature, there are many methods for estimating the FP of organic compounds; for instance, Vidal et al. [2] reviewed the prediction methods for FP and flammability limits, of which the quantitative structure-property relationships (QSPR) methods have been used in large amount of research work [3-9]. As the name suggests, QSPR methods investigate the quantitative relation between the descriptors of the chemical structures and their properties. Consequently, FPs of unknown chemical compounds can be predicted [3]. The first QSPR work utilized two descriptors to calculate the FPs of 400 compounds, and the average absolute error (AAE) for all 400 compounds was 10.3 K [4]. After that, Tetteh et al. [5] applied the radial basis function neural network to construct the FP prediction model for 400 compounds and the inputs of the model were 25 functional groups and their molecular connectivity index. In addition Katritzky et al. [1] proposed a three-descriptor linear equation for 271 compounds. Later they upgraded their model for 758 compounds [6]. In Katritzky's study, multiple linear regression and neural network were employed, and the AAEs were 13.9 and 12.6 K for 158 test compounds, respectively. Furthermore Gharagheizi and Alamdari [7] employed a generic algorithm based on the multiple linear regression method for 1030 compounds and selected four molecular descriptors out of the 1664 molecular descriptors. Moreover Patel et al. [8] used 16 molecular descriptors for 236 solvents with the AAE being 20.44 K. Additionally our previous work employed partial least square (PLS) and support vector regression (SVR) with the inputs of 65 functional groups and the logarithm of molecular weight, and realized the predictability of the SVR is far better than that of the PLS.

Table 1 summarizes these previous studies, and the statistical parameters used, that are AAE, R^2 (squared correlation coefficient), and RMS (root mean square), are defined as follows.

$$AAE = \frac{\sum_{i=1}^{n} |y_p - y_a|}{n} \quad (1)$$

†To whom correspondence should be addressed.
E-mail: glee@cjnu.ac.kr

Table 1. Previous works and their results

Works	Inputs	Modeling method	AAE (train/test)	R^2 (train/test)	RMS (train/test)	No. of whole data	Ratio of training data
Tetteh et al. [5]	Molecular connectivity index and 25 functional groups	Neural network	7.1/11.2	0.96/0.92	10.1/14.0	400	33%
Katritzky et al. [1]	3 Molecular descriptors	Multi-parameter regression	-	0.95	12.2	271	100%
Katritzky et al. [6]	4 Molecular descriptors	Neural network	-/12.6	0.88/0.98	-	758	79%
Gharagheizi et al. [7]	4 Molecular descriptors	GA-MLR	-/10.2	0.97/0.97	12.0/12.7	1030	80%
Pan et al. [3]	57 Functional groups	SVM	6.12/9.99	0.98/0.95	9.95/15.81	1282	80%
Patel et al. [8]	16 Molecular descriptors	Neural network	20.44 (whole)	0.90/0.66	-	236 (solvents)	80%
Lee et al. [9]	65 Functional groups and molecular weight	SVM	5.6-9.43 /10.85-18.07	0.94-0.98 /0.78-0.94	7.64-15.28 /15.56-30.14	893	80%

$$R^2=\frac{\left(n\sum_{i=1}^{n}y_p y_e-\sum_{i=1}^{n}y_p\cdot\sum_{i=1}^{n}y_e\right)^2}{\left[n\sum_{i=1}^{n}y_p^2-\left(\sum_{i=1}^{n}y_p\right)^2\right]\left[n\sum_{i=1}^{n}y_e^2-\left(\sum_{i=1}^{n}y_e\right)^2\right]} \quad (2)$$

$$RMS=\sqrt{\frac{\sum_{i=1}^{n}(y_p-y_e)^2}{n}} \quad (3)$$

In the above equations, n is the number of compounds, and y_p and y_e are the predicted and experimental values, respectively. Table 1 shows dramatic differences in the parameters for these studies. However, these differences could rise from many reasons, for example, the compounds used in each work are not same. In addition, these works divided the whole data into training and test data, and the selection of training data is one of important factors that can affect the prediction result. Therefore, it is not possible to conclude whether a work is superior to other works simply based on the statistics. For fair comparisons, our previous work randomly generated 100 training and test data sets, and compared their average, minimum, and maximum prediction parameter values.

The selections of molecular descriptors and modeling techniques are the key issues in the QSPR calculations. In this work, SVR model is adopted to predict the FP of organic compounds, where 65 functional groups and two molecular properties were implemented as the inputs of the model. Also, particle swarm optimization technique (PSO) was applied to find the optimal parameters of the SVR model. In the next section, the proposed methods will be introduced followed by comparison and contrast with literature results.

THE PROPOSED METHODS

1. Group Contribution Method

Molecular descriptors can be classified into molecular properties, constitutional descriptors, topological descriptors, geometrical descriptors, functional group counts and so on [10]. The property estimation method based on functional groups is well known as the group contribution method, which has been used to predict various physical properties of organic compounds [11-15]. In this method, the property of a compound is estimated from the functional groups which are made from atoms and bonds. As the data containing a few hundred chemical compounds is needed in order to train the prediction model, it has an advantage to greatly reduce the number of required data. However, it could result in poor prediction when the molecular structure is oversimplified or there is insufficient information in the molecular property database [11,15]. As a consequence, to improve the accuracy of FP prediction, it is necessary to obtain reliable and sufficient FP data and to employ enough functional groups.

There are many sources of experimental FP data such as the literature, handbook, and database. This study used the FP data from the DIPPR 801 database (2009 v. 2), which is one of the most reliable physical property databases [16]. Among the total 1973 compounds, there are 1765 organic compounds but the ones with experimental FP data are only 893.

As the input of the prediction model, 65 functional groups and the logarithm of molecular weight were implemented [9]. These functional groups were chosen by analyzing the chemical structures of 893 organic compounds, 55 functional groups proposed by Lee et al. [15], and 57 functional groups proposed by Pan et al. [3]. These 65 groups are grouped into four types consisting of 18 end groups, 24 middle groups, 13 aliphatic rings, and 10 aromatic rings. After our previous study, we focused on experimental boiling points (BP) as an input of the prediction model. Several studies have proposed the nonlinear equation to estimate FP from BP [17-22]. Patil [17] proposed the following quadratic equations for the estimation of the FP.

$$T_f=4.656+0.844T_b-0.234\times10^{-3}T_b^2 \quad (4)$$

Where T_f and T_b indicate the FP and BP in K, respectively. Satyanarayana and Kakati [18] found that the equation doest not predict the FP correctly. Hshieh [21] suggested another quadratic equation.

$$T_f=-54.5377+0.5883T_b+0.00022T_b^2 \quad (5)$$

The standard error of Eq. (6) was 11.66 K for 494 compounds. Following Satyanarayana and Kakati [18], Metcalfe and Metcalfe [19] proposed a regression equation to calculate the FP from the BP and

liquid density.

$$T_f = -84.794 + 0.6208T_b + 37.8127\rho \quad (6)$$

Where ρ is the liquid density in g/cm^3. They reported a standard deviation of 10.2 K for 250 compounds. Katritzky et al. [1] used three descriptors including experimental boiling point for the FP prediction of 271 compounds. Satyanarayana and Rao [20] suggested the following nonlinear equation for 13 groups of compounds.

$$T_f = a + b(c/T_b)e^{-c/}/(1-e^{-c/})^2 \quad (7)$$

Where a, b, c are constants, which are different for each group. For 1221 organic compounds, the equation showed the AAE of less than 1%. Most recently among these studies, Catoire and Naudet [22] suggested the following nonlinear equation expressed as a function of T_b in K, the standard enthalpy of vaporization at 298.15 K (ΔH^0_{vap}) in kJ/mol, and the total number of carbon atoms in the molecule (n).

$$T_f = 1.477 \times T_b^{0.79686} \times \Delta H^{0}_{vap}{}^{0.16845} \times n^{-0.05948} \quad (8)$$

They obtained the AAE of about 3 K, the standard deviation of about 2 K, and a maximum absolute deviation of 10 K.

The number of compounds with known experimental FP and BP data is 792, and the squared correlation coefficient of FP on BP is 0.91 (Fig. 1). Due to the strong correlation between BP and FP, BP was added into the inputs of the prediction model in this study as shown in Table 2.

792 organic compounds are classified into 63 classes as defined in the DIPPR database (Table 3). If the prediction model for each class is built separately, the variations in the compound structures for each class will be less than the ones in the whole data, and subsequently the predictability will be enhanced. Patel et al. [8] classified 236 solvents into five major classes, which are monohydric alcohols, polyhydric alcohols, amines, ethers, and aliphatic and aromatic hydrocarbons, and for each class a FP prediction model was built.

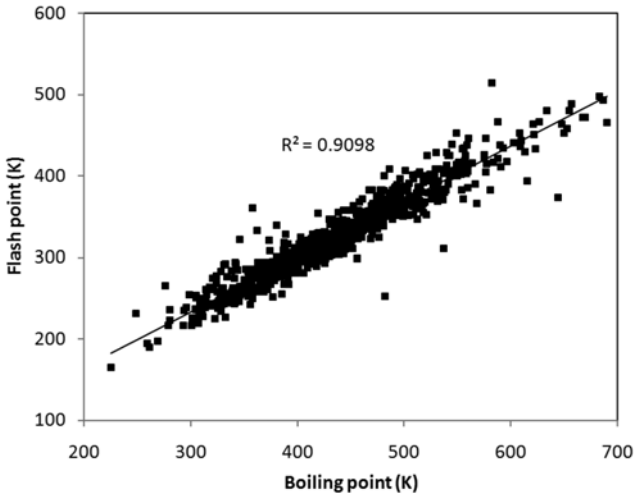

Fig. 1. Plot of experimental boiling points versus experimental flash points.

As the input of our SVR model includes 13 aliphatic rings and 10 aromatic rings, this study classified organic compounds into ring and non-ring classes, where the ring class is divided further into aromatic and aliphatic rings. Overall, there are 475 non-ring compounds with functional group number reduced to 42, since two groups in the ending group as well as all in aromatic and aliphatic ring groups are not used in non-ring compounds. The number of compounds and functional groups of each class is shown in Table 4.

2. Support Vector Regression (SVR)

In general, regression methods are used to represent the relationships between the functional groups and their physical properties. Unfortunately, these relationships are highly nonlinear. Moreover, the general regression methods are less suitable when a large dimen-

Table 2. The inputs of the model

No.	Input	No.	Input	No.	Input	No.	Input
1(E1)	-CH3	18(E18)	-H	35(M17)	-Al-	52(R10)	O
2(E2)	=CH2	19(M1)	>C<	36(M18)	-B-	53(R11)	=C
3(E3)	CH	20(M2)	>C=	37(M19)	>C<(-X)*	54(R12)	S
4(E4)	N	21(M3)	=C=	38(M20)	>C=(-X)*	55(R13)	>Si<
5(E5)	-NH2	22(M4)	-C	39(M21)	-CH2-(-X)*	56(A1)	=CH-
6(E6)	-NO2	23(M5)	-CH2-	40(M22)	>CH-(-X)*	57(A2)	=C<
7(E7)	-SH	24(M6)	>CH-	41(M23)	-CH=(-X)*	58(A3)	=C<(-X)*
8(E8)	-Br	25(M7)	-CH=	42(M24)	>Si<	59(A4)	>N-
9(E9)	-F	26(M8)	>N-	43(R1)	-CH2-	60(A5)	NH
10(E10)	-Cl	27(M9)	-N=	44(R2)	=CH-	61(A6)	O
11(E11)	-I	28(M10)	-NH-	45(R3)	>CH-	62(A7)	S
12(E12)	-COH	29(M11)	-O-	46(R4)	>C<	63(A8)	o-B, m-B, p-B
13(E13)	-COOH	30(M12)	-S-	47(R5)	=C<	64(A9)	3-branched benzene**
14(E14)	=O	31(M13)	-CO-	48(R6)	-N<	65(A10)	4-branched benzene***
15(E15)	-OH(alcohol)	32(M14)	-CO2-	49(R7)	NH	66	Logarithm of molecular weight
16(E16)	-OH(phenol)	33(M15)	-SO2-	50(R8)	=N-	67	Boiling point
17(E17)	=S	34(M16)	-SO-	51(R9)	CO		

*-X: attached to halogen atoms, **3-Branched benzene: (1,2,3), (1,2,4), or (1,3,5), ***4-Branched benzene: (1,2,3,4), (1,2,3,5), or (1,2,4,5)

Table 3. The number of compounds in 63 classes

Class	No. of Comp.	Class	No. of Comp.	Class	No. of Comp.
Acetates	18	Amines, n-Aliphatic Primary	9	Esters, Unsaturated Aliphatic	14
Acids, Aromatic	22	Amines, Other Aliphatic	18	Ethers/Diethers, Other	14
Acids, Aromatic Carboxylic	1	Amines/Imines, Other	27	Formates	6
Acids, n-aliphatic	6	Anhydrides	7	Halides, Polyfunctional C,H,O	27
Acids, Other-Aliphatic	9	C,H,Multihalogen Compounds	12	Isocyanates/Diisocyanates	4
Acids, Polyfunctional	1	C,H,NO2 Compounds	9	Ketones	35
Alcohols, Cycloaliphatic	9	C,H,O, Other Polyfunctional	31	Mercaptans	18
Alcohols, n-	2	C,H,O,S, Polyfunctional	20	Methylalkenes	14
Alcohols, Other Aliphatic	14	Chlorides, Aromatic	15	Monoaromatics, Other	9
Aldehydes	18	Chlorides, C1/C2 Aliphatic	6	Naphthalenes	6
Alkanes, n-	14	Chlorides, C3 & Higher Aliphatic	21	Nitriles, Polyfunctional	17
Alkanes, Other	4	Cycloalkanes	3	Organics, Other Polyfunctional	2
Alkenes, 1-	5	Cycloalkanes, Multiring	3	Peroxides	2
Alkenes, 2,3,4-	7	Cycloalkenes	5	Polyols	22
Alkenes, Ethyl & Higher	8	Dialkenes	16	Propionates/Butyrates	11
Alkylbenzenes, n-	10	Dimethylalkanes	8	Rings, Other Condensed	3
Alkylbenzenes, Other	24	Diphenyl/Polyarmatics	8	Rings, Other Hydrocarbon	6
Alkylcyclopentanes	7	Epoxides	13	Salts, Organic	12
Alynes	8	Esters, Aromatic	30	Silanes/Siloxanes	29
Amides/Amines, Polyfunctional	15	Esters, Other Saturated Aliphatic	14	Sulfides/Thiophenes	14
Amines, Aromatic	26	Esters, Polyfunctional	17	Terpenes	7

Table 4. The number of compounds and functional groups for each class

Class	No. of compounds	Functional groups (no. of functional groups)
Non-ring	475	E1-E8, E9-E15, E17, E18, M1-M24 (42)
Ring	317	E1-E10, E12-E16, M1-M11, M13, M14, M19, M21, M24, R1-R13, A1-A10 (56)
Aromatic ring	207	E1-E10, E12-E16, M1-M11, M13, M14, M19, M21, M24, R1, R3, R7, R9, R10, R12, A1-A10 (49)
Aliphatic ring	110	E1, E2, E5, E7, E10, E12-E15, M1-M3, M5-M7, M9-M11, M14, M21, R1-R13 (35)

sion of input data is involved, such as in our case. In addition, the functional groups included in most of the chemical compounds are much fewer than 65. Thus, most of the input data values would be zero, and this means the input data set is very sparse. To tackle these, SVR is adopted to construct empirical models since it is built from statistical learning theory and is also based on the classification paradigm, namely, support vector machine [23]. From the literature SVR is shown to be an effective method for handling linear or nonlinear sparse data sets [23]. In SVR, the inputs are first linearity or nonlinearity mapped into a high-dimensional feature space wherein the linearity is correlated with the output [24]. Such a linear regression in a high-dimensional feature space reduces the algorithm complexity enabling high predictive capabilities of both training and test set of data [24]. In this regard, SVR is one of the most suitable tools for constructing empirical models. A more detailed description of SVR is available in the literature [23].

The goal of SVR is to find the optimal hyperplane, from which the distance to all of the data points is minimum [3]. The problem of linear SVR in the given n training data (x_i, y_i) is to find the optimal hyperplane, $f(x)=\omega \cdot x+b$ that can estimate y. With the given ε-insensitive loss function, the distance from the hyperplane to any data point is less than ε. The value of ε can affect the number of support vectors used to construct the regression function. After additional slack variables ξ and ξ^* are introduced, the optimization problem can be written as

$$\begin{aligned} &\min J(\omega, \xi_i, \xi_i^*) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{subject to } y_i - \omega x - b \le \varepsilon + \xi_i \\ &\qquad \omega x + b - y_i \le \varepsilon + \xi_i^* \\ &\qquad \xi_i, \xi_i^* \ge 0 \end{aligned} \tag{9}$$

where C is the penalty parameter that determines the trade-off between the training error and the model complexity.

By solving Eq. (9), the optimal linear regression function is obtained as follows.

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \tag{10}$$

where $0 \le \alpha_i, \alpha_i^* \le C$.

The nonlinear SVR is to map the data onto a high-dimensional feature space using kernel function, $K(x, x_i)$ so that a linear model can be constructed in this feature space. Therefore, it allows the con-

Table 5. The prediction performances and comparison with our previous model

Statistical parameter	The proposed model			Our previous model			Ratio (proposed/previous)		
	Average	Min.	Max.	Average	Min.	Max.	Average	Min.	Max.
AAE (train)	6.30	5.12	7.15	7.86	4.24	12.19	0.80	1.21	0.59
AAE (test)	8.02	5.84	10.16	12.63	10.32	15.15	0.63	0.57	0.67
AAE (whole)	6.65	5.74	7.34	8.82	5.94	12.46	0.75	0.97	0.59
Max Error (train)	77.22	50.14	86.8	68.49	31.52	110.88	1.13	1.59	0.78
Max Error (test)	68.76	33.55	106.36	94.98	54.45	144.46	0.72	0.62	0.74
Max Error (whole)	81.96	71.11	106.36	97.85	54.45	144.46	0.84	1.31	0.74
R^2 (train)	0.964	0.954	0.979	0.953	0.892	0.985	1.01	1.07	0.99
R^2 (test)	0.948	0.897	0.976	0.884	0.811	0.932	1.07	1.11	1.05
R^2 (whole)	0.961	0.956	0.969	0.939	0.889	0.969	1.02	1.08	1.00
RMS (train)	10.31	8.12	11.60	11.43	6.63	18.38	0.90	1.22	0.63
RMS (test)	12.49	8.51	16.80	18.73	14.51	23.60	0.67	0.59	0.71
RMS (whole)	10.82	9.59	11.47	13.34	9.67	18.36	0.81	0.99	0.62

version of nonlinear problems into linear regression. Consequently, Eq. (10) can be rewritten as follows.

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x, x_i) + b \quad (11)$$

Among several kernel functions, this study used radial basis function (RBF) of $K(x, x_i)=\exp(-\gamma|x-x_i|^2)$.

User can select penalty parameter C, ε of ε-insensitive loss function, and the width, γ of RBF. The optimal values of three parameters are determined by particle swarm optimization, and the objective function of the optimization is the mean squared error of 10-fold cross validation.

The SVM model was calculated by the Matlab library of LibSVM v.2.91 [25], which is one of the fastest SVM libraries.

3. Particle Swarm Optimization (PSO)

PSO is one of the heuristic optimization methods such as generic algorithm (GA) and simulated annealing (SA) [26]. The word "heuristic" means that the methods mimic the successful optimization strategies found in nature. Although it is not guaranteed to find the global optimum solution, a good approximation can often be achieved. They do not need derivatives of the objective functions and are also not sensitive to the initial parameter guesses. These advantages are very useful in determining SVM parameters. However extensive function evaluation is a well known disadvantage of this method. As the evaluation time for the cross validation of SVR is not small and PSO required less computational effort than GA and SA, this study employed PSO to determine three parameters of SVR.

PSO was originally developed based on the social behavior of collection of animals such as birds [26]. It starts with randomly generated swarms, called particles, and remembers the best solution found. The particles move around the solution space with adjusted velocities and have a tendency to swim towards the global optimal solution over the optimal procedure. The detailed algorithm of PSO is available in the paper by M. Schwaab et al. [26].

RESULTS AND DISCUSSION

First, one model was built to cover the FPs of the total 792 organic compounds. The SVM model was trained with 633 compounds (80% of 792) and the rest of the compounds were used for the test set. This study randomly generated 100 data sets of training data and all sets were tested, since prediction results depended on training data set. The parameters of SVR in all models were optimized by PSO.

Table 5 presents the results obtained from the proposed model. In this table, there are average, minimum and maximum errors obtained from 100 training data sets. It can be found that most errors of the proposed models have been improved when compared with those from our previous study [9]. However, the comparison may not be entirely fair since the number of compounds (792) is different from our previous study (893). To obtain a fair comparison, our previous model was re-tested using the same training sets, and the results are listed in Table 5. 29 statistical parameters in the 36 parameters were better than our previous model.

For the data set with the minimum AAE for test data, the predicted and experimental values are plotted in Fig. 2(a). When the plot is compared with the one obtained from our previous model shown in Fig. 2(b), it is certain that the proposed model provided higher accuracy.

For the data set with the minimum AAE of 6.65 for the whole data, the percent error of all the 792 compounds was calculated. The maximum relative percent error was 21%, and the average percent error was 1.77%. The detailed relative errors are plotted in Fig. 3(a). As shown in the figure, the error of 426 compounds is less than 1%, and the result is better than our previous study (Fig. 3(b)). Nevertheless, the error of 16 compounds, which is about 2% of the 792 compounds, is more than 10%.

To improve the accuracy, all compounds are classified into different major classes. Firstly, 792 organic compounds are divided into ring and non-ring classes, and 100 training sets of each class were randomly generated and prediction models were built as to optimize the parameters for SVR. The prediction results are listed in Table 6. For 575 compounds in the non-ring class, the average AAE of whole data is 5.77, which is 87% of one obtained by the model without class (one model). Additionally, the results of non-ring class are greatly improved compared to those of one model and ring class, while the results of ring class are not as comparable as the one model.

Ring compounds are classified into aromatic ring and aliphatic

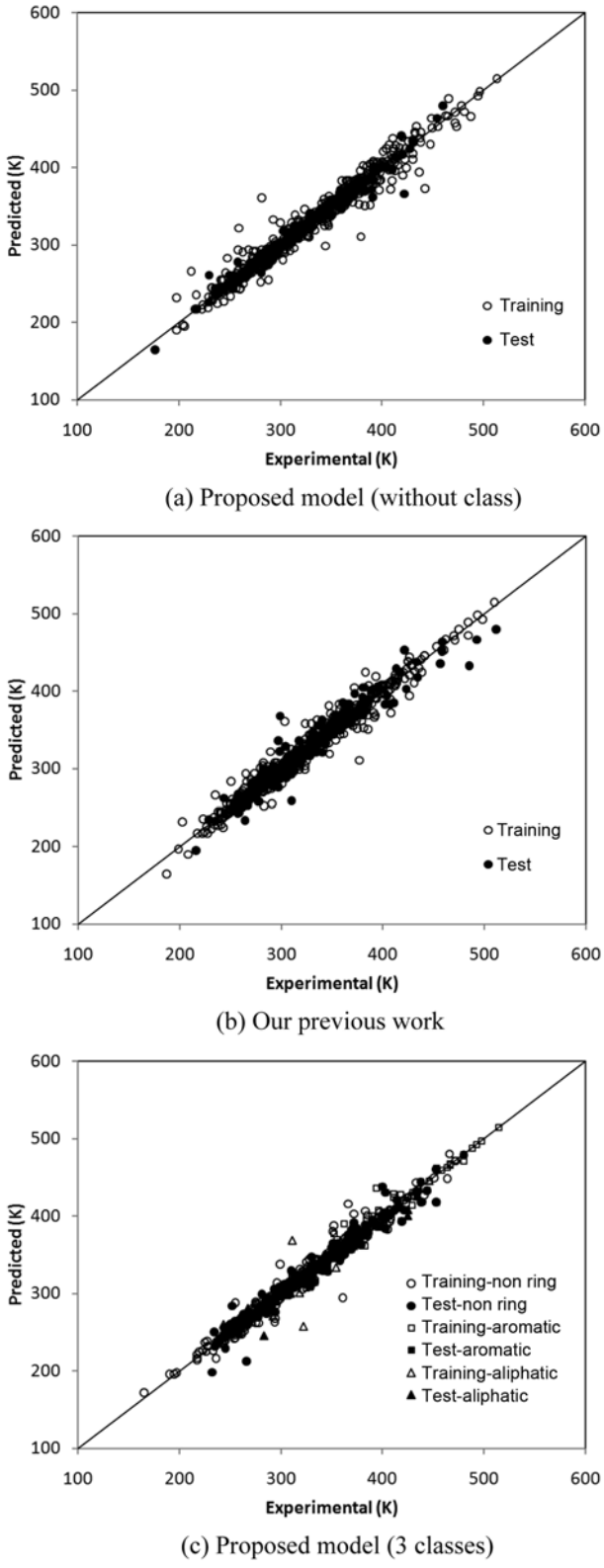

Fig. 2. Comparison between the predicted and experimental flash points.

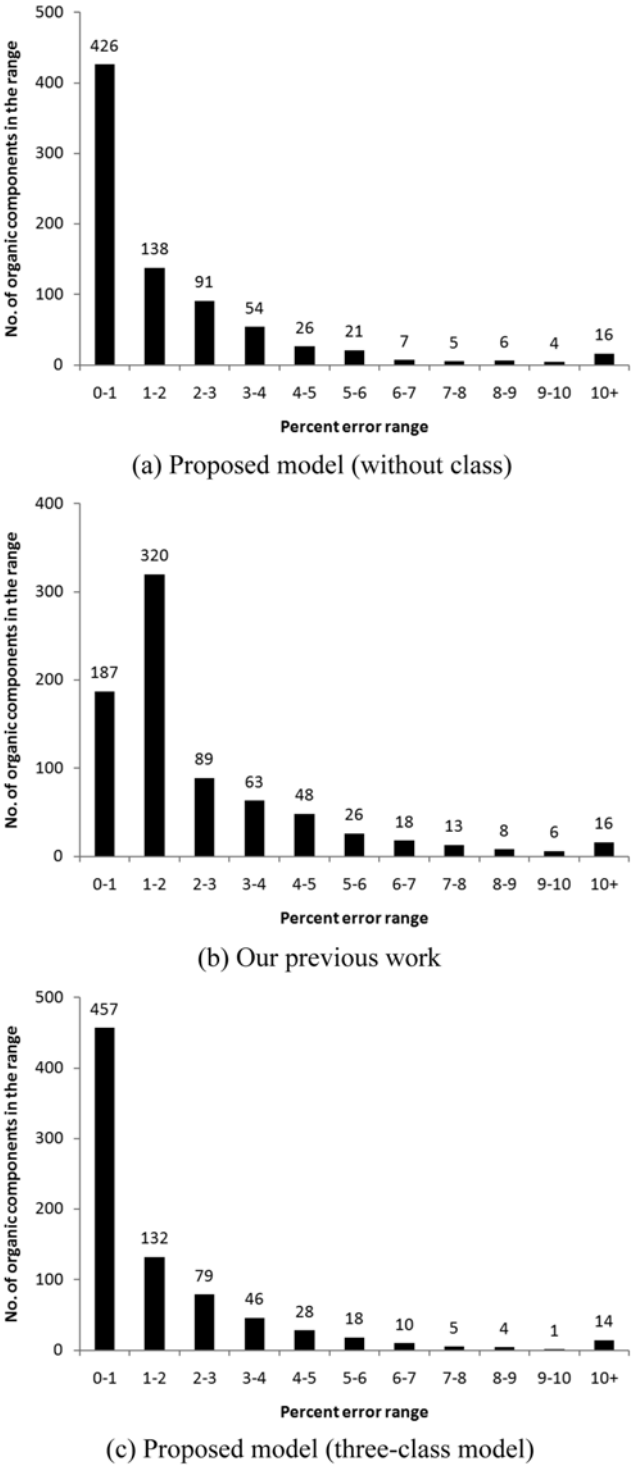

Fig. 3. The number of organic compounds in each range of the percent errors.

ring classes, and their prediction results are shown in Table 6. It further proves the classification system is a key point to enhance the prediction performance.

To compare the prediction results of one model with three-class model, data sets of non-ring (475 compounds), aromatic ring (207), and aliphatic ring (110) should be merged. 100 data sets of each class were combined to be 1,000,000 (=100^3) data sets, and the results are summarized in Table 7. On comparison, 28 statistical parameters out of 36 are better for the three-class model than one model, while 6 parameters are worse.

Table 6. Flash point prediction performance of each class

Statistical parameter	Non-Ring			Ring			Aromatic Ring			Aliphatic Ring		
	Average	Min.	Max.	Average	Min.	Max.	Average	Min.	Max.	Average	Min.	Max.
AAE (train)	5.33	4.32	6.96	6.05	4.63	7.60	6.01	3.67	7.57	5.22	3.01	7.11
AAE (test)	7.50	5.31	10.02	9.53	6.83	12.38	9.77	6.73	13.03	9.69	5.17	19.68
AAE (whole)	5.77	5.17	6.85	6.75	5.69	8.18	6.79	5.02	8.01	6.11	5.03	7.65
Max error (train)	66.60	38.58	80.86	77.14	55.09	96.10	38.52	25.17	47.23	61.76	22.63	68.34
Max error (test)	51.31	19.89	87.89	60.67	27.18	100.39	41.83	19.65	83.72	45.76	15.80	75.75
Max error (whole)	73.68	54.61	87.89	81.04	64.78	100.39	45.94	37.77	83.72	66.28	58.69	75.75
R^2 (train)	0.970	0.955	0.978	0.958	0.938	0.979	0.967	0.953	0.985	0.944	0.916	0.988
R^2 (test)	0.952	0.897	0.976	0.925	0.849	0.974	0.925	0.846	0.971	0.898	0.717	0.984
R^2 (whole)	0.967	0.956	0.973	0.951	0.934	0.964	0.958	0.946	0.972	0.933	0.907	0.945
RMS (train)	8.97	7.68	11.13	10.61	7.44	12.63	8.68	5.77	10.40	10.55	4.95	12.70
RMS (test)	11.33	7.15	15.09	14.36	9.78	20.18	13.37	8.64	18.76	14.85	6.90	27.29
RMS (whole)	9.53	8.57	10.93	11.54	9.76	13.36	9.87	8.10	11.27	11.79	10.69	14.01

Table 7. Comparison between one model and three-class model

Statistical parameter	One model			Three-class model		
	Average	Min.	Max.	Average	Min.	Max.
AAE (train)	6.30	5.12	7.15	5.49	3.97	7.14
AAE (test)	8.02	5.84	10.16	8.43	5.67	12.15
AAE (whole)	6.65	5.74	7.34	6.08	5.11	7.27
Max error (train)	77.22	50.14	86.8	70.29	38.58	80.86
Max error (test)	68.76	33.55	106.36	62.16	19.89	87.89
Max error (whole)	81.96	71.11	106.36	74.65	58.69	87.89
R^2 (train)	0.964	0.954	0.979	0.972	0.958	0.984
R^2 (test)	0.948	0.897	0.976	0.948	0.877	0.980
R^2 (whole)	0.961	0.956	0.969	0.967	0.956	0.974
RMS (train)	10.31	8.12	11.60	9.17	6.89	11.18
RMS (test)	12.49	8.51	16.80	12.59	7.54	18.23
RMS (whole)	10.82	9.59	11.47	9.97	8.78	11.49

When three-class model was used, the percent errors of 792 compounds for the data set with minimum AAE for whole data are plotted in Fig. 3(c). The average percent error was 1.59%, and the percent error of 457 compounds is less than 1%. The result obtained is better than one model shown in Fig. 3(a). Fig. 2(c) compares the predicted and experimental values for the data set used in Fig. 3(c), and it shows that the proposed model predicted more accurately than our previous model.

Although different data sets were used in other studies of Table 1, the obtained results illustrate that the proposed model provides definitely the best measure. Unlike the previous studies shown in Table 1, it is easy to compare the proposed model with the previous studies of using empirical linear or nonlinear equations. Because Catoire and Naudet [22] reported the minimum AAE among these studies [17-22], this study compared the result of the proposed model with their three-parameter nonlinear equation. The required data including the standard enthalpy of vaporization were obtained from DIPPR 801, and Eq. (8) was used to estimate the FP of 792 compounds. The calculated AAE, maximum absolute error, R^2, and RMS were 9.57, 103.68, 0.931, and 14.81. The maximum absolute error, 103.68 is for ethyl aluminum sesquichloride (experimental and estimated FPs are 253.15 K and 356.83 K). The results were worse than the worst result of the proposed model, and the AAE by Eq. (8) was almost two-times of our best AAE, 5.11 (Table 7). For the data set with minimum AAE for whole data, the FPs of the 15 compounds having the biggest absolute errors calculated by the proposed model were calculated by Eq. (8) (Table 8). The AAE of 15 compounds by the proposed model, 43.2 K, is a bit smaller than the one by Eq. (8), 43.5 K. In Table 9, the FPs of the 15 compounds having the biggest absolute errors calculated by Eq. (8), are compared with the ones by the proposed model. The AAE of 15 compounds by Eq. (8), 62.8 K, is very bigger than the one by the proposed model, 29.3 K. The comparison illustrated the proposed model improved the prediction accuracy. The higher predictive accuracy could be attributed to more proper inputs of the model and the optimization of the SVR parameters.

CONCLUSION

This study has built the SVR model to estimate the FP by using

Table 8. Comparison between the proposed model and Eq. (8) for 15 compounds having the biggest absolute errors calculated by the proposed model

Compound	Experiment value	The proposed model				Eq. (8)	
		Estimated value	Absolute error	Training or test	Class	Estimated value	Absolute error
Chloroacetaldehyde	361	294.8	66.2	Training	Non-ring	283.4	77.6
Methylcyclopentadiene	322.039	257.7	64.3	Training	Aliphatic	251.9	70.1
2-Cyclohexyl Cyclohexanone	311	368.7	57.7	Training	Aliphatic	390.3	79.3
Trimethylamine	266	212.8	53.2	Test	Non-ring	205.0	61.0
Adiponitrile	366.15	415.8	49.7	Training	Non-ring	425.9	59.8
Anthracene	394	436.1	42.1	Training	Aromatic	430.1	36.1
1,5-Dichloropentane	299	338.3	39.3	Training	Non-ring	342.2	43.2
1,3-Benzenediol	400	438.1	38.1	Test	Aromatic	410.1	10.1
Thiacyclopropane	283.15	245.9	37.2	Test	Aliphatic	254.8	28.3
1,3-Propylene Glycol	352.15	387.9	35.7	Training	Non-ring	395.1	42.9
Isophthaloyl Chloride	453.15	418.1	35.1	Test	Aromatic	403.1	50.0
Dimethyl Ether	232	198.5	33.5	Test	Non-ring	187.7	44.3
2-Methylthiophene	255.15	288.4	33.3	Training	Non-ring	286.0	5.9
Acetal	252.15	284.0	31.9	Test	Non-ring	315.7	2.6
1,6-Hexanediol	372	402.9	30.9	Training	Non-Ring	413.0	41.0

Table 9. Comparison between Eq. (8) and the proposed model for 15 compounds having the biggest absolute errors calculated by Eq. (8)

Compound	Experiment value	Equation 8		The proposed model			
		Estimated value	Absolute error	Estimated value	Absolute error	Training or test	Class
Ethyl Aluminum Sesquichloride	253.15	356.8	103.7	254.1	1.0	Training	Non-Ring
Phenothiazine	373.15	470.9	97.7	374.1	0.9	Training	Aromatic
2-Cyclohexyl Cyclohexanone	311	390.3	79.3	368.7	57.7	Training	Aliphatic
Chloroacetaldehyde	361	283.4	77.6	294.8	66.2	Training	Non-ring
Methylcyclopentadiene	322.039	251.9	70.1	257.7	64.3	Training	Aliphatic
1,3,5-Trichlorobenzene	400	337.9	62.1	399.1	0.9	Training	Aromatic
Trimethylamine	266	205.0	61.0	212.8	53.2	Test	Non-ring
Adiponitrile	366.15	425.9	59.8	415.8	49.7	Training	Non-ring
Hexachlorobenzene	515	459.8	55.2	514.1	0.9	Training	Aromatic
Isophthaloyl Chloride	453.15	403.1	50.0	418.1	35.1	Test	Aromatic
Dichloroacetaldehyde	333.15	286.2	47.0	318.2	15.0	Training	Non-Ring
Formic Acid	321.15	274.4	46.8	320.2	1.0	Training	Non-Ring
3,4-Dichloroaniline	439	394.4	44.6	418.8	20.2	Test	Aromatic
Dimethyl Ether	232	187.7	44.3	198.5	33.5	Test	Non-ring
1,5-Dichloropentane	299	342.2	43.2	338.3	39.3	Training	Non-ring

the experimental FPs of 792 organic compounds from DIPPR 801. As the input of the prediction model, we used 65 functional groups, the logarithm of molecular weight, and their BPs. In addition, all of the 792 compounds are classified into three major classes that are non-ring, aliphatic ring, and aromatic ring, and hence three prediction models were built individually for these classes.

The average AAE obtained by the proposed model was 6.08 K, which was much improved compared to the previous studies. Furthermore, the result was comparable to the accuracy of experimental FP determination (5-8 K). Thus, the proposed model is expected to produce a reliable FP data estimation for any new organic compound.

ACKNOWLEDGEMENT

Thr research was supported by a grant from the Academic Research Program of Chungju National University in 2009.

REFERENCES

1. A. R. Katritzky, R. Petrukhin, R. Jain and M. Karelson, *J. Chem. Inf. Comput. Sci.*, **41**, 1521 (2001).

2. M. Vidal, W. J. Rogers, J. C. Holste and M. S. Mannan, *Process Saf. Prog.*, **23**, 47 (2004).
3. Y. Pan, J. Jiang, R. Wang, H. Cao and J. Zhao, *QSAR Comb. Sci.*, **27**, 1013 (2008).
4. T. Suzuki, K. Ohtaguchi and K. Koide, *J. Chem. Eng. Japan*, **24**, 258 (1991).
5. J. Tetteh, T. Suzuki, E. Metcalfe and S. Howells, *J. Chem. Inf. Comput. Sci.*, **39**, 491 (1999).
6. A. R. Katritzky, I. B. Stoyanova-Slavova, D. A. Dobchev and M. Karelson, *J. Mol. Graph. Model.*, **26**, 529 (2007).
7. F. Gharagheizi and R. F. Alamdari, *QSAR Comb. Sci.*, **27**, 679 (2008).
8. S. J. Patel, D. Ng and M. S. Mannan, *Ind. Eng. Chem. Res.*, **48**, 7378 (2009).
9. C. J. Lee, J. W. Ko and G. Lee, *Korean Chem. Eng. Res.*, **48**, 717 (2010).
10. http://michem.disat.unimib.it/mole_db/.
11. L. Constantinou and R. Gani, *AIChE J.*, **40**, 1697 (1994).
12. X. Wen and Y. Qiang, *Ind. Eng. Chem. Res.*, **40**, 6245 (2001).
13. T. A. Albahri, *Ind. Eng. Chem. Res.*, **42**, 657 (2003).
14. Z. K. K. Zbransk and V. Rika, *Ind. Eng. Chem. Res.*, **47**, 2075 (2008).
15. C. J. Lee, G. Lee, W. So and E. S. Yoon, *Korean J. Chem. Eng.*, **25**, 568 (2008).
16. http://dippr.byu.edu/.
17. G. S. Patil, *Fire and Mater.*, **12**, 127 (1988).
18. K. Satyanarayana and M. C. Kakati, *Fire and Mater.*, **15**, 97 (1991).
19. E. Metcalfe and A. E. M. Metcalfe, *Fire and Mater.*, **16**, 153 (1992).
20. K. Satyanarayana and P. G. Rao, *J. Hazard. Mater.*, **3**, 81 (1992).
21. F.-Y. Hshieh, *Fire and Mater.*, **21**, 277 (1997).
22. L. Catoire and V. Naudet, *J. Phys. Chem. Ref. Data*, **33**, 1083 (2004).
23. V. N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York (1995).
24. A. B. Gandhi, J. B. Joshi, V. K. Jayaraman and B. D. Kulkarni, *Chem. Eng. Sci.*, **62**, 7078 (2007).
25. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
26. M. Schwwab, E. C. Biscaia, J. L. Monteiro and J. C. Pinto, *Chem. Eng. Sci.*, **63**, 1542 (2008).