Korean J. Chem. Eng., **25**(3), 568-574 (2008)

A new estimation algorithm of physical properties based on a group contribution and support vector machine

Chang Jun Lee, Gibaek Lee*, Won So and En Sup Yoon†

Department of Chemical and Biological Engineering, Seoul National University,
151-742 Shilim-dong, Gwanak-gu, Seoul 151-742, Korea
*Department of Chemical and Biological Engineering, Chungju National University,
380-702 Chungju, Chungbuk 380-870, Korea
(*Received 12 December 2006 • accepted 5 August 2007*)

Abstract−There are two ways to evaluate the properties of unknown chemical compounds. One is by traditional approaches, which measure the desired data from the experiments and the other is by predicting them in the theoretical approaches using a kind of prediction model. The latter are considered to be more effective because they are less time consuming and cost efficient, and there is less risk in conducting the experiments. Besides, it is inconvenient to conduct experiments to obtain experimental data, especially for new materials or high molecular substances. Several methods using regression model and neural network for predicting the physical properties have been suggested so far. However, the existing methods have many problems in terms of accuracy and applicability. Therefore, an improved method for predicting the properties is needed. A new method for predicting the physical property was proposed to predict 15 physical properties for the chemicals which consist of C, H, N, O, S and Halogens. This method was based on the group contribution method that was oriented from the assumption that each fragment of a molecule contributes a certain amount to the value of its physical property. In order to improve the accuracy of the prediction of the physical properties and the applicability, we extended the database, significantly modifying the existing group contribution methods, and then established a new method for predicting the physical properties using support vector machine (SVM) which is a statistical theory that has never been used for predicting the physical properties. The SVM-based approach can develop nonlinear structure property correlations more accurately and easily in comparison with other conventional approaches. The results from the new estimation method are found to be more reliable, accurate and applicable. The newly proposed method can play a crucial role in the estimation of new compounds in terms of the expense and time.

Key words: Group Contribution Method, Functional Group, Support Vector Machine, Property Estimation

INTRODUCTION

When designing new chemical compounds and processes, it is very important to assess and evaluate properties for understanding risks of the unknown reactions without experiments, because there may be loss of life and property and damage to the surrounding environment as well. Especially, when handling a dangerous reaction, unexpected chemical reactions may cause a runaway reaction which can lead to explosion as a result in abnormal rise in temperature. Therefore, it is necessary to evaluate the potential risks in new chemical compounds.

Information about the physical properties of chemical compounds is one of the best elements to evaluate and understand the chemical reactions. Especially, the heat release of the reactions is a good criterion as the preliminary screening procedure for a reaction of chemical compounds. The potential heat energy can be calculated by the physical properties such as the enthalpy of formation at standard state, H_f (298.15 K), and the heat capacity at constant pressure, CP (T) of the materials included in the reactions.

There are two ways to evaluate the amount of heat released from the reactions and the adiabatic temperature rise. One is the traditional approach, which measures the desired data from the experiments and the other is by predicting them in the theoretical approaches using a kind of prediction model. The latter is considered to be more effective because less time consuming and cost efficient, and there is less risk in conducting the experiments. Besides, it is inconvenient to conduct the experiments to obtain the experimental data, especially for new materials or high molecular substances.

Several methods for predicting the physical properties have been suggested so far. However, the existing methods have many problems in terms of accuracy and applicability. Therefore, an improved method for predicting the properties is needed.

In this study, we developed a new model for predicting the physical properties. In order to improve the accuracy of predicting of the physical properties and the applicability, we extended the database, modified the existing group contribution methods and then established new method for predicting the physical properties using SVM as a statistical method. The proposed method can be applied to predict 16 physical properties for the chemicals that consist of C, H, N, O, S and Halogens.

The main objective of this paper is to establish a predictive model for accurate physical properties with extended data base and SVM. At first the basic concepts of the theories used herein are explained and we explain the data base used herein and the way to construct the model. Finally, the performance of the proposed model is discussed and our conclusions are presented.

†To whom correspondence should be addressed.
E-mail: esyoon@pslab.snu.ac.kr

Table 1. 55 Functional groups as the input variables

I-Ending Group								
-CH3	≡CH	-COH	-NH2	-Br	-F	=O	≡N	-phenyl
=CH2	-OH	-CO2H	-NO2	-Cl	-I	=S	-SH	-H
II-Middle Group								
>C<	>C=	=C=	-C≡	=N-	-SO2-	-O-	o-B	m-B	p-B
-CH2-	>CH-	-CH=	>N-	-NH-	-SO-	-CO-	-CO2-	-S-	
III-Ring Group								
CH2	-CH	>C	N	NH	=C	R-C-R		
CH	CO	-C	-N	S	O	R-CH-R		
IV-Molecular Weight								
V-The Distinction between the structural isomers (consisting of 3-branch benzenes)								
The Distinction between the structural isomers (consisting of 4-branch benzenes)								
The Distinction between the geometric isomers								

ESTIMATION OF PHYSICAL PROPERTIES

By estimating the physical properties theoretically without any experiments, not only can the expenses but also the potential risk be reduced [1]. Several studies have been done on methods about theoretical estimations of the physical properties. The methodologies based on the molecular structure and the computational quantum chemistry are widely known to be reliable [2-7].

Although these methodologies coupled with the statistical thermodynamics can predict the properties more accurately, they are limited to only the low molecular substances. Even though the software such as GAUSSIAN or GAMESS can be used to predict many kinds of properties, they still cannot be applied to many cases, especially the high molecular substances.

Group contribution techniques, based on the assumption that each fragment of a molecule contributes a certain amount to the value of its physical property, are most widely used because they are relatively easy to use and have better abilities to make more accurate predictions. Joback [2-7] developed commercial software called CRANIUM, based on Joback's group contribution method. Constantinou [8] proposed a new group contribution method correcting the existing one by adding the second-order groups to the first-order ones. Lee [9] developed a group contribution method using neural network based on the same database that Joback used. But the major disadvantage of group contribution methods is that they perform unevenly on predicting. That is, although they may show good performance for specific compounds, they sometimes could be too poor to predict properties for other ones. This is a problem that results from an insufficient database about material properties. For many materials, physical properties were not available for the data which Joback [3] used, especially for the ideal heat capacity. Therefore, in order to improve the performance, it is necessary to expand the database. We improved the performance and the reliability using the Design Institute for Physical Properties (DIPPR), which is three times larger than those used in Joback's [3] and Lee's [9] method. The data used in this study can be downloaded from http://dippr.byu.edu.

In addition, we modified the representative functional groups, found more important ones by analyzing the correlations between the specific functional groups of the materials and their physical properties, and then proposed the final distinct groups for isomers. These are largely classified to five groups consisting of 18 ending groups, 19 middle groups and 14 ring groups, molecular weights and 3 isomer groups as shown in Table 1.

SUPPORT VECTOR MACHINE

The support vector machine (SVM) is a universal constructive learning procedure based on statistical learning theory. Its basic idea is to transform the signal to a higher dimensional feature space and find the optimal hyper-plane in the space that maximizes the margin between the classes. In addition to classification, the support vector methodology has also been introduced for linear and nonlinear function estimation problems [11,12]. A brief introduction of the principle of SVM for nonlinear function estimation is presented below

Consider the regression model that takes the form

$$f(x)=\omega^T\varphi(x)+b \tag{1}$$

with given training data $\{x_k, y)_{k=1}^N$ and, $\varphi(x)$: $R^n \rightarrow R^{n_k}$ a mapping to a high dimensional feature space which can be infinite dimensional and is only implicitly defined. For empirical risk minimization in the case of Vapnik one employs the following cost function.

$$R_{emp}=\frac{1}{N}\sum_{k=1}^{N}\left|y_k-\omega^T\varphi(x_k)-b\right|_\varepsilon \tag{2}$$

Here the so-called Vapnik's ε-insensitive loss function is defined as

$$\left|y-f(x)\right|_\varepsilon=\begin{cases}0, & \text{if}\left|y-f(x)\right|\le\varepsilon\\ \left|y-f(x)\right|-\varepsilon, & \text{otherwise}\end{cases} \tag{3}$$

The value ε in the Vapnik ε-insensitive loss function is the accuracy that one requires for the approximation. Additional slack variables ξ_k and ξ^*_k for k=1, ···, N are introduced. The optimization problem in the primal weight space becomes

$$\left[\begin{array}{l}\min\limits_{\omega,b,\xi,\xi^*} J_p(\omega,\xi,\xi^*)=\frac{1}{2}\omega^T\omega+c\sum\limits_{k=1}^{N}(\xi_k+\xi_k^*)\\ \text{such that } y_k-\omega^T\varphi(x_k)-b\le\varepsilon+\xi_k\\ \qquad\omega^T\varphi(x_k)+b-y_k b\le\varepsilon+\xi_k^*\\ \qquad\xi_k,\xi_k^*\ge 0, k=0,\ldots,N\end{array}\right] \tag{4}$$

After taking the Lagrangian and conditions for optimality one obtains the following dual problem:

$$\begin{aligned} \min_{\alpha,\alpha^*} J_D(\alpha,\alpha^*) &= -\frac{1}{2}\sum_{k,l=1}^{N}(\alpha_k-\alpha_k^*)(\alpha_l-\alpha_l^*)K(x_k,x_l) \\ &\quad -\varepsilon\sum_{k=1}^{N}(\alpha_k+\alpha_k^*)+\sum_{k=1}^{N}y_k(\alpha_k-\alpha_k^*) \\ \text{such that } &\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)=0,\ \alpha_k,\alpha_k^*\in[0,c] \end{aligned} \tag{5}$$

Here the kernel trick has been applied with $K(x_k, x_l)=\varphi(x_k)^T\varphi(x_l)$ for k=1, ···, N. The dual representation of the model becomes

$$f(x)=\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)K(x,x_k)+b \tag{6}$$

where α_k and α_k^* are the solution to the quadratic programming (QP) problem, Eq. (6), and b follows from the complementarily Karush Kuhn Tucker (KKT) conditions. The solution to the QP problem is global and unique provided that the chosen kernel function is a positive definite.

CONSTRUCTING THE PROPOSED ESTIMATION ALGORITHM

1. Preparing the Data

When developing the new model, the most important factor to ensure good performance is to obtain a high quality database. Joback's group contribution method [3] was built using about 480 chemical compounds. In consideration of the number of the input variables, the number of data was insufficient. For instance, if some missing data were excluded, there would only be 400 compounds remaining for the enthalpy of formation. In order to improve the reliability and the applicability, it is important to obtain a high quality database and expand the database.

The newly proposed model was built by using the data included in DIPPR. From this, we selected the useful data fields such as normal boiling point, normal melting point, critical pressure, etc. We included Gibbs energy of formation at standard conditions, enthalpy of formation at standard conditions and the gas heat capacity at constant pressure, which were considered to be important for evaluating the chemical hazards, and excluded some data fields that have many missing values. Finally, the database of the proposed model included about 16 physical properties for 1400 chemical compounds. The proposed method can predict 16 physical properties as shown in Table 2.

We define 55 representative functional groups, which are modified from the functional groups of the Joback Method [3]. The expansion of the database may add new compounds that have more various functional groups and it is necessary to modify the distinct functional groups. More important groups are found by analyzing the correlations between the specific functional groups of the materials and their physical properties.

Fifty-five distinct functional groups were classified into five groups: 18 ending groups, 19 middle groups and 14 ring groups, molecular

Table. 2. Predictable physical property from the proposed method

Physical Properties		Units
H_f (298.15 K)	Enthalpy of Formation	KJ/(gmol)
G_f (298.15 K)	Gibbs Energy of Formation	KJ/(gmol)
H_C (298.15 K)	Enthalpy of Combustion	KJ/(gmol)
CP_{gas} (298.15 K)	Gas Heat Capacity at Constant Pressure	KJ/(gmol·K)
$CP_{liq.}$ (298.15 K)	Liquid Heat Capacity at Constant Pressure	KJ/(gmol·K)
H_{fusion} (298.15 K)	Enthalpy of Fusion	KJ/(gmol)
S (298.15 K)	Entropy	KJ/(gmol)
T_B	Boiling Point	K
T_C	Critical Temperature	K
T_F	Flash Point	K
V_C	Critical Volume	m^3
P_C	Critical Pressure	atm
T_{TP}	Triple Point Temperature	K
P_{TP}	Triple Point Pressure	atm
T_M	Melting Point	K
AF	Acentric Factor	

Table 3. The examples of the structural isomers (consisting of 3-branch benzenes)

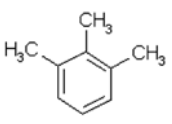
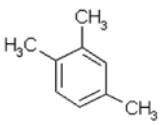
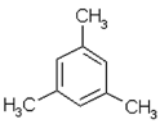
Structure [C9H12]			
Name	1,2,3-tri-methyl benzene	1,2,4-tri-methyl benzene	1,3,5-tri-methyl benzene
Input value	1	2	3

Table 4. The examples of the structural isomers (consisting of 4-branch benzenes)

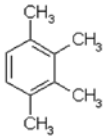
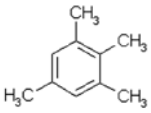
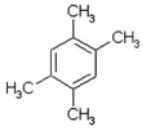
Structure [C10H14]			
Name	1,2,3,4-tetra-methyl benzene	1,2,3,5-tetra-methyl benzene	1,2,4,5-tetra-methyl benzene
Input value	1	2	3

Table 5. The example of the geometric isomers

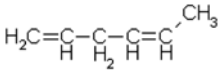
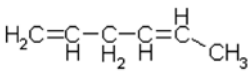
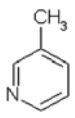
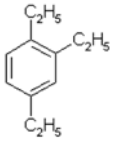
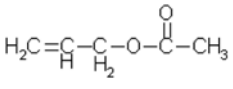
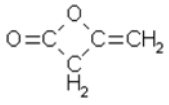
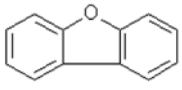
Structure [C6H10]		
Name	Cis-1,4-hexadiene	Trans-1,4-hexadiene
Input value	1	2

Table 6. The examples for the structure formulae of some compounds

		
3-Methylpyrudune [A]	1,2,4-Triethylbenzene [B]	2,5-Dihydrofuran [C]
		
3-Acetoxypropene [D]	Diketene [E]	Dibenzofuran [F]

weights and 3 distinct classifiers as shown in Table 1.

The examples of 3 distinct classifiers are shown in Table 3, Table 4 and Table 5. In the proposed model, the input variables for these classifiers are set as "ordinal variable," when running the model. As the numbers in these variables are used merely for distinguishing among the structural or the geometric isomers, the method for determining the order would be varied in the individual case.

2. Data Transformation

First, the data must be transformed to a form that is suited to build the model. After obtaining the data, we search the structural information of each compound in order to use the data as the input variable for modeling, using the chemical name, CAS number, molecular formula or molecular weight from the raw data. Table 6 shows the examples of some structural formulae.

With these structural formulae we need to transform this graphical information into numerical information according to the rules based on the new group contribution method. The 55 input variables would be determined for each molecule. Table 7 shows the 55 input variables for chemical compounds included in Table 6. Input variables consist of 52 distinct functional groups and 3 classifiers for the isomers as explained. The results of the transformation chemical compounds included in Table 6 are presented in Table 8.

Through these procedures, we can obtain the informative dataset containing the input value and the target value, 16 physical properties, for each chemical. This dataset was used in the input mode for the modeling procedure in this study.

3. Model Building

The goal of the present work is to construct the estimation based on the SVM. When constructing this, 55 distinct functional groups were used as input variables and 16 physical properties were used as output variables. If we used chemical compounds in Table 6, information of 55 input variables as shown in Table 8 was obtained, and these were used as input X. And 16 physical properties of chemical compounds called as output Y would be obtained from DIPPR database.

Table 7. The input variables of the model

In1	In2	In3	In4	In5	In6	In7	In8	In9	In10
-CH3	=CH2	≡CH	≡N	-NH2	-NO2	-SH	-Br	-F	-Cl
In11	In12	In13	In14	In15	In16	In17	In18	In19	In20
-I	-Phenyl	-COH	-CO2H	=O	-OH	=S	-H	>C<	>C=
In21	In22	In23	In24	In25	In26	In27	In28	In29	In30
=C=	-C≡	-CH2-	>CH-	-CH=	>N-	=N-	-NH-	-O-	-S-
In31	In32	In33	In34	In35	In36	In37	In38	In39	In40
-CO-	-SO2-	-SO-	o-B	m-B	p-B	-CO2-	*	**	CH2
In41	In42	In43	In44	In45	In46	In47	In48	In49	In50
CH	-CH	>C	-C	-N	NH	N	CO	O	=C
In51	In52	In53	In54	In55	* (1): Table 3 ** (2): Table 4 *** (3): Table 5				
S	R-C-R	R-CH-R	M.W	***					

Table 8. Transformation of the graphical information of Table 6 into the numerical values

Group	Functional group	A	B	C	D	E	F
Ending Group	-CH3	0	3	0	1	0	0
	=CH2	0	0	0	1	1	0
Middle Group	-CH2-	0	3	0	1	0	0
	-CH=	0	0	0	1	0	0
	-CO2-	0	0	0	1	0	0
	Structural isomers (consisting of 3-branch benzenes)	0	2	0	0	0	0
Ring Group	CH2	0	0	2	0	1	0
	CH	4	3	2	0	0	8
	-C	1	3	0	0	0	0
	N	1	0	0	0	0	0
	CO	0	0	0	0	1	0
	O	0	0	1	0	1	2
	=C	0	0	0	0	1	0
	R-C-R	0	0	0	0	0	4

Table 9. Comparisons of accuracy levels for the predicted physical properties

Physical properties	Units	Average absolute percent errors	
		Joback method [3]	New model
H_f (298.15 K)	KJ/(gmol)	16.78	3.40
G_f (298.15 K)	KJ/(gmol)	14.39	10.5
CP_{gas} (298.15 K)	KJ/(gmol·K)	5.52	0.75
T_C	K	4.08	1.90
T_F	K	11.07	0.18
V_C	m^3	6.16	0.14
T_B	K	2.99	1.83

SVM model was trained for 16 physical properties with X and Y. To construct an accurate model, a data set of 1400 chemical compounds data was obtained. Among them, the 470 samples were selected as training data randomly and others were used for the test data, respectively. A total of 16 physical property models were constructed and we tested the accuracy of the model. The proposed method was programmed by using MATLAB 6.1.

CASE STUDIES, RESULTS AND DISCUSSION

We improved the prediction model in terms of the following two ways. First, we improved the applicability by extending the database, using the DIPPR database that is three times larger than those used in Joback's [3] and Lee's method [9]. Second, we improved the accuracy and the reliability by modifying the existing group contribution methods and constructing the predictive model using SVM. The newly proposed method can distinguish the isomers, both the structural isomers and the geometric isomers. The new method based on SVM proved to show very good performance in estimating the physical properties compared to existing methods.

We compared the results from our model with those from the Joback method [3] for some predicted physical properties in Table 9. As shown in Table 9, we could obtain more accurate results and the most absolute average percent errors were within 2 percent for all the physical properties of 930 chemical compounds.

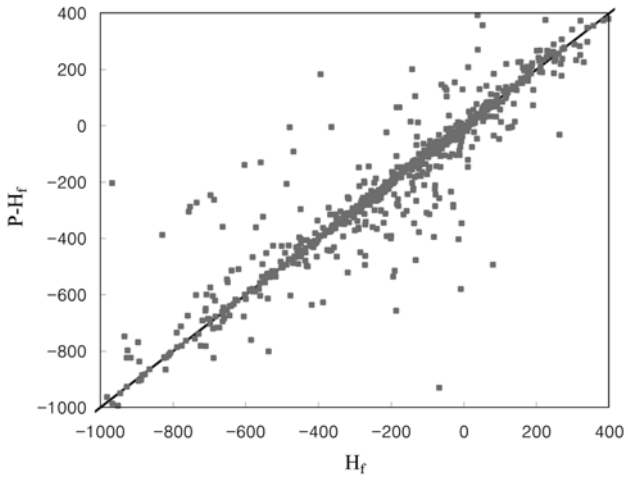


Fig. 1. Comparison between the predictive values and experimental ones of the standard enthalpy of formation (P-H_f represents the predicted result and H_f represents the experimental one).

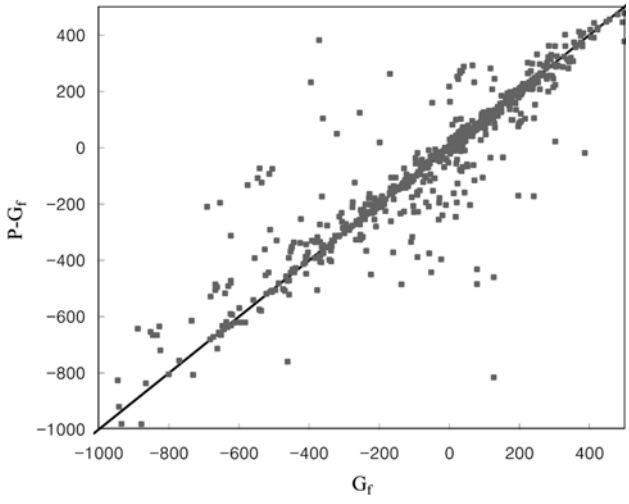


Fig. 2. Comparison between the predictive values and experimental ones of the standard Gibbs energy of formation (P-G_f represents the predicted result and G_f represents the experimental one).

Fig. 1 to Fig. 6 show a comparison with predictive values and experimental ones for the graphs of predictive values for the standard enthalpy of formation, the standard Gibbs energy of formation, the critical temperature, the flash point, the critical volume, and the normal boiling point. Absolute average percent errors and these figures verify that the proposed model has good performance. We can conclude that the proposed model has good predictive abilities for the physical properties. We expect that the developed model would be applied to evaluate the thermal hazards quantitatively. In the next section, we introduce a case study.

CONCLUSIONS

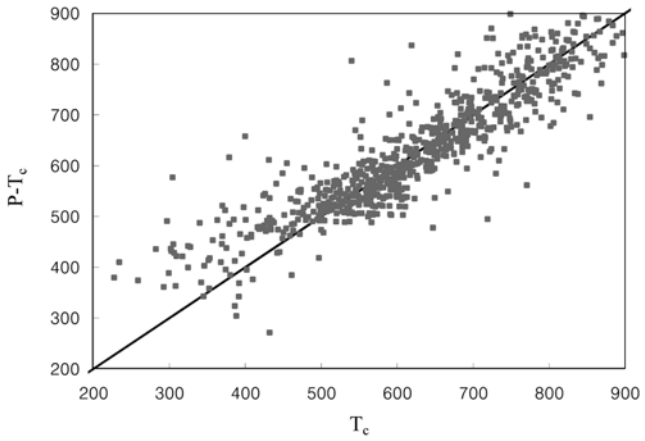


Fig. 3. Comparison between the predictive values and experimental ones of the critical temperature (P-T_c represents the predicted result and T_c represents the experimental one).

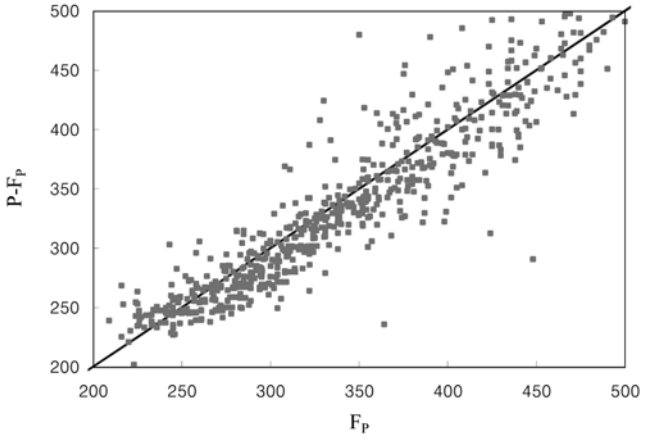


Fig. 4. Comparison between the predictive values and experimental ones of the flash point (P-F_P represents the predicted result and F_P represents the experimental one).

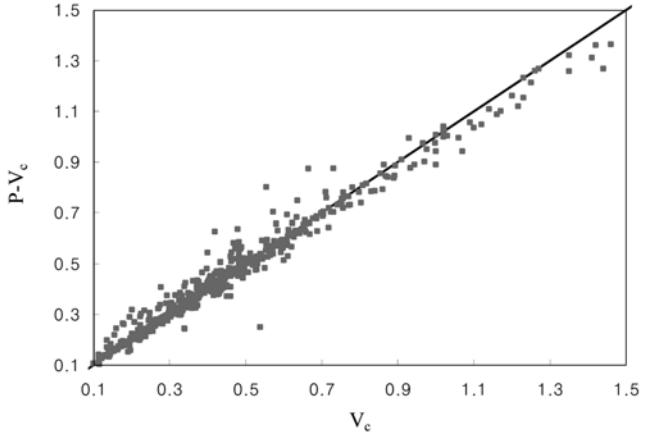


Fig. 5. Comparison between the predictive values and experimental ones of the critical volume (P-V_c represents the predicted result and V_c represents the experimental one).

With extended data base and SVM, which is a statistical theory, we proposed an algorithm for estimating properties of chemical compounds. To improve efficiency, we define 55 representative functional groups, which are modified from the functional groups of the Joback Method. Because the expansion of the database would be related to the reliability of an algorithm, we modified the existing data base and used more various functional groups. The application of SVM of functional groups has led to a new-group contribution method for the estimation of important physical and thermodynamic properties.

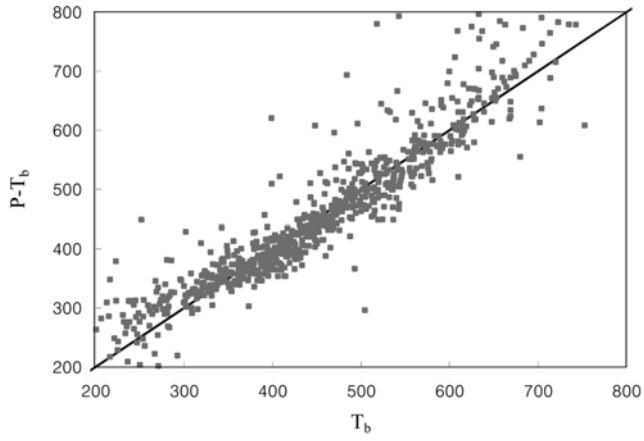


Fig. 6. Comparison between the predictive values and experimental ones of the normal boiling point (P-T_b represents the predicted result and T_b represents the experimental one).

Compared to other used group contribution methods, the proposed method exhibited improved accuracy and a wider range of applicability. Developed prediction method for physical properties of chemical compounds can be used to evaluate the properties of new chemical compounds wich leads to reducing the hazards of experiments.

Especially, with our proposed algorithm chemical reactions with hazardous runaway-reaction potential would be expected, and it would offer important information for new chemical manufacturing processes, relevant and appropriate ones in the context of process safety design.

ACKNOWLEDGEMNT

This work was supported by grant No. (R05-2002-000-00057-0) from the Basic Research Program of the Korea Science & Engineering Foundation and BK21.

NOMENCLATURE

ω : weight vector
$\varphi(x)$: map into feature space
b : constant of offset
y_k : target variables
x_k : input variables
R : the set of real numbers
R_{emp} : empirical risk function
N : number of training samples
ε : margin
ξ_k : slack variables

K() : kernel function
α_k : Largrange multipliers
H_f (298.15 K) : enthalpy of formation [KJ/gmol]
G_f (298.15 K) : Gibbs energy of formation [KJ/gmol)
H_C (298.15 K) : enthalpy of combustion [KJ/gmol]
CP_{gas} : gas heat capacity at constant pressure [KJ/gmol·K]
$CP_{liq.}$: liquid heat capacity at constant pressure [KJ/gmol·K]
H_{fusion} (298.15 K) : enthalpy of fusion [KJ/gmol]
S (298.15 K) : entropy [KJ/gmol]
T_B : boiling point [K]
T_C : critical temperature [K]
T_F : flash point [K]
V_C : critical volume [m^3]
P_C : critical pressure [atm]
T_{TP} : triple point temperature [K]
P_{TP} : triple point pressure [atm]
T_M : melting point [K]
AF : acentric factor

REFERENCES

1. M. S. Mannan, W. J. Rogers and A. Aldeeb, *A systematic approach to reactive chemicals analysis*, Proc. HAZARDS XVI, Manchester, U.K. 41-58 (2001).
2. C. Bruneton, C. Hoff and P. Barton, *Computers and Chemical Engineering*, **22**(6), 735 (1998).
3. K. G. Joback, *Unified approach to physical property estimation using multivariate statistical techniques*, S.M. Thesis, Massachusetts Institute of Technology, Cambridge (1984).
4. K. G. Joback, *Designing molecules possessing desired physical property values Vol.1*, Ph. D. Thesis, Massachusetts Institute of Technology, Cambridge (1989a).
5. K. G. Joback, *Designing molecules possessing desired physical property values Vol. 2*, Ph. D. Thesis, Massachusetts Institute of Technology, Cambridge (1989).
6. K. G. Joback, *Fluid Phase Equilibria*, **185**(1), 45 (2001).
7. H. J. Liaw, C. C. Yur and Y. F. Lin, *J. of Loss Prevention in the Process Industries*, **13**(6), 499 (2000).
8. L. Constantinou and R. Gani, *AIChE Journal*, **40**(10), 1697 (1994).
9. K. H. Lee, J. Y. Jung and I. B. Lee, *Korea J. Chem. Eng.*, **31**, 744 (1993).
10. DIPPR (Design Institute for Physical Properties), http://dippr.byu.edu
11. V. N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, New York, U.S.A. 53-67 (1995).
12. V. Cherkassky and F. Muler, *Learning from data: Concepts, theory, and methods*, John Wiley & Sons, New York, U.S.A. 353-387 (1998).
13. H. J. Liaw, C. J. Chen and C. C. Yur, *J. of Loss Prevention in the Process Industries*, **14**(5), 371 (2001).