Printed in the Republic of Korea

ANALYTICAL SCIENCE & TECHNOLOGY

Vol. 28, No. 3, 187-195, 2015

http://dx.doi.org/10.5806/AST.2015.28.3.187

QSPR analysis for predicting heat of sublimation of organic compounds

Yu Sun Park, Jong Hyuk Lee, Han Woong Park and Sung Kwang Lee★

Department of Chemistry, Hannam University, Daejeon 305-811, Korea

(Received April 24, 2015; Revised May 12, 2015; Accepted May 12, 2015)

유기화합물의 승화열 예측을 위한 QSPR분석

박유선 · 이종혁 · 박한웅 · 이성광★

한남대학교 화학과

(2015. 4. 24. 접수, 2015. 5. 12. 수정, 2015. 5. 12. 승인)

Abstract: The heat of sublimation (HOS) is an essential parameter used to resolve environmental problems in the transfer of organic contaminants to the atmosphere and to assess the risk of toxic chemicals. The experimental measurement of the heat of sublimation is time-consuming, expensive, and complicated. In this study, quantitative structural property relationships (QSPR) were used to develop a simple and predictive model for measuring the heat of sublimation of organic compounds. The population-based forward selection method was applied to select an informative subset of descriptors of learning algorithms, such as by using multiple linear regression (MLR) and the support vector machine (SVM) method. Each individual model and consensus model was evaluated by internal validation using the bootstrap method and y-randomization. The predictions of the performance of the external test set were improved by considering their applicability to the domain. Based on the results of the MLR model, we showed that the heat of sublimation was related to dispersion, H-bond, electrostatic forces, and the dipole-dipole interaction between inter-molecules.

요 약: 승화열은 대기 유기 오염물질의 확산에 관련된 환경적인 문제를 해결하거나, 위험한 화학 물질의 위해성을 평가하는 데에 중요한 변수이다. 하지만 실험적으로 승화열을 측정하려면 많은 시간과 비용이 소모 되며, 그 실험자체도 복잡하고 위험하다. 따라서 본 연구에서는 유기화합물의 승화열을 간단하게 예측하는 모델을 개발하기 위하여 정량적 구조-물성 상관관계 연구를 이용하였다. 군기반 전진선택방법을 적용하여 다중선형회귀방법과 서포트 벡터 머신과 같은 학습방법에 적합한 분자표현자들을 선택하도록 하였다. 개별 모델과 복합모델들은 부스트래핑 방법과 y-임의추출법에 의해 내부검증이 되었다. 외부 테스트 데이터의 예측 성능은 적용범위를 고려하므로서 개선되었다. 다중선형회귀모델에 따르면, 승화열은 분자간의 분산력, 수소결합, 정전기적 상호작용, 쌍극자-쌍극자 상호작용과 관련이 있는 것을 나타낼 수 있었다.

Key words: Heat of sublimation, QSPR, MLR, SVM, consensus model

★ Corresponding author
Phone : +82-(0)42-629-8874 Fax : +82-(0)42-629-8811
E-mail : leesk@hnu.kr

1. 서 론

화합물의 상변화에 따른 엔탈피 변화량(생성열, 증기열, 승화열)들은 분자 내 및 분자간의 상호작용을 나타내므로 화학, 물리학, 화학공학, 환경공학 등 다양한 분야에서 매우 중요하다. 그 중에서도 표준 승화열은 표준상태(1 bar, 298.15K)에서 고체상과 기체상 사이의 상 변화에 따른 1 mol 당 엔탈피 변화라고 정의되는 열역학적인 양으로 고체상태에서 분자간의 상호작용을 나타내며, 또한 결정구조의 안정성을 나타내기도 한다. 표준 승화열은 대기 중으로 오염물질의 이동을 평가하거나, 환경에서의 수명, 물질의 변색이나 염료의 확산 정도를 나타낼 때도 사용된다[1]. 또한 화약류와 같이 반응성과 위험성이 큰 화합물들의 폭발 성능을 평가하는데 있어서도 중요한 변수가 된다. 하지만 다종의 화합물에 대한 승화열을 실험적으로 측정하기 위해서는 시간과 비용도 많이 소모되므로 컴퓨터를 이용하여 이론적으로 계산하려는 연구들이 진행되어 왔다. Politzer교수[2]는 분자표면적에서 양자역학적으로 계산한 정전기적 포텐셜을 계산하고 그 값들의 통계적 편차 값을 이용하여 승화열 및 증발열을 예측하는 경험적인 모델을 개발하였다. Gharagheizi 등은[3] 각 분자의 작은 작용기 그룹에 대하여 승화열의 부분 기여도 값을 계산하여 전체 화합물의 승화열을 계산하는 그룹 기여도 모델을 개발하여 실온에서의 승화열을 예측하였다. 이러한 연구들처럼 통계를 기반으로 경험적인 모델을 구현할 때는 훈련에 사용되는 데이터가 적을 경우 실제로 모델을 적용할 수 있는 화합물의 범위가 적어지게 되고, 많은 실험데이터를 사용할 경우에도 신뢰도가 부족한 데이터를 포함하면 모델의 예측결과가 왜곡되는 문제가 있다. 일부 모델은 예측된 결과데이터를 오인하여 훈련데이터로 사용된 경우[4]도 있었다. 본 연구팀에서는 이전 연구에서 신뢰도가 높은 대량의 실험 데이터를 수집하여 2 차원 화학구조로부터 간단하게 계산할 수 있는 분자 표현자(descriptor)를 통해 선형/비선형 통계적 모델을 구현한 Quantitative Structure Property Relationships (QSPR)분석을 수행한 예[5,6]가 있다. 본 연구에서는 표준 승화열 데이터를 이용한 QSPR모델을 개발하여 승화열과 관련된 화학구조적 특성 분석 및 미지 화합물 예측을 수행하고자 한다. 이러한 접근은 상대적으로 복잡하게 계산하여야 하는 양자화학적 결과와 비교하여 간단하면서도 쉽고 빠르게 계산할 수 있는 장점이 있으며, 신뢰도 높은 다량의 데이터를 활용함으로써 외부데이터에 의한 검증의 신뢰도를 높일 수 있다고 판단이 된다.

2. 연구 방법

Fig. 1은 본 연구에서 수행한 QSPR 과정을 도표로 나타낸 것이다. 데이터 수집, 표현자 계산, 데이터전처리, 표현자 선택, 모델내부검증, 복합모델 생성, 모델 외부검증, 적용범위 설정 순으로 진행하였다.

2.1. 승화열 데이터

표준 승화열은 문헌[7,8]으로부터 총 1291 개의 화합물에 대한 실험값을 수집하였다. 신뢰할 수 있는 데이터를 사용하기 위하여 중복 실험된 화합물의 실험값은 평균처리를 하였고, 상대표준편차(RSD)를 계산하여 5%이하의 화합물만을 사용하였다. 또한 Fullerene 같은 특수 화합물은 제거하여 총 923 개의 유기 화합물에 대한 표준 승화열의 실험값을 정리하였고, 추가지원정보(supporting information)의 *Table* 1S에 나타내었다. 걸러진 923 종의 승화열 데이터는 무작위로 선택하여 모델 개발에 사용하는 훈련데이터로 462 종, 모델의 예측능력을 평가하고 검증하는데 사용하는 외부테스트 데이터로 461 종으로 나누었다. 구분된 훈련데이터와 외부 테스트 데이터들은 분자량, logP(옥탄올-물 분배계수), 수소결합 주게, 수소결합 받게,

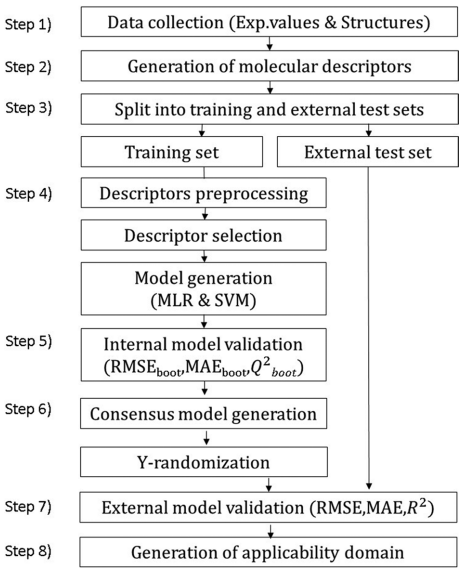

Fig. 1. Schematic diagram of QSAR analysis for predicting heat of sublimation.

Table 1. Molecular descriptor range and means for all data set

	All data set (n=923)	Training set (n=462)	External test set (n=461)
MW	59.068 ~ 532.683 (183.084)	60.055 ~ 428.532 (183.676)	59.068 ~ 532.683 (182.491)
SlogP (Predicted logP)	-3.835 ~ 11.701 (2.356)	-3.585 ~ 10.779 (2.357)	-3.835 ~ 11.701 (2.356)
No. H-bond acceptors	0 ~ 12 (1.986)	0 ~ 12 (2.011)	0 ~ 6 (1.961)
No. H-bond donors	0 ~ 6 (0.802)	0 ~ 6 (0.801)	0 ~ 6 (0.803)
No. Rotatable bonds	0 ~ 25 (1.369)	0 ~ 24 (1.433)	0 ~ 25 (1.306)
Topological Polar Surface Area	0 ~ 220.200 (40.264)	0 ~ 220.200 (41.131)	0 ~ 137.460 (39.395)

회전 가능한 결합수, 극성표면적을 비교하였을 때, 범위와 평균값을 보아서 두 데이터 그룹은 적절하게 잘 나누어진 것을 확인할 수 있었다(*Table* 1).

2.2. 분자 표현자(Molecular Descriptor)

QSPR 분석은 화학구조와 예측하고자 하는 물성 간의 정량적인 상관관계를 통하여 화학구조로부터 물성 값을 예측하는 것을 기초로 한다. 화학구조는 그 자체적으로 물성과 상관관계를 표현하기 어려우므로, 화학구조의 특징을 수치로 표현하는 분자 표현자(molecular descriptor)를 통해서 물성을 예측하는 모델을 구현하게 된다. 본 연구에서는 PreADMET[9] 프로그램을 이용하여 2 차원 화학 구조로부터 다양한 분자 표현자들을 계산하였다. 사용된 표현자의 종류로는 구성요소 표현자(constitutional descriptors), 기하학적 표현자(geometrical descriptors), 정전기적 표현자(electrostatic descriptors), 위상학적 표현자(topological descriptors), 물리화학적 표현자(physicochemical descriptors), 화학적 상호 작용점(chemical feature) 간의 결합한 거리 정보를 포함하는 CATS(Chemical Advanced Template Search) 표현자[10]를 계산하여 576 종을 사용하였다. 계산된 분자 표현자중에서 모델개발에 적합하지 않는 표현자들을 미리 선별하여 제거하기 위하여 RapidMiner[11] 프로그램을 사용하였다. 부적합한 표현자들은 3 단계를 통해서 제거되었는데, 첫 번째 단계는 훈련데이터 화합물 전체에 대하여 각 표현자들의 표준편차(SD)가 낮아서 구조적 특성을 잘 표현할 수 없는 경우(SD<0.01), 두 번째 단계는 표준 승화열과의 1:1상관관계가 매우 낮은 경우(Y vs X : R^2 < 0.01), 세 번째 단계에서는 표현자 들간의 상관관계가 높아 정보가 중복되는 경우(X vs X : R^2 > 0.9)에는 그 중 한가지를 제거하였다. 이 과정을 거쳐 부적합한 표현자들은 제거하였고, 남은 356 개의 표현자들을 다음단계인 표현자 선택과정을 통해 QSPR모델 개발에 사용하였다.

2.3. 표현자 선택 과정 (descriptors selection)

데이터 전처리 과정을 통해 미리 선별된 표현자들로부터 QSPR모델을 구현할 때, 학습 방법에 적합한 표현자를 선택하는 과정이 필요하다. 본 연구에서는 모델의 적합도 평가에 따라 표현자를 선택하는 wrapper 형태의 전진 선택 방법(forward selection method)을 사용하였는데, 모델의 적합도를 가장 크게 증가시키는 표현자들 순으로 모델에 포함하는 방법이다. 이 방법은 처음 표현자가 선택된 이후에 추가적으로 표현자가 모델에 포함될 때는 포함되기 이전과 비교하여 적합도 향상이 가장 많이 되는 표현자를 선택하도록 한다. Rapidminer프로그램에서 구현되는 이 방법은 기존의 전진선택방법처럼 한 종류의 모델에 대하여 적용되지 않고, 처음 모델의 수는 적용되는 표현자의 수만큼 개별 모델군(population)으로 지니고 있고, 개별 모델에 추가적으로 전진선택방법으로 추가 표현자들을 포함하게 되므로, 최종적으로는 다양한 개별 모델 군들을 생성되게 한다. 다른 표현자 선택방법보다 빠르게 최적 표현자 집합을 선택할 수 있으며, 추후에 모델 내부검증과정과 동시에 수행하여 내부검증에 적합한 최적의 모델을 선택할 수 있다. 모델개발에 사용된 기계학습방법으로는 선형방법인 다중 선형 회귀(Multiple Linear Regression, MLR)방법과 비선형 방법인 서포트 벡터 머신(Support Vector Machine, SVM) 방법을 사용하였고, 반복표본추출 방법인 부스트래핑(Bootstrapping)기법을 이용하여 내부검증을 수행하였다. 부스트래핑 기법을 적용하여 내부검증데이터에 대하여 얻은 적합도(Q^2_{boot}, $RMSE_{boot}$, MAE_{boot})는 훈련데이터만 이용하여 얻은 적합도(R^2, RMSE, MAE)과 구분되며, 그 계산식은 다음과 같다.

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N} (y_i - y_{cal,i})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \right)^2$$

$$Q^2_{boot} = 1 - \left(\frac{\sum_{i=1}^{N_{iv}} (y_i - y_{pred,i})^2}{\sum_{i=1}^{N_{iv}} (y_i - \bar{y})^2} \right)^2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y_{cal,i})^2}{N}}$$

$$RMSE_{boot} = \sqrt{\frac{\sum_{i=1}^{N_{iv}}(y_i - y_{red,i})^2}{N_{iv}}}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - y_{cal,i}|}{N}$$

$$MAE_{boot} = \frac{\sum_{i=1}^{N_{iv}}|y_i - y_{pred,i}|}{N_{iv}}$$

N은 훈련데이터의 수를 나타내며, R^2, RMSE(root mean square error), MAE(mean absolute error)는 훈련데이터 전체에 대하여 결정되는 결정계수, 평균 제곱근 편차, 평균 절대 오차를 나타낸다. N_{iv}는 부스트래핑기법에 의해 추출되지 않은 내부검증데이터의 수를 나타내며, Q^2_{boot}, $RMSE_{boot}$, MAE_{boot}는 마찬가지로 부스트래핑기법에 의해 추출되지 않은 내부검증데이터의 결정계수, 평균 제곱근 오차, 평균 절대 오차를 나타낸다.

2.4. 다중 선형 회귀 (Multiple Linear Regression, MLR)

MLR 방법은 여러 개의 독립변수(X)들을 이용하여 종속변수(Y)를 다음과 같이 다중 선형 회귀 모델을 만드는 방법으로 기본 형태는 다음과 같다.

$$Y = a_0 + a_1X_1 + a_2D_2 + \ldots + a_nX_n$$

이 식에서 $X_1, X_2, \ldots X_n$ 들은 독립변수로 분자의 특징을 나타내는 표현자(descriptor)를 사용하게 되며, 독립변수의 회귀계수인 $a_1, a_2, \ldots, a_n$ 은 해당 독립변수를 제외한 다른 모든 독립변수를 고정시켰을 때 해당 독립변수를 단위 별로 증가시킬 때의 종속변수(Y)의 평균 변화량을 의미한다.

2.5. 서포트 벡터 머신 (Support Vector Machine, SVM)

SVM 방법은 주어진 독립변수 데이터들을 kernel함수를 사용하여 고차원의 공간으로 변환하여 고차원의 공간에서 오차를 최소한으로 하기 위해 최대 차이(maximal margin)를 지니는 최적의 초평면(Optimal hyperplan)을 통하여 분류를 하는 방법[12]이다. 초기에는 2진형태의 분류데이터를 예측하는데 사용되어졌으나, 현재는 ε-무감도 손실함수(ε-insensitive loss function)을 도입하여 비선형 회귀 예측에 사용할 수 있게 되었다. 본 연구에서는 SVM방법 중에서 추가적인 매개 변수를 제거할 수 있는 γ-SVM방법[13]을 적용하였다. SVM방법에서 최적화 해야 할 매개변수인 C, γ, ν에 대해서는 C는 0.1, 1, 10, 100, γ는 0.001~0.1에 0.001씩, 마지막으로 ν는 0.1~0.5에 0.1씩 변환하면서 부스트래핑 방법에 의한 $RMSE_{boot}$값이 가장 낮은 조건으로 최적 매개변수 조건을 결정하였다. 이 과정을 통해서 최적 매개 변수 조건은 C =100, γ=0.1, ν=0.4이였다.

2.5. 모델 검증 과정

QSPR예측모델 구현할 때는 훈련데이터의 실험값들에 대한 적합도 계산과 더불어 모델의 예측 가능성을 검증하는 과정도 매우 중요하다. 본 연구에서는 앞에서 설명한 부스트래핑 방법을 사용한 내부 검증(internal validation)과정과 Y-randomization과정을 통한 모델의 우연 일치 가능성을 확인하였다. 이와 더불어 모델 개발에 전혀 사용되지 않은 외부테스트 데이터를 사용하여 외부 검증(external validation)을 수행하여 최종적으로 예측 능력을 평가하였다.

부스트래핑 방법은 훈련데이터로부터 중복을 허용하여 원래 훈련데이터와 동일한 수의 데이터를 표본추출하는 방법이다. 표본으로 추출된 데이터는 내부훈련데이터로 모델을 구현하는데 사용하며, 선택되지 않은 데이터는 내부검증데이터로 사용하여 학습방법에 대한 모델의 적합도를 계산하는데 사용한다. 이 방법을 통해 훈련데이터의 63%는 내부 훈련 데이터로, 37%는 내부 검증 데이터로 대체적으로 나누어지게 된다. 본 연구에서는 부스트래핑 기법으로 100번정도 반복하여 표본 추출을 하며, 매번 내부검증데이터의

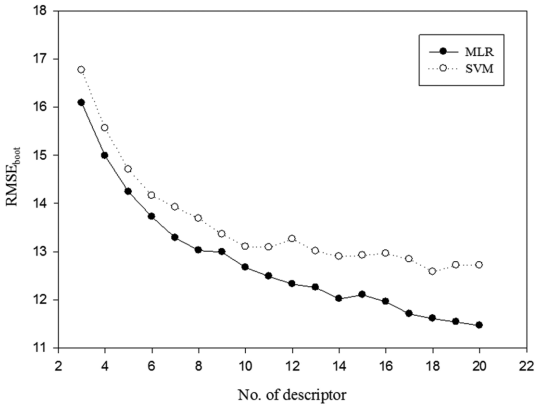

Fig. 2. Influence of the number of descriptors in MLR and SVM on $RMSE_{boot}$.

Table 2. Comparative statistical performance of MLR, SVM and consensus models

Model (#descriptors)	Training Set (n=462)			Bootstrapping validation set			External test set (n=461)		
	R^2	RMSE[a]	MAE[b]	R^2_{boot}	$RMSE_{boot}$	MAE_{boot}	R^2_{ext}	$RMSE_{ext}$	MAE_{ext}
MLR(14)	0.819	11.422	8.301	0.797±0.027	12.018±0.944	8.718±0.0497	0.744	14.430	10.544
MLR(20)	0.840	10.736	7.830	0.817±0.024	11.463±0.842	8.387±0.463	0.769	13.728	9.822
SVM(10)	0.874	9.627	7.192	0.752±0.033	13.403±1.213	9.591±0.657	0.719	15.137	10.713
SVM(18)	0.912	8.144	6.185	0.766±0.029	13.037±1.136	9.305±0.583	0.743	14.468	10.078
MLR(14)+MLR(20)	0.839	10.783	7.858	0.815±0.024	11.586±0.977	8.394±0.508	0.766	13.785	9.972
SVM(10)+SVM(18)	0.899	8.706	6.571	0.768±0.031	12.993±1.192	9.248±0.615	0.736	14.656	10.274
MLR(14),(20)+SVM(10)	0.874	9.637	7.051	0.830±0.023	11.152±0.995	8.033±0.506	0.781	13.351	9.463
MLR(14),(20)+SVM(18)	0.882	9.317	6.858	0.831±0.022	11.118±0.963	8.033±0.480	0.783	13.289	9.447
SVM(10),(18)+MLR(14)	**0.896**	**8.859**	**6.617**	**0.812±0.025**	**11.751±1.059**	**8.438±0.537**	**0.768**	**13.740**	**9.747**
SVM(10),(18)+MLR(20)	0.900	8.651	6.428	0.818±0.024	11.559±1.058	8.278±0.530	0.775	13.544	9.529
Y-Scrambling_MLR(14)	0.070	25.902	19.975	0.042±0.022	26.022±1.480	20.307±1.017	-	-	-
Y-Scrambling_SVM(10)	0.071	25.950	20.097	0.037±0.023	26.018±1.502	20.279±1.029	-	-	-

[a]root mean square error, [b]mean absolute error.

모델 적합도를 계산하여 그 평균값으로 모델의 예측도를 검증하는 내부 검증결과로 사용하였으며, 이를 이용하여 모델에 사용되는 표현자의 수에 따른 변화를 확인하였다(*Fig.* 2).

많은 수의 표현자로부터 학습방법에 적합한 표현자를 선별하는 과정에서는 우연하게 실험데이터와 모델결과가 일치할 가능성도 있다. 이를 확인하기 위하여 Y-randomization과정을 수행하는데, 모델 개발에 사용되는 종속변수(승화열 데이터)를 화합물간에 무작위로 섞은 뒤에 모델개발과정을 동일하게 진행하여 그 적합도를 비교하는 방법으로, 무작위로 섞어 반복한 평균 적합도 결과($R^2_{Y\text{-}random}$)와 실제 모델의 결과가 큰 차이가 나는지를 확인하도록 하였다. 개발된 모델의 최종적인 평가는 모델개발에 전혀 사용하지 않는 외부 테스트 데이터를 통하여 예측 능력과 견고성을 평가하였다.

3. 결과 및 토론

3.1. 학습 방법에 따른 모델 및 복합모델

2 가지 학습방법(MLR, SVM)과 함께 연동하여 표현자를 선택하게 하는 래퍼(wrapper)방식의 전진선택(forward Selection)방법을 통하여 분자표현자를 3 개부터 20 개까지 선택하게 하였다. 부스트래핑방법에서 내부검증데이터의 $RMSE_{boot}$값이 최소가 되는 모델을 최적모델로 선택하고자 하였다. 대체적으로 적은 수의 표현자 집합으로 구성된 모델일수록 통계적으로 의미가 있다. *Fig.* 2에서 나타난 바와 같이 MLR 방법과 SVM방법에서는 대체적으로 표현자 수가 증가 할수록 부스트래핑기법에 의해 내부 검증데이터의 $RMSE_{boot}$값도 낮아지는 것을 알 수 있었다. MLR모델들은 대체적으로 14 종 이상의 표현자를 포함할 때까지 모델 성능이 개선되는 것을 볼 수 있었으나, SVM모델은 10 종 이상의 표현자를 포함하여도 모델 성능이 개선되지는 않았다. 그러므로 본 연구에서는 표현자수에 대하여 $RMSE_{boot}$값의 기울기 변화가 감소하기 시작하는 시점들의 모델들을 중심으로 MLR모델과 SVM모델의 성능결과를 *Table* 2에 나타내었다. MLR 모델은 14 종, 20 종의 모델을, SVM모델은 10 종, 18 종을 포함하는 모델에 대하여 나타내었다. MLR모델들은 461종의 외부테스트 데이터에 대하여 R^2_{ext}값이 0.720~0.769의 범위를 나타내었고, SVM모델들도 외부테스트 데이터에 대하여 R^2_{ext}값이 0.719~0.743의 범위 정도에 비교적 잘 예측되는 것을 확인할 수 있었다.

각각의 학습방법(MLR, SVM)에서 선별된 개별 모델들에 대하여 그 예측결과를 혼합하여 평균 처리한 다양한 복합모델(consensus model)을 생성하여 기존의 개별 모델들과 훈련데이터, 부스트래핑 방법, 외부테스트 데이터에 대한 성능결과도 함께 비교하였다. 대체적으로 복합모델들은 개별 MLR 또는 SVM모델보다 훈련데이터의 뿐만 아니라, 부스트래핑의 내부검증데이터, 외부테스트 데이터에 대해서도 예측능력이 더 뛰어난 것을 확인할 수 있었다. 복합모델 중에서 표현자 14종과 20종을 포함하는 MLR모델, 표현자 18종를 포함하는 SVM모델을 혼합한 복합모델의 내부검증 결과가 가장 좋은 결과를 확인할 수 있었고, 이 복합모델에 대한 외부테스트 데이터의 예측결과도 가장 우수한 결과를 결과를 볼 수 있었다. 최적의 복합모델을 이용하여 훈련데이터와 외부테스트 데이터

의 실험값과 예측값은 *Table* 1S(추가자료, Supporting information)에 나타내었다. 최종적으로 모델의 표현자가 우연하게 적합하였을 가능성을 검증하기 위하여 Y-randomization 과정을 수행하였다. 본 연구에서는 이 과정을 100회 정도 반복하여 진행하였고, 그 성능평가의 평균결과를 *Table* 2의 아래에 나타내었다. 14종의 표현자로 개발된 MLR모델의 성능결과는 R^2_{boot}=0.764~0.817를 나타내었고, 10종의 표현자로 개발된 SVM모델의 성능결과는 R^2_{boot}=0.752~0.766의 결과를 나타낸 반면에, Y-randomization을 수행한 모델들의 평균 결과는 R^2_{boot} =0.037~0.042를 나타내었다. 실제 개발된 모델과 Y-randomization과정으로 수행한 모델의 성능결과가 큰 차이를 나타내므로 우연 상관(chance correlation)에 의한 결과가 아님을 증명할 수 있었다.

3.2. QSPR 모델 분석

일반적으로 선형 모델인 MLR방법은 표현자들의 선형적인 관계를 수식으로 표현하므로, 모델 자체가 black box로 여겨 지는 SVM, ANN방법보다 표현자를 통해서 모델을 분석하고 해석하기가 용이하다. 본 연구에서도 14종의 표현자를 포함한 MLR모델을 바탕으로 각각의 표현자에 대한 설명과 회귀계수 및 t-ratio값을 *Table* 3에 나타내었다. MLR수식에서 회귀계수의 부호는 승화열과의 상관관계를 나타내고, t-ratio 값은 회귀 계수와 그 표준 오차와의 비율을 나타낸 값으로 이 값의 절대값이 크면 상대적으로 해당 표현자의 기여도가 크다고 평가할 수 있다. *Table* 3에 나타낸 표현자 중에서 t-ratio의 절대값이 큰 순서는 Chi0 > NoHobndD > FrVSA_ch_gr > Chi3c > Ldipole_i 순이며 이 순서대로 모델에서 중요한 역할을 하는 것으로 확인할 수 있다. 가장 큰 기여도를 나타낸 표현자는 Chi0이며 분자연결지수로 화합물 구조를 나타낸 고유값 이다. 각각의 원자에서 인접한 결합원자수(정점도, vertex degree)값에 대하여 역수 제곱근의 총합

Table 3. Descriptors included in MLR model for regression coefficients and intercept in the best MLR model for HOS

No.	Molecular descriptors	Description	Coefficient	t-ratio
1	Nring5	The number of aromatic bonds composed a molecule	-4.232	-3.397
2	NHbondD	The number of functional groups to donor lone pair electrons to form hydrogen bond	11.140	16.934
3	Monocyclic_compounds_carbocycles	The number of functional groups of monocyclic compounds carbocycles	-9.157	-7.692
4	FrVSA_ch_gr	Fraction of 2D Van der Waals chargable groups surface area	101.752	11.074
5	Ldipole_i	The average of the charge differences over all boned atom pairs	46.286	8.813
6	WPSA2	This is connected with the electrostatic feature and shape to represent Interactions of the polarity between molecules	-0.113	-5.635
7	Estate_SssssC	Information about atom-atom electrostatic interactions and topological environment in the molecule	4.644	4.640
8	Estate_SdsN		3.151	6.582
9	Distance_E_state_min_max		2.315	5.032
10	Chi0	A general scheme based on the Randic index to calculate also zero-order and higher-order descriptors and became known as the molecular connectivity indices.	10.455	21.378
11	Chi3c		-11.573	-10.000
12	SC4p	The number of subgraphs of a given type and order	-0.454	-3.881
13	CATS_binary_Hyd_Acc_2	Chemical advanced template search descriptor	5.780	4.340
14	CATS_binary_Pos_Neg_3		30.957	5.220

Table 4. Comparison of statistical performance of single models and consensus models using applicability domain

Model (#descriptors)	#.Compound (coverage)	R^2_{ext}	$RMSE_{ext}$	MAE_{ext}
MLR(14) – within AD	442(95.88%)	0.752	14.041	10.251
SVM(10) – within AD	448(97.18%)	0.732	13.979	10.171
MLR(20) – within AD	441(95.66%)	0.765	13.836	9.863
SVM(18) – within AD	454(98.48%)	0.756	14.150	9.926
CONSENSUS (all data)	461(100%)	0.783	13.289	9.447
CONSENSUS (within AD)	458(99.35%)	0.785	13.235	9.385

으로 분자 내 원자들의 수가 많을수록 커지게 된다. *Table* 3에서 이 이 모델에서는 회귀계수값이 양의 값을 지니므로, 분자의 크기가 증가할수록 일반적으로 분산력이 증가되면서 승화열이 증가하는 것으로 예상할 수 있다. 두 번째로 기여도가 큰 것은 NoHbondD 표현자로 분자 내 수소결합 주게의 수를 의미하며 O-H, N-H, S-H와 같이 수소결합을 형성하기 위해 수소를 제공할 수 있는 원자들의 수를 나타낸다. 회귀계수 값이 양이 값을 지니므로, 수소결합 받게의 수가 증가할수록 분자간 수소결합의 증가로 승화열이 증가하는 것을 알 수 있다. 특히 분자크기와 더불어 승화열을 증가시키는 중요한 요인임을 확인 할 수 있었다. 그 다음으로 FrVSA_ch_gr으로 분자 내 전하를 띠고 있는 원자들의 Van der Waals표면적 비율을 나타낸 표현자이다. 이 표현자의 회귀계수는 양의 값을 나타내므로 이 표현자 값이 증가할수록 승화열이 증가하는 것을 의미한다. 이것은 즉 분자 내 전하를 지닐 원자들의 표면적이 높으므로 상대적으로 분자간 정전기적 상호작용이 증가함에 따라 각 분자간의 상호작용을 끊어내고 승화하기 위한 승화열이 더 필요한 것으로 볼 수 있다. 그 다음으로 기여도가 큰 Chi3c 표현자는 세 개의 원자가 한 원자에 결합되어 있는 형태인 클러스터 형태의 부분 구조 내 원자들의 결합원자수간 곳의 역수 제곱근으로 주로 화학구조의 모양이 선형인 형태보다는 가지 형태의 모양을 지닐수록 그 값이 커지게 된다. 이 표현자의 회귀계수는 음의 값을 지니므로, 분자 내 결합이 가지 형태로 나열된 구조보다 선형형태로 나타낸 구조가 승화열이 더 높다는 것을 의미한다. 선형구조일수록 분자간 접촉 면적이 더 증가할 수 있기 때문으로 추정된다. Ldipole_i표현자는 부분 쌍극자 지수를 나타내는 것으로 쌍극자-쌍극자 상호작용을 설명할 수 있다. 각 결합된 원자간 전하의 차이에 대한 평균값으로 이 값이 커질수록 분자 내 결합된 원자간 전하차이가 크다는 것을 의미하며 이것은 분자 내 쌍극자로 편중될 수 있음을 나타내는 것이고, 인근 동일 분자와의 쌍극자-쌍극자 상호작용이 커지게 될 것으로 예상된다. 이들의 상관계수가 양의 값을 지니므로 이를 반영한다고 볼 수 있다.

본 MLR모델을 통해선 화학분자의 승화열은 분산력, 수소결합, 정전기적 상호작용, 쌍극자-쌍극자 상호

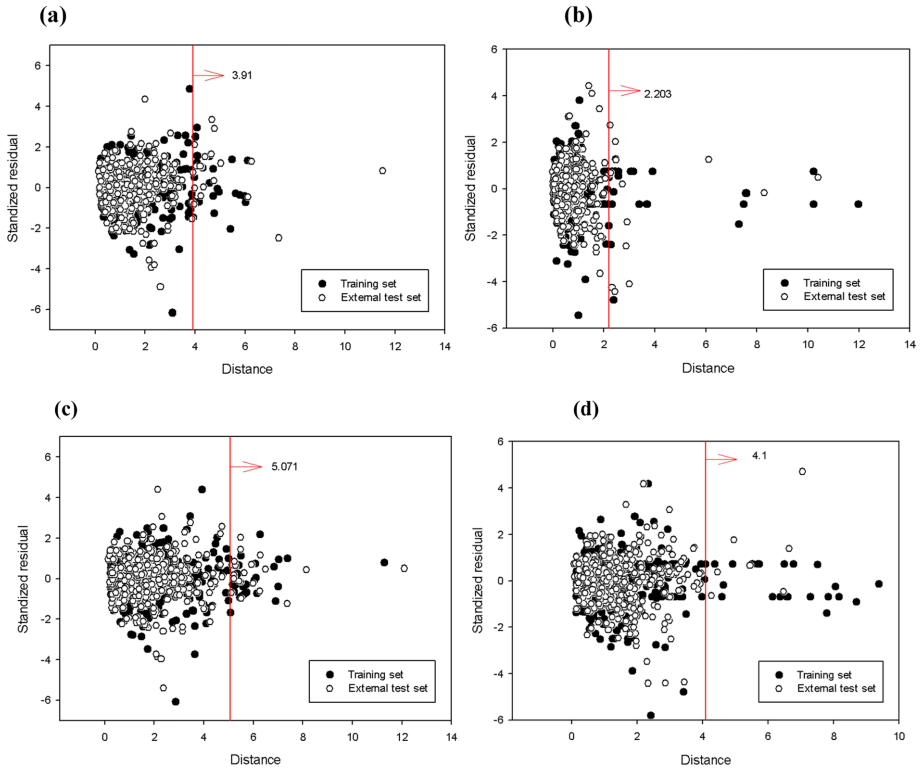

Fig. 3. Plot of standard residuals versus distance between external test compounds and nearest training set using descriptors in (a) 14-variable MLR, (b) 10-variable SVM-10, (c) 20-variable MLR, and (d) 18-variable SVM model.

작용 등에 의해서 증가하는 것을 확인할 수 있었다.

3.3. 모델의 적용 범위(applicability domain)

개발된 모델들은 실제로 훈련데이터로 예측모델을 개발하였기 때문에 훈련데이터와 유사한 화합물 범위에서는 신뢰성 있는 예측결과를 나타낼 수 있다. 이러한 모델의 신뢰성 있게 예측이 가능한 적용범위를 훈련데이터의 분자표현자를 기반으로 나타낸 분자유사성으로 설정하고자 하였다. 모델의 적용범위는 처음 문헌[14]에서 제안한 k-최근접 이웃방법(k-nearest neighbor method)을 통해서 외부테스트 데이터에 대하여 훈련데이터에서 가장 가까운 거리를 계산하여 일정 거리 내에 있는 화합물들을 적용할 수 있는 범위로 나타내도록 하였다. 분자간 거리 계산은 모델에 사용된 표현자들을 통해서 하며, 계산 전에 훈련데이터를 기반으로 z-score방법을 이용한 정규화(normalization)을 시행하여 표현자들의 평균값과 표준편차를 동일하게 조절한 뒤에 거리를 계산하도록 한다. 적용범위내의 기준은 훈련데이터 간에 가장 근접 화합물 5 종과의 거리를 화합물 별로 계산하고, 낮은 거리 순으로 누적화하여 전체 95%에 해당하는 지점을 적용범위의 한계점으로 결정하도록 하였다. *Table* 3에서 내부검증데이터에 대하여 가장 좋은 결과를 나타낸 모델은 14 종 표현자의 MLR모델, 20 종 표현자의 MLR모델, 그리고 18 종 표현자의 SVM모델을 이용한 복합모델이며, 적용범위를 적용한 성능결과는 *Table* 4에 나타내었다. 각 개별 모델의 경우 적용범위에 벗어나는 외부테스트 화합물들을 제외한 결과는 원래의 모델보다 성능결과가 더 향상된 것을 확인 할 수 있었고, 복합모델의 결과도 적용범위에 벗어나는 화합물의 예측결과를 제외하여 평균값으로 예측한 결과도 다소 향상되는 것을 확인할 수 있었다. 적용범위를 벗어난 화합물들의 구조를 확인한 결과 *Fig.* 4(b)에 나타낸 바와 같이 여러 개의 다중고리를 포함하는 방향족 탄화수소(Polycyclic Aromatic Hydrocarbons, PAHs) 종류의 화합물임을 확인할 수 있었다. 이러한 경우는 승화열을 예측하는데 있어서 이러한 다중고리화합물을 잘 표현할 수 있는 표현자를 모델에 사용하지 못하였거나, 메커니즘적으로 특이성을 보일 수 있다는 점을 예상할 수 있고, 추후 예측모델의 적용범위를 고려할 때, 이러한 다중고리 방향족 탄화수소 화합물들은 제외될 필요가 있어 보인다.

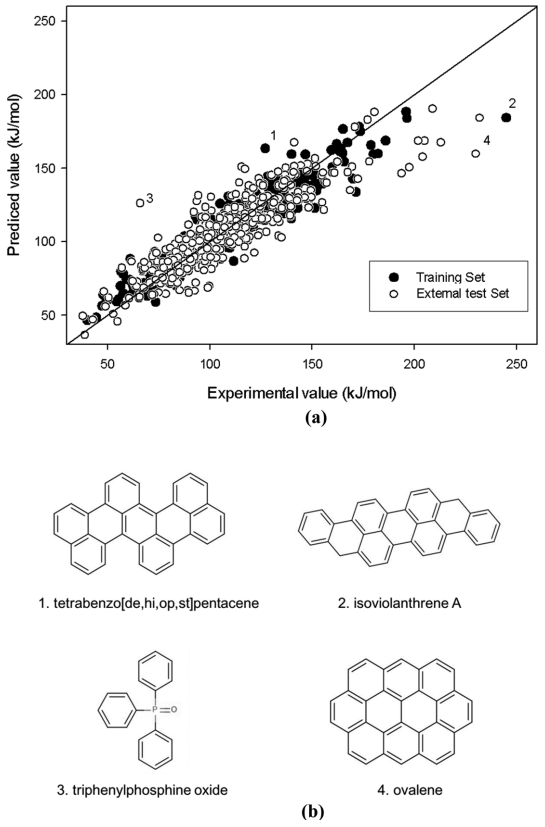

Fig. 4. (a) Plots of experimental versus predicted HOS value for consensus model, (b) structure of the compounds showed largest difference from experimental value.

4. 결 론

본 연구에서는 QSPR 분석을 통해서 상온에서 유기화합물의 승화열을 예측 할 수 있는 적합한 모델을 개발하였다. 복잡한 계산과정을 거치지 않고 통계적인 방법을 바탕으로 간단한 2 차원 화학구조로부터 물성을 예측 할 수 있었고, 개발된 모델은 검증 과정을 통해서 적합성, 우연 일치 가능성, 신뢰성, 예측 능력을 확인할 수 있었다. 그리고 검증된 모델들은 서로 보완할 수 있도록 다양한 복합모델을 적용하여 검증하였으며, 최종적으로 2 종의 MLR모델과 1종의 SVM모델을 적용한 복합모델(R^2_{boot}:0.831, $RMSE_{boot}$:11.118, MAE_{boot}:8.033)의 예측 능력이 단일 모델들 보다 향상된 결과를 나타냄을 확인할 수 있었다. 화합물의 분자 유사성을 통해 모델의 예측 가능한 적용 범위를 확인할 수 있었고, 적용범위를 적용하여 예측 능력을 보다 향상됨을 확인 할 수 있었다. 본 복합모델에서는 다중고리 방향족 탄화수소의 화합물이 오차가 큰 것을 확인할 수 있었고,

적용범위를 고려할 때 이러한 다중고리 방향족 탄화수소는 예측도가 감소하므로 제외하여야 함을 알 수 있었다.

따라서 앞으로 사용자의 필요에 따라서 적용 가능한 범위를 설정하고, 검증된 QSPR모델을 선택해서 사용한다면 시간과 비용과 위험을 줄일 수 있는 효율적인 연구를 돕거나, 환경적인 문제를 해결하는데 가장 적합한 방법이 될 수 있다는 것을 확인 할 수 있었다. 본 연구결과는 추후 개별적으로 프로그램화하여 일반인들에게 제공할 예정이다.

감사의 글

본 연구는 민・군겸용기술과제의 지원으로 수행되었으며, 이에 감사 드립니다.

이용 가능한 지원정보

Table 1S는 승화열 실험데이터가 있는 923종 화합물의 이름, 실험값, 예측모델의 예측값을 포함하고 있으며, 이에 대한 정보는 학회지 홈페이지를 통해서 무료로 받을 수 있다.

References

1. K. Nakajoh, E. Shibata, T. Todoroki, A. Ohara, K. Nishizawa and T. Nakamura, *Environ. Toxicol. Chem.,* **25**(2), 327-336 (2006).
2. P. Politzer, Y. Ma, P. Lane and M. C. Concha, *Int. J. Quantum Chem,* **105**(4), 341-347 (2005).
3. F. Gharagheizi, P. Ilani-Kashkouli, W. E. Acree, A. H. Mohammadi and D. Ramjugernath, *Fluid Phase Equilib.,* **354**, 265-285 (2013).
4. F. Gharagheizi, *Thermochim. Acta,* **469**(1-2), 8-11 (2008).
5. E. H. Jean, J. H. J. Park, Jin Hee and S. K. Lee, *Anal. Sci. Technol.,* **24**(6), 533-543 (2011).
6. I. S. Song, J. Y. Cha and S. K. Lee, *Anal. Sci. Technol.,* **24**(6), 544-555 (2011).
7. W. Acree and J. S. Chickos, *J. Phys. Chem. Ref. Data,* **39**(4), 043101 (2010).
8. M. A. V. Roux, M. Temprado, J. S. Chickos and Y. Nagano, *J. Phys. Chem. Ref. Data,* **37**(4), 1855-1996 (2008).
9. PreADMET Ver.2.0.2.0, BMDRC: Seoul Korea, 2007.
10. G. Schneider, W. Neidhart, T. Giller and G. Schmid, *Angew. Chem. Int. Ed. Engl.,* **38**(19), 2894-2896 (1999).
11. RapidMiner Ver.5.3.012, Rapid-I: Stockumer, Germany.
12. C. Cortes and V. Vapnik, *Mach. Learn.,* **20**(3), 273-297 (1995).
13. B. Schölkopf, A. J. Smola, R. C. Williamson,and P. L. Bartlett, *Neural Comput.,* **12**(5), 1207-1245 (2000).
14. A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR Combi. Sci.,* **22**(1), 69-77 (2003).